ISSN 2277-3061

Motif Discovery and Data Mining in Bioinformatics

Nooruldeen Nasih Qader, Hussein Keitan Al-Khafaji
University of Sulaimani, Computer Science, Iraq
nnq1@yahoo.co.uk
Alrafidain University College, Computer Communication, Iraq
dr.hkm1811@yahoo.com

ABSTRACT

Bioinformatics analyses huge amounts of biological data that demands in-depth understanding. On the other hand, data mining research develops methods for discovering motifs in biosequences. Motif discovery involves benefits and challenges. We show bridge of the two fields, data mining and Bioinformatics, for successful mining of biological data. We found the motivation and justification factors lead to preferring naturalistic method research for Bioinformatics, because naturalistic method depends on real data. The method empowers Bioinformatics techniques to handle the true properties and reducing assumptions for un-modeled or uncover biodata phenomena. The empowerment comes from recognizing and understanding biodata properties and processes.

Indexing terms/Keywords

Bioinformatics; biosequence; biodata; naturalistic; motif; mining; DNA; database; algorithm; TFBS

Academic Discipline And Sub-Disciplines

Computer Science; Bioinformatics; Biology;

SUBJECT CLASSIFICATION

Data Mining in Bieinformatics;

TYPE (METHOD/APPROACH)

Literary Analysis; Survey;

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol 13, No. 1

editor@cirworld.com

www.cirworld.com, www.ijctonline.com

INTRODUCTION

Bioinformatics involves the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, and biochemistry to solve biological problems usually at the molecular level. Recent progress in biology, medical science, Bioinformatics, and biotechnology has led to the accumulation of tremendous amounts of biodata that demands in-depth analysis. On the other hand, data mining research has led to the development of numerous efficient and scalable methods for mining interesting patterns in large DB's. The question becomes how to bridge the two fields, data mining and Bioinformatics, for successful mining of biological data [63].

Studies that depend mainly on hypotheses models lead to the derivation of imperfect biological models such as models used to generate the sequences (i.e., the Bernoulli Model, hidden Markov model) and motifs representation.The motivation and justification factors lead to preferring naturalistic method research for Bioinformatics, because naturalistic method depends on real data [47].

The rest of this paper is organized as follows:basic concepts of Bioinformatics in a way that helps the biodata analysis in section 2. Benefits of motif finding arise in many fields; in this context Bioinformatics field is coming in advance that explained in section 3. Follow by presenting some challenges of motif discovery in section 4. Motif discovery is presented with gene regulation is section 5. Biodata analysis from a data mining perspective is explained in section 6. Some biosequences DB's have introduced in 7. Several approaches for motif discovery are demonstrated in 8 and 9. Finally the paper ends with conclusion and future work.

BASIC CONCEPTS OF BIOINFORMATICS

The ability to predict the behavior, the function, or the structure of biological entities (such as genes and proteins), as well as interactions among them, plays a fundamental role in the discovery of information to help biologists to explain biological mechanisms [44]. The following sub-sections describe some basic concepts of Bioinformatics:

Biological Preliminary Notions

Biologists and computer scientists use different terms to denote the same concepts. Hence "character", "nucleotide", "base" and "symbol" are to be equivalent, and use the term "motif" as a synonym for "pattern". Bioinformatics is a combination of biology and computer sciences; therefore it is necessary to introduce some basic concepts of biology as follows:

Cells are the fundamental working units of every living system. All the instructions needed to direct their activities are contained within the chemical DNA.

DNA from all organisms is made up of the same chemical and physical components. The DNA sequence is the particular side-by-side arrangement of bases along the DNA strand (e.g., ATTCCGGA). This order spells out the exact instructions required to create a particular organism with its own unique traits.

The **genome** is an organism's complete set of DNA. Genomes vary widely in size: the smallest known genome for a free-living organism (a bacterium) contains about 600,000 DNA base pairs (bp), while human and mouse genomes have 3 billion bp[26]. Except for mature red blood cells, all human cells contain a complete genome. Differences between two living organisms are mostly due to the differences in their genomes [28]. DNA in the human genome is arranged into 24 distinct **chromosomes**-physically separate molecules that range in length from about 50 million to 250 million bp. Cell, chromosome and DNA are shown in **Error! Reference source not found.**.

Each chromosome contains many **genes**, the basic physical and functional units of heredity. Genes are specific sequences of bases that encode instructions on how to make proteins. The human genome is estimated to contain 30,000 to 40,000 genes.

Proteins perform most life functions and even make up the majority of cellular structures. Proteins are large complex molecules made up of smaller subunits called amino acids. The constellation of all proteins in a cell is called its **proteome**. Unlike the relatively unchanging genome, the dynamic proteome changes from minute to minute in response to tens of thousands of intra- and extracellular environmental signals [26].

Exon is the DNA segments that are present in mature ribonucleic acid (RNA). **Promoters** are normally present close to the first exon and serve as binding sites for the transcription factors. **Enhancers**, on the other hand, are normally present at a much greater distance from the first exon and aid in the binding of transcription factors to the promoters. Both these regulatory elements thus work together to bring about the expression or suppression of genes [46].

Gene Expression

Three kinds of chain are the central molecular building blocks of life: DNA, RNA and proteins. The DNA molecule is a double-stranded long sequence composed of four types of nucleotides (A, C, G and T). It has the double-helix structure, and stores the hereditary information. RNA molecules are very similar to DNAs, composed also of four nucleotides (A, C, G and U). Proteins are chains of 20 different basic units, called amino acids. The gene is the fundamental unit on the genomic DNA which contains the required information to carry out the biological functions of cells. Proteins carry out almost all essential functions in a cell. In order to make a protein, the corresponding gene has to be *transcribed* into mRNA, and then the mRNA is *translated* into protein. Transcription and translation of the cell [15] are shown in **Error!**

Reference source not found.. Understanding gene expression, i.e., how protein factors interact with DNA regions to affect transcription is an important problem in molecular biology [28].

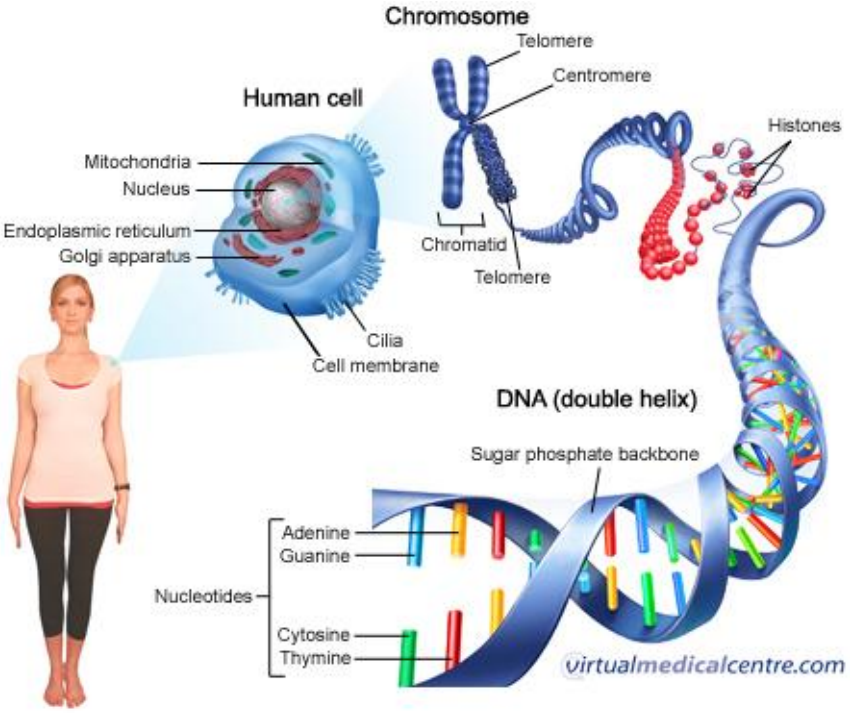


Fig 1: Cell, chromosome and DNA [6]

Motifs Types

Motifs are frequently occurring patterns. Motifs in biological sequences can indicate the presence of certain biological characteristics. In general, these could represent patterns in any kind of biological sequences such as DNA sequences, RNA sequences, protein sequences, etc[46]. In DNA, a motif may correspond to a protein binding site; in proteins, a motif may correspond to the active site of an enzyme or a structural unit necessary for proper folding of the protein [23].

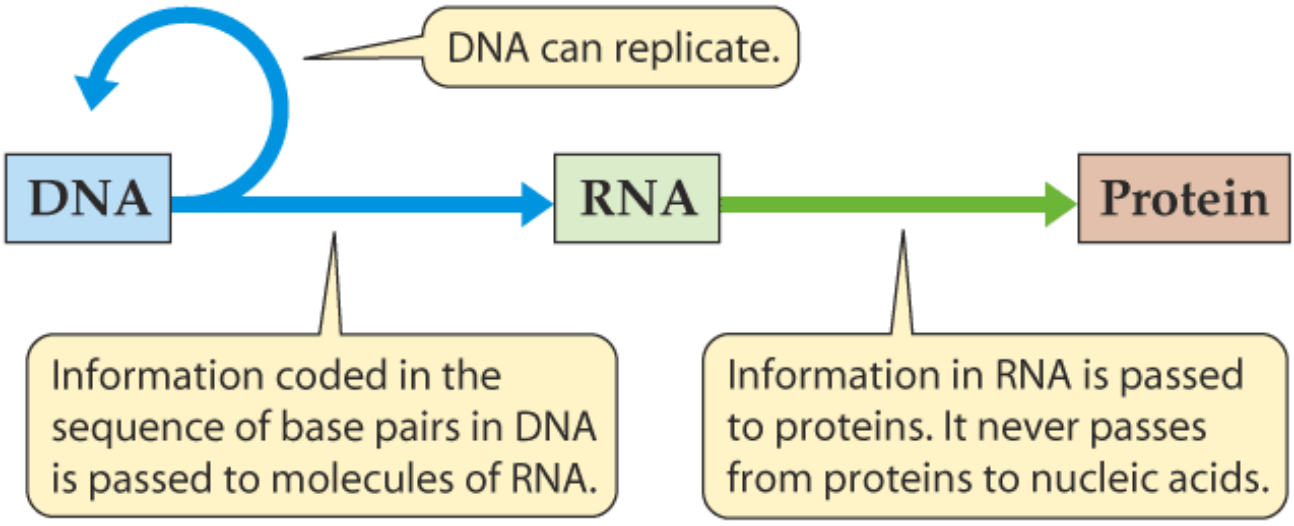


Fig 2: Transcription and translation [10]

Motifs could be monad or composite; a pattern with a single component is also called a single pattern, most approaches to pattern discovery focus on monad patterns that correspond to relatively short contiguous string with some mismatch that appear unexpectedly many times in a statistically significant way. But in the search for more complex methods have appeared that extract DNA sites composed of two binding sites. The first attempts to identify several binding sites, called multiple motifs. However, many of the actual regulatory signals are composite patterns that are groups of nomad signals [5].

Composite motifs can be thought of as “compound patterns” made of a list of mended motifs, or patterns, and a list of intervals that specifies at what distances adjacent motifs should occur. It can be classified into two main types. If no

variable gaps are allowed in the motif, it is called a *simple motif*. If variable gaps are allowed in a motif, it is called a *structured motif*[10, 50].

Transcription Factors Binding Site (TFBS)

An important part of gene regulation is mediated by specific proteins, called transcription factors (TFs), which influence the transcription of a particular gene by binding to specific sites on DNA sequences, called TFBS. TFs are protein products of certain genes that serve as regulators of the expression of other genes. These proteins diffuse into the cell, recognize and bind to certain sequence segments in DNA. Upon binding, they can induce changes of chromatin structure or interact with basal transcriptional machinery, and thereby *initiate, repress or modulate* transcription of genes close to the binding site. Such binding sites are relatively short stretches of DNA, normally 5 to 30 nucleotides long, and are located in the so-called promoter regions. One believes that the interaction between TF and TFBS plays a key role for regulation of gene expression [12, 56]. In many cases, more than one TF may cooperatively regulate a gene. Such patterns are called composite regulatory patterns. This type of pattern can occur when a gene is regulated by two TF[28, 65].

The locations on the DNA where proteins bind to are called cis-acting elements, also referred to as cis-regulatory elements or cis-elements. Different TFs may have different binding motifs, and multiple TFs can bind cooperatively to a cis-element that contains several different binding motifs. At any time, the particular composition of TFs active in the nucleus of a cell determines which subset of cis-elements is bound and activated in this cell. This combinatorial binding allows a few hundred TFs to control the spatial and temporal expression patterns of tens of thousands of genes. The development of a fertilized egg to an advanced embryo with complex body plans and organs may be, to a first approximation, regarded as the successful implementation of the transcription programs encoded in these cis-elements [67].

Benefits of Motifs Discovery

Motifs are believed to be extremely important in biology and Bioinformatics. The discovery of information encoded in biological sequences is assuming a distinguished role in identifying genetic diseases and in deciphering biological mechanisms. This information is usually encoded in patterns frequently occurring in the sequences [16].

Motif discovery is the critical step to understand the regulatory mechanism of genes. The motifs can represent patterns which activate or inhibit the transcription process and are responsible for regulating gene expression. In Bioinformatics, motif discovery is becoming very important because they represent conserved sequences which can be biologically meaningful. It could be essential to the analysis and understanding of the biological data. *If a pattern occurs frequently, it ought to be important or meaningful in some way*[57]. Motifs are recurring patterns in biodata that are presumed to have a biological function. Often they indicate sequence specific binding sites for proteins such as nucleases and TFs. Others are involved in important processes at the RNA level, including ribosome binding, mRNA processing and transcription termination, growing usefulness in defining genetic regulatory networks and deciphering the regulatory program of individual genes make them important tools for computational biology in the post-genomic era [13]. Also, Motifs are important in understanding the underlying cause of amyloid illnesses, pharmaceutical and industrial purposes [42].

The discovery of patterns in DNA, RNA, and protein sequences has led to the solution of many vital biological problems. For instance, the identification of patterns in nucleic acid sequences has resulted in understanding biological processes as basic as finding binding sites in amino acids, finding regulatory information within either DNA or RNA sequences and protein domains, the RNA transcription, the determination of open reading frames, identification of promoter elements of genes, identification of intron/exon splicing sites, location of RNA degradation signals; identification of alternative splicing sites. Finding motifs can be equally crucial for analyzing interactions between viruses and cells or identification of disease-linked patterns. In protein sequences, patterns have proven to be extremely helpful in domain identification, location of protease cleavage sites, identification of signal peptides, protein interactions, determination of protein degradation elements, identification of protein trafficking elements, discovery of short functional motifs, …etc[49]. Motif discovery for protein sequences is important for identifying structurally or functionally important regions and understanding proteins' functional components, or active sites [15].

CHALLENGES OF MINING MOTIFS

Biology is encoded in molecular sequences but deciphering this encoding remains a grand scientific challenge. Functional regions of DNA, RNA, and protein sequences often exhibit characteristic but subtle motifs; thus, computational discovery of motifs in sequences is a fundamental and much-studied problem [20].

Motif discovery is a fundamental problem in molecular biology because it has significant applications, refer to section 0 and [25]. On the other hand, motif discovery is a challenging task, mainly because they exhibit a high degree of degeneracy making their distinction difficult from random artifacts. For this reason, algorithms for motif discovery often suffer from impractical high false positive rates and return noisy models that are not useful [3].

Motif discovery in DNA data sets is a challenging problem domain because of lack of understanding of the nature of the data, and the mechanisms to which proteins recognize and interact with its binding sites are still complicated to biologists [35]. If the motif is TFBS; recent studies have shown that the underlying regulatory mechanisms of TFBS are complex, dynamic (especially in higher organisms) and can be arranged in multiple hierarchical levels [43]. Algorithms for motif discovery often are not useful to characterize TFBS's[3, 7, 40]. In TFBS, where the motifs are mostly monad, the challenges present in motif discovery include:

The motifs are never exactly the same as the actual conserved sequence. There is always considerable sequence variability present with respect to a monad motif.

Motifs are very short signals as compared to the size of the DNA sequence under consideration.

The regulatory sequences containing the motifs may sometimes be located very far from coding regions that they regulate. This makes it difficult to determine the portion of the DNA sequence that should be analyzed.

The regulatory sequences may, at times, be present on the opposite strand from the coding sequence they regulate.

A difficulty in discovering composite motifs is more than of monad motifs; because one of the components in the group may be too weak. Since the traditional monad based motif finding algorithms usually output one or a few high scoring patterns, they often fail to find regulatory signals consisting of week monad parts. For example, a set of yeast S. serivisiae genes regulated by two TFs with experimentally verified sites, URS1 and UASH that occur relatively near each other. Although URS1 is strongly conserved and easily found with monad motif search approaches, UASH is too weak to be discovered with these approaches [40]. A possible approach to detecting composite patterns is to attempt to detect each part of the pattern separately and then reconstruct the composite pattern. Unfortunately, some parts may be too weak to discover with scoring based approaches and consume huge memory resource with suffix tree based approaches in the way they are failing to complete the job with current desktop computers. In addition, exhaustive pattern search has a drawback of excessive time requirement. Some algorithms (i.e., SMaRTFinder, SMOTIF2) followed this approach without considering monad scoring and results to consume huge memory resource, sometimes goes to fail the operation.

Although most genes in the human genome have been identified and annotated, the cis-elements that control their expression are largely unknown. Thus a challenging problem whose solution requires not only new experimental data but also new statistical and computational methods [15, 67]. The ability to generate sequence and catalog interactions between DNA and the proteins that bind it has increased dramatically. Computational ability to deal with this data has also grown, albeit more slowly. Despite all these efforts, the problem has not satisfactorily been solved yet, as shown in the assessment of 13 common motif discovery algorithms by [55]. Recently, steps have been taken to precisely understand what makes the problem so difficult. The ability of popular motif models turn out to have comparable discriminative power, but are not sufficient to capture all motifs [38].

MOTIF DISCOVERY AND GENE REGULATION

Biologists today are interested in understanding how different genes in the genome are regulated and the way they interact with each other. Motifs discovery in DNA is an important step in the process towards understanding the transcriptional program in cells [62]. Discovery of TFBSs or DNA motifs in promoter regions of genes plays a key role in understanding the regulations of gene expression [56]. A gene is a segment of DNA that is the blueprint for protein; gene can be decoded to produce functional products like protein [11]. Genes seldom work alone; rather, they cooperate to produce different proteins for a particular function [12, 18, 54]. One TF is usually involved in the regulation of many genes, and the TFBSs that the TF recognizes often exhibit strong sequence specificity and conservation. In order to understand how genes' mRNA expression levels are regulated in the cell, it is crucial to identify TFBSs. Although much progress has been made in developing experimental techniques for identifying these TFBSs, these techniques are typically expensive and time-consuming. They are also limited by experimental conditions, and cannot pinpoint the binding sites exactly [15].

Efforts of motifs discovery have long been frustrated by the limited availability and accuracy of TFBS motifs. It is also known that the transcription machinery will recognize binding sites even if the motifs do not occur exactly, i.e., allowing for some nucleotide substitutions. A TFBS may contain a number of highly degenerate positions (known as wildcards in pattern recognition). In principle, even more powerful data mining techniques can be developed to integrate different types of data and provide more accurate predictions of TFBS. In particular, more robust methods of finding composite motifs may help to unravel the regulatory structure of promoters [8, 36].

BIODATA ANALYSIS FROM A DATA MINING PERSPECTIVE

Analyzing and interpreting sequence data is an important task in Bioinformatics. In Bioinformatics, the motifs discovery could be essential to the analysis and understanding of the biological data. If a pattern occurs frequently, it ought to be important or meaningful in some way [4, 57, 65].

Existing sequence mining algorithms mostly focus on mining for exact motifs. However, a large class of applications, such as biological DNA and protein motif mining, require efficient mining of "approximate" patterns that are contiguous. This approximate is of particular importance in Bioinformatics, where the challenge is to detect motifs that occur frequently in a given set of biodata. These motifs can provide clues regarding the cis-regulatory elements, which are important repeated patterns along the biological sequence. The repeated occurrences of these motifs are not always identical, and some copies of these sequences may differ from others in a few positions [18, 19, 54]. Composite motifs constructed by gluing together distant parts of the original sequence, it is more difficult than monad motifs. Algorithm on composite motifs models cannot easily incorporate noise tolerance in the way that monad motif models can [18, 19, 54]. The following subsections will present biological data mining, transactional sequence pattern and biosequence mining comparison, and SPM in biodata:

ISSN 2277-3061

Biological Data Mining

Bioinformatics aim to understand the detailed mechanism of the cell. Bioinformatics is the science of managing, storing, extracting, analyzing, interpreting, and utilizing information from biological data such as sequences, molecules, pathways, etc. [28, 63]. As shown in the Figure 1, Bioinformatics involve the use of multi disciplines. The huge amount of data involved in these research fields makes the usage of data mining techniques very promising. These techniques, starting from many sources, such as the results of high throughput experiments or clinical records, aims at disclosing previously unknown knowledge and relationships [17]. While tremendous progress has been made over the years, many of the fundamental problems in Bioinformatics, such as protein structure prediction, gene-environment interaction, and regulatory pathway mapping, are still open[63].

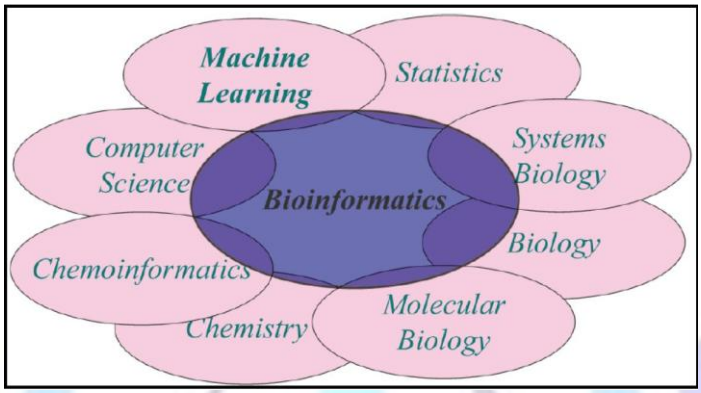


Figure 1 Bioinformatics disciplines [37, 57]

Advances in technology such as microarrays have launched the subfields of genomics and proteomics to study the genes, proteins, and the regulatory gene expression circuitry inside the cell. What characterizes the state of the field is the flood of data that exists today or that is anticipated in the future; data that need to be mined to help unlock the secrets of the cell. The gaps between data collection and knowledge discovery have collectively created exciting opportunities for data mining researchers. Data mining will play an essential role in understanding these rapidly expanding sources of biological data, fundamental problems and development of novel therapeutic/diagnostic solutions in post-genome medicine. New data mining techniques are needed to analyze, manage and discover sequence, structure and functional patterns/models from large sequence and structural DB's, as well as for structure prediction, gene finding, gene expression analysis, biochemical pathway mining, biomedical literature mining, drug design and other emerging problems in genomics and proteomics [9, 63].

Pattern Mining in Transaction Sequences and Biosequences

A transaction sequence primarily motivated in market-basket analysis, it might be any purchase sequence, a web link click stream, etc. The focus of those works is on the scalability on large DB's. Biosequences typically have a small alphabet, a long length, and patterns containing gaps (i.e., "don't care") of arbitrary size. Mining patterns in such sequences faces a different type of explosion than in transaction sequences, this explosion affects the mining process. SPMdeveloped in data mining searches for all frequent patterns in "transaction sequences". Experiments of [58] show that classical SPM does not scale for biosequences because of the following features for biosequences[9, 27, 31]; as explained in the following:

1. *Small alphabet*, biosequences have a very small alphabet, i.e., 4 for DNA and RNA sequences and 20 for protein sequences, and many patterns occur in most sequences. In contrast, transaction sequences have a large alphabet, ranging from 1,000 to 10,000, and only a tiny fraction of items occurs in a transaction sequence.

2. *Long sequence length*, a biosequence has a typical length of a few hundreds, sometimes millions. In contrast, a transaction sequence has a typical length from 10 to 20. A long sequence (especially, with a small alphabet) often contains long patterns. The classic sequential pattern growth of one item at a time as in [1, 24, 64] requires many DB scans and high frequency of pattern matching.

3. *Gappedpatterns over long regions*,biosequence patterns have the form of X_i *….*X_k spanning over a long region, where each Xi is a short region of consecutive items, and * denotes a variable length gap corresponding to not conserved region. The presence of * implies that pattern matching is more permissible and involves the whole range in a sequence.

4. Biosequence is a subject of *mutation, insertion, and deletion*.Most classical SPM is based on the Apriori idea (a subset of a frequent item set must be frequent), Apriori idea does not work with biosequence motif due to allowing missing items within biosequences pattern [1].

Sequential Pattern Mining in Biodata

SPM is the process of extracting frequent sequential patterns from sequential events or transactions which occurred in a specific order. SPM as well as association rule mining, classification and clustering are the most important data mining techniques. Sequential patterns are similar to association rules and the main difference between them is that sequential

patterns indicate the correlation among transactions in a certain order while association rules represent intra transaction relationships. SPM is applicable in a wide range of applications such as the analysis of customer purchase behavior, web access patterns, disease treatments, DNA sequences, and so on [2, 32]. Pattern mining is widely used in the Bioinformatics domain for:

1. Pattern mining for biosequences
2. Patterns in DNA sequences
3. Patterns in genes for predicting gene organization rules
4. Patterns for predicting protein sequence function
5. Patterns for analysis of gene expression data
6. Patterns for protein fold recognition
7. Patterns for protein family detection

BIO SEQUENCESDB

The strings of DNA/RNA molecules are called nucleotide sequences, and each element in such a sequence is called a base. Similarly, the strings of protein molecules are called protein sequences, and each element in such a sequence is an amino-acid (residue). Collectively nucleotide and protein sequences, are called biosequences[6]. A brief description about some biosequence DB is presented below:

Saccharomyces Cerevisiae Genome DB

S. cerevisiae was the first eukaryotic genome that was completely sequenced. The genome sequence was released in the public domain on April 24, 1996. Since then, regular updates have been maintained at the Saccharomyces genome DB. Another important S. cerevisiae DB is maintained by the Munich Information Center for Protein Sequences (MIPS). The genome is composed of about 12,156,677 bp and 6,275 genes, compactly organized on 16 chromosomes. Only about 5,800 of these are believed to be true functional genes. Yeast is estimated to share about 23% of its genome with that of humans [45, 51].

Arabidopsis Thaliana

A. thaliana is a small flowering plant native to Europe, Asia, and northwestern Africa. It is a spring annual with a relatively short life cycle, usually growing to 20–25 cm tall. Arabidopsis is a popular model organism in plant biology and genetics. It is used for studying plant sciences, including genetics, population genetics, and plant development. It plays the role in plant biology that mice and fruit flies (Drosophila) play in animal biology. Although A. thaliana has little direct significance for agriculture, it has several traits that make it a useful model for understanding the genetic, cellular, and molecular biology of flowering plants. The small size of its genome makes A. thaliana useful for genetic mapping and sequencing; it has about 157 Mbp and five chromosomes. It was the first plant genome to be sequenced, completed in 2000. The most up-to-date version of the A. thaliana genome is maintained by the Arabidopsis Information Resource (TAIR). Much work has been done to assign functions to its 27,000 genes and the 35,000 proteins they encode. Also DB's, such as ATHAMAP, offer information about the chromosomal positions of genes of interest and possible location of their TFs and TFBS[45, 52, 53].

Homo Sapience

The human genome contains 3164.7 million chemical nucleotide bases (A, C, T, and G). The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene (2.4 million bases). Almost all (99.9%) nucleotide bases are exactly the same in all people. The functions are unknown for over 50% of discovered genes. Less than 2% of the genome codes for proteins. Repeated sequences that do not code for proteins ("junk DNA") make up at least 50% of the human genome. Repetitive sequences are thought to have no direct functions, but they shed light on chromosome structure and dynamics. The human genomes gene-dense are predominantly composed of the DNA building blocks G and C. In contrast, the gene-poor are rich in the DNA building blocks A and T. GC- and AT-rich regions usually can be seen through a microscope as light and dark bands on chromosomes. Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between. Chromosome 1 contains most of the genes (2968), and the Y chromosome has the fewest (231). Human genomes DB is available and updated freely at GenBank[21] and [22].

Growth of GenBank

The consequent explosion in the availability of raw genomic data is well described by the exponential growth of the sequences deposited in public DB's such as GenBank[41]. **Error! Reference source not found.** and Figure 2 show the top 20 organisms in the latest GenBank release, their genome sizes, bases in GenBank, and number of entries [14].

One of the greatest impacts of having the sequence may be in enabling an entirely new approach to biological research. In the past, researchers studied one or a few genes at a time. With whole-genome sequences and new high-throughput technologies, they can approach questions systematically and on a grand scale. They can study all the genes in a genome, for example, how tens of thousands of genes and proteins work together in interconnected networks to orchestrate the chemistry of life.

The avalanche of genome data grows on a daily basis. The new challenge will be to use this vast reservoir of data to explore how DNA and proteins work with each other and the environment to create complex dynamic living systems. Systematic studies of function on a grand scale-functional genomics-will be the focus of biological explorations in this century and beyond [34, 61].

Table 1. GenBank genome sizes

Species	Genome size	Bases	Entries
Homo sapiens	3,400,000,000	13,669,851,495	12,838,795
Mus musculus	3,454,200,000	8,445,993,792	7,347,636
Rattus norvegicus	2,900,000,000	6,284,206,670	1,997,976
Bos taurus	3,651,500,000	5,319,815,212	2,135,747
Zea mays	5,000,000,000	5,007,807,286	3,870,406
Sus scrofa	3,108,700,000	4,229,790,475	2,556,492
Danio rerio	1,900,000,000	3,074,615,557	1,695,362
Strongylocentrotus purpur	900,000,000	1,352,840,985	228,153
Nicotiana tabacum	900,000,000	1,184,330,809	1,752,654
Oryza sativa Japonica Gro	900,000,000	1,176,024,629	1,217,983
Xenopus (Silurana) tropic	900,000,000	1,146,732,476	1,423,046
Drosophila melanogaster	180,000,000	1,038,512,618	1,202,127
Pan troglodytes	3,577,500,000	997,816,950	213,217
Arabidopsis thaliana	100,000,000	950,139,115	2,240,601
Canis lupus familiaris	100,000,000	931,176,470	1,434,100
Vitis vinifera	100,000,000	910,760,908	655,658
Gallus gallus	1,200,000,000	884,489,747	806,871
Glycine max	1,115,000,000	846,429,180	1,828,912
Macaca mulatta	3,543,000,000	808,403,289	78,410
Ciona intestinalis	200,000,000	748,153,905	1,216,132
Total		106,533,156,756	108,431,692

MOTIFS DISCOVERY ALGORITHMS

Automatizing motif discovery has become a very active research area [16]. The aim is the extraction of all kinds of biological “meaning" of these sequences. At present the main bottleneck to progress in molecular biology is the analysis of data, and not the acquisition of sequence data. The problem is how automatically discover motifs, the domain will be protein or nucleotide sequences, i.e., strings over an alphabet ∑, where ∑ contains 20 (4) different symbols when protein (nucleotide) sequences are analyzed [29].

The problem of motif discovery has been tackled extensively over the past two decades [31]. More than a hundred methods have already been proposed, and new methods are published nearly every month. There is a large diversity in the algorithms and models used, and the field has not yet reached agreement on the optimal approach. Most methods search for short, statistically overrepresented patterns in a set of sequences believed to be enriched in binding sites for particular TFs.

In spite of high availability of algorithms for motif discovery, the researcher could not find the perfect one among them due to the high complexity in the field (motif types, long sequence and pattern, gap, mutation) and different models to handle the issue. These algorithms differ in how motifs are defined and modeled. Each approach looks at a different facet of motifs. No single model or technique can identify all possible motifs. In the following, two general types of motifs discovery algorithms are discussed monad and composite motif discovery:

Monad Motif Discovery Algorithms

Planted motif search algorithms PMS0, PMS1, PMS2, SMP3, SMP4, and PMSPrune are designed for monad motif. They take as input *n* sequences of length *m* each and two integers *l* and *d*. The problem is to identify a string *M* of length *l* such that *M* occurs in each of the *n* sequences with a hamming distance (the number of mismatches between two strings of equal length) of at most d. For example, if the input sequences are GCGCGAT, CACGTGA, and CGGTGCC; *l* = 3 and *d* = 1, then GGT is a motif of interest. These algorithms attempt to find sequential patterns based on specifying motif related information, such as length of each sequence and length of motifs. They currently suffer from limitations of sequence length and input file size, and they need a long time of execution exceeding months or years [9, 49, 59]. For example,

DNA monad motif discovery tools need information about data such as DNA sequences in FASTA format, and the number of sequences must be between 5 and 500, the length of each sequence must be between 15 and 1000 DNA letters, motif length up to 23 characters may require days to process. Namely, the currently best known algorithms for PMS and edit motif search (on a PC) are expected to take more than five years for patterns of length 31[48].

Composite Motif Discovery Algorithms

In an attempt to make facilities to control the problems related to motif mining, some algorithms deal with the idea of searching and extraction of motif based on providing motifs template such as SMOTIF1, SMOTIF2, SMaRTFinder, andEXMOTIF, which made good progress in reducing the time and space required to perform motifs mining. Some motifs mining algorithms allow certain constraints. Constraints which limit the maximum gap between two items in the subsequence make it possible to use these algorithms to mine for contiguous patterns. FLAME does not target the composite motif problem, but can be used as a building block for composite motif mining [18]. The diversity of approaches to composite motif discovery is even greater than that for monad motif discovery, and methods vary widely in what they expect as input and what they provide as output. On the other hand, some algorithms demand as input a collection of already known unit sites to serve as training data. The known positive sites are used along with negative sequence examples to build a model representation which can then be compared to new sequences in order to identify novel module instances [30, 55, 60].

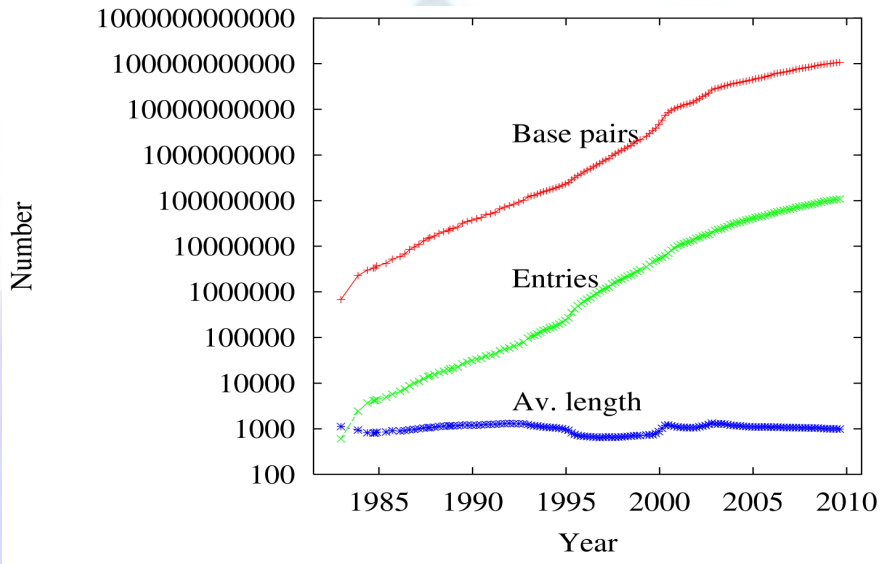


Figure 2Growing GenBank DB

MOTIFS DISCOVERYREPRESENTATION AND NEW TRENDS

Most of the motif discovery algorithms look for pre-specified patterns [33]. Some basic concepts related to motif are presented in the following:

Sequence File Format

A wide used sequence file format is FASTA, FASTA stand of Fast Alignment. Various programs require FASTA format as input file, it may contain protein or DNA sequences. The format is very simple. Every entry consists of a sequence identifier (ID), an optional comment (COMMENT), and a sequence (SEQUENCE). The format looks like this:

>ID COMMENT

SEQUENCE

The special character ">" marks the beginning of a new sequence. The ">" character is followed immediately by the sequence identifier. The rest of that line is occupied by the optional comment. Subsequent lines contain the sequence itself. Rules about representing sequences in FASTA format include:

1. Upper case and lower case doesn't matter.

2. White space (spaces and new lines) within the sequence are ignored.

3. Characters should be from a valid alphabet.

Here is an example of a sequence in FASTA format:

>ICYA_MANSE

GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLPLENENQGKCTIAEYKY

ISSN 2277-3061

Motif Representation

A critical step of the process of motif discovery is the choice of an appropriate structure to model the motifs [44]. This choice is a trade-off between the expressiveness of the model to describe particular biological properties, and the efficiency of the algorithms that can be applied when that model is chosen [55]. Arguably the most important distinction between motif discovery tools is the model that is used. A motif can be represented by two popular models: string representation, consensus or pattern and matrix representation (Position Frequency Matrix (PFM), Position Weight Matrix (PWM) or profile). Fig 3 displays an example of these models. The consensus sequence gives the most frequent nucleotide in each position. To allow for degeneracy, the characters that are used to describe a motif can be extended from {A, C, G, T} to IUPAC characters, e.g., "TATRNT" is a consensus where "R" stands for a purine (A or G) and "N" stands for a base of any type. Table 2 shows conversion to IUPAC. The PFM represents the frequencies of each base type at each position. The PWM computes a log-ratio between observed frequencies in the frequency matrix and base occurrence frequencies in random DNA (background frequency).In the PWM, motifs of length *l* are represented by size 4 × *l* with the four entries in the j^{th} column of the matrix [11][12][66].

Table 2 The IUPAC alphabet

Bases	A	C	G	T	U	A,G	C,T	G,T	A,C	G,C	A,T	C,G,T	A,G,T	A,C,T	A,C,G	A,C,G,T
Symbol	A	C	G	T	U	R	Y	K	M	S	W	B	D	H	V	N

Nucleotides Positions Interdependency

Motif representation models, the string and the matrix share an important common weakness: they assume the occurrence of each nucleotide at a particular position of a binding site is independent of the occurrence of nucleotides at other positions. Thus, motif representations cannot model biological issues well because they fail to capture nucleotide interdependence. It has been pointed out by many researchers that the nucleotides of the DNA binding site cannot be treated independently, e.g. the binding sites of zinc finger in proteins and the TFCSRE, which activates the gluconeogenic structural genes, can bind to the following binding sites:

CGGA**T**G**A**ATGG

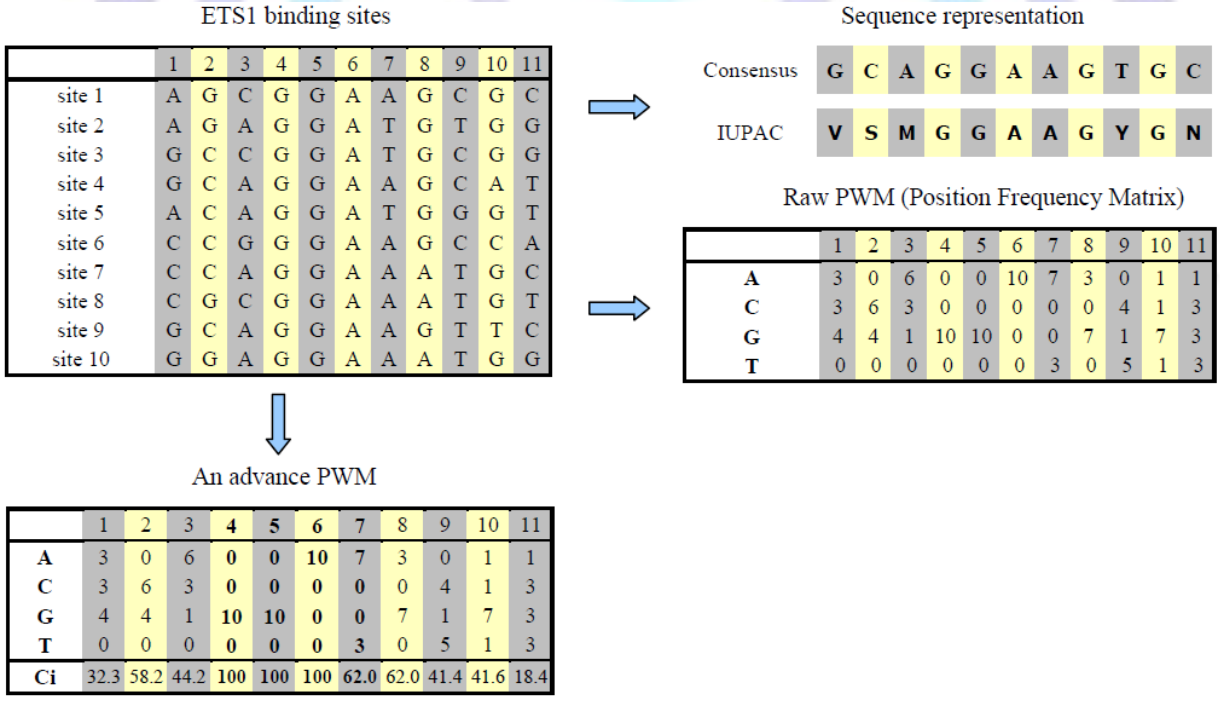
CGGA**T**G**A**ATGG

CGGA**T**G**A**AAGG

CGGA**C**G**G**ATGG

CGGA**C**G**G**ATGG

Note that there is a dependence between the fifth and the seventh symbols [11, 12, 28].



	1	2	3	4	5	6	7	8	9	10	11
site 1	A	G	C	G	G	A	A	G	C	G	C
site 2	A	G	A	G	G	A	T	G	T	G	G
site 3	G	C	C	G	G	A	T	G	C	G	G
site 4	G	C	A	G	G	A	A	G	C	A	T
site 5	A	C	A	G	G	A	T	G	G	G	T
site 6	C	C	G	G	G	A	A	G	C	C	A
site 7	C	C	A	G	G	A	A	A	T	G	C
site 8	C	G	C	G	G	A	A	A	T	G	T
site 9	G	C	A	G	G	A	A	G	T	T	C
site 10	G	G	A	G	G	A	A	A	T	G	G

	1	2	3	4	5	6	7	8	9	10	11
A	3	0	6	0	0	10	7	3	0	1	1
C	3	6	3	0	0	0	0	0	4	1	3
G	4	4	1	10	10	0	0	7	1	7	3
T	0	0	0	0	0	0	3	0	5	1	3

	1	2	3	4	5	6	7	8	9	10	11
A	3	0	6	0	0	10	7	3	0	1	1
C	3	6	3	0	0	0	0	0	4	1	3
G	4	4	1	10	10	0	0	7	1	7	3
T	0	0	0	0	0	0	3	0	5	1	3
Ci	32.3	58.2	44.2	100	100	100	62.0	62.0	41.4	41.6	18.4

Fig 3: Motifs representing forms [43]

Probability Analysis

In Bioinformatics, two models have been exhaustively used to generate sequence according to:

First, the Bernoulli Model, it is assumed that symbols of a sequence are generating according to an independent identically distributed process; hereunder, there is no dependency between the probability distribution of the symbols, but this argument is not entirely true, since sequences are believed to be biologically related.

Second, hidden Markov model (relays on basic Markov process), It's a simplified state of reality because it states that the probability of an event is only dependent on the event that occurred in the previous time step, and is not affected by events that happened two or more steps previously. Most events in the real world do depend on what happened two or more steps in the past.

Both models used assumptions not entirely true, but they simplify the problem and give results [39].

Naturalistic Method for Bioinformatics Research

This method proposes shifting the direction of researches in Bioinformatics to rely more on real biodata to deduce knowledge. Itavoids assumption-driven model that restrains the researcher to see the real picture. This method enables the researchers to dive further into the data to understand biodata properties, ground their research on a meaningful theory with a meaningful purpose, seeks to discover and describe biodata properties, configure arguments to explain properties of biodata, and they all theorize about how a structure of biodata can be used to deduce their features. In-depth studies of biodata structure gain more understanding of biodata. The goal of the method is recognize biodata reality and comprehend its nature. It selects and uses analytical techniques to gain maximum meaning of biodata and processes. It emphasizes on discovering biodata characteristics by analyzing the real data nature to reflect its de facto and to be as far as possible from Bioinformatics theoretical assumptions. Characteristics and properties of any object form corner stone and powerful method to understand the object. Therefore, this method depends on discovering properties of biodata. Simple example of applying this method in DNA motif discovery, DNA shows de facto properties such as small alphabet, long sequence, containing gaps, and mutation. But we know that DNA is full of information, therefore, they have more properties. Naturalistic method is calling to concentrate more on biodata in order to discover hidden properties. We expect following naturalistic method will increase our understanding of biodata.

The motivation and justification factors presented in [47] lead to preferring naturalistic method research for Bioinformatics, because naturalistic method depends on real data. The method empowers Bioinformatics techniques to handle the true properties and reducing assumptions for un-modeled or uncover biodata phenomena. The empowerment comes from recognizing and understanding biodata properties and processes.

CONCLUSION AND FUTURE WORK

This paper presents an overview of some basics of Bioinformatics and SPM in Bioinformatics, namely motif discovery. Although motifs discovery involves many challenges; its benefits and important applications continuously motivated researchers. It has been shown some algorithms of motif discovery. Limitations of current algorithms and motif model encourage future research to investigate the structural properties of biosequences. Naturalistic method empowers Bioinformatics techniques to handle the true properties and reducing assumptions for un-modeled or uncover biodata phenomena.Therefore, the next paper focuses on finding new properties and characteristics of motifs and biosequence.

BIBLIOGRAPHY

[1] Agrawal, R. and Srikant, R. 1995. Mining Sequential Patterns. *Proceedings of the Eleventh International Conference on Data Engineering (1995)*. 41, 1 (1995), 3–14.

[2] Alberto, A. et al. 2011. Efficient algorithms for the discovery of gapped factors. *Algorithms for Molecular Biology*. 6, 1 (2011), 5.

[3] Arlindo, O. 2011. GRISOTTO: A greedy approach to improve combinatorial algorithms for motif discovery with prior knowledge. *Algorithms for Molecular Biology*. 6, 1 (Apr. 2011), 13.

[4] Bajcsy, P. et al. 2005. Survey of biodata analysis from a data mining perspective. *Data Mining in Bioinformatics*. (2005).

[5] Boer, C. de and Hughes, T. 2012. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Research*. 40, November 2011 (2012), 169–179.

[6] Brazma, A. and Jonassen, I. 2009. Approaches to the automatic discovery of patterns in biosequences. *Journal of computational biology a journal of computational molecular cell biology (2009)*. 5, 2 (2009), 279–305.

[7] Carvalho, A.M. and Freitas, A.T. 2006. An Efficient Algorithm for the Identification of Structured Motifs in DNA Promoter Sequences. *IEEEACM Transactions on Computational Biology and Bioinformatics (2006)*. 3, 2 (2006), 126–140.

[8] Chen, C.Y. and Tsai, H.K. 2008. Discovering gapped binding sites of yeast transcription factors. *Proceedings of the National Academy of Sciences*. 105, 7 (2008), 2527.

[9] Chen, Z. and Wang, L. 2011. A Fast Exact Algorithm for the Closest Substring Problem and Its Application to the Planted (L, d)-Motif Model. *rnc.r.dendai.ac.jp*. (2011), 0–19.

[10] Chen-Ming, H. et al. 2011. WildSpan: mining structured motifs from protein sequences. *Algorithms for Molecular Biology*. 6, 1 (2011), 6.

[11] Chin, F. and Leung, H. 2008. Optimal algorithm for finding dna motifs with nucleotide adjacent dependency. *Proceedings of APBC*. (2008), 343–352.

[12] Chin, F. and Leung, H.C.M. 2008. DNA Motif Representation with Nucleotide Dependency. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*. 5, 1 (2008), 110–9.

[13] D'haeseleer, P. 2006. What are DNA Sequence Motifs? *Nature biotechnology*. 24, 4 (2006), 423–425.

[14] DOGS - Database Of Genome Sizes: *http://www.cbs.dtu.dk/databases/DOGS/GBgrowth.php*. Accessed: 2012-06-13.

[15] Fan, X. et al. 2010. The EM Algorithm and the Rise of Computational Biology. *Statistical Science*. 25, 4 (2010), 476–491.

[16] Fassetti, F. et al. 2008. Mining Loosely Structured Motifs from Biological Data. *IEEE Transactions on Knowledge and Data Engineering (2008)*. 20, 11 (2008), 1472–1489.

[17] Fiori, A. and Baralis, E. 2010. *Extraction of Biological Knowledge by Means of Data Mining Techniques*. Politecnico di Torino.

[18] Floratou, A. et al. 2010. Efficient and Accurate Discovery of Patterns in Sequence Datasets. *Data Engineering (ICDE)*. (2010).

[19] Floratou, A. et al. 2010. Finding Hidden Patterns in Sequences. *Sciences-New York*. (2010).

[20] Frith, M. et al. 2008. Discovering Sequence Motifs with Arbitrary Insertions and Deletions. *PLoS computational Biology*. 4, 4 (Apr. 2008), e1000071.

[21] GenBank human genome: *ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/Assembled_chromosomes/seq/*. Accessed: 2011-05-12.

[22] GoldenPath of currentGenomes -Homo_sapiens chromosomes: *ftp://hgdownload.cse.ucsc.edu/goldenPath/currentGenomes/Homo_sapiens/chromosomes/*. Accessed: 2011-04-05.

[23] Grant, C. et al. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. (2011), 5–6.

[24] Han, J. et al. 2001. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. *Proceedings of the 17th …*. (2001), 215–224.

[25] Huang, C.-W. et al. 2011. An Improved Heuristic Algorithm for Finding Motif Signals in DNA Sequences. *Computational Biology and …*. 8, 4 (2011), 959–75.

[26] Human Genome Project Science: *http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml#draft*. Accessed: 2012-06-13.

[27] Ichinose, N. et al. 2012. Large-scale motif discovery using DNA Gray code and equiprobable oligomers. *Bioinformatics*. 28, 1 (2012), 25–31.

[28] Ji, H. and Wong, W.H. 2006. Computational biology: toward deciphering gene regulatory information in mammalian genomes. *Biometrics*. 62, 3 (2006), 645–663.

[29] Jonassen, I. 1996. Methods for finding motifs in sets of related biosequences. *Methods*. (1996).

[30] Klepper, K. et al. 2008. Assessment of Composite Motif Discovery Methods. *BMC bioinformatics*. 9, 1 (2008), 123.

[31] Kuksa, P.P. and Pavlovic, V. 2010. Efficient motif finding algorithms for large-alphabet inputs. *BMC bioinformatics*. 11 Suppl 8, 1471-2105 (Jan. 2010), S1.

[32] Kumar, P. et al. 2012. *Pattern Discovery Using Sequence Data Mining: Applications and Studies*. IGI Global.

[33] Kumar, S. 2004. *Finding Patterns in Sequences: Comparison of Motif Extraction, Dynamic Time Warping, and Hidden Markov Model Approaches*. University of Illinois.

[34] Lander, E. 2011. Initial impact of the sequencing of the human genome. *Nature*. 470, 7333 (Feb. 2011), 187–97.

[35] Lee, N.K. and Wang, D. 2011. SOMEA: self-organizing map based extraction algorithm for DNA motif identification with heterogeneous model. *BMC bioinformatics*. 12 Suppl 1, Suppl 1 (Jan. 2011), S16.

[36] Leslie, C. 2011. High-resolution sequence and chromatin signatures predict transcription factor binding in the human genome. *Cancer*. (2011), 61284–61284.

[37] Machine Learning in Bioinformatics: *http://www.slideshare.net/sherry89/machine-learning-in-bioinformatics-6250547*. Accessed: 2012-09-08.

[38] Marschall, T. and Rahmann, S. 2009. Efficient exact motif discovery. *Bioinformatics*. 25, 12 (Jun. 2009), i356–64.

[39] Masseglia, F. et al. 2008. *Successes and New Directions in Data Mining*. Information Science Reference.

[40] Medina-Rivera, A. et al. 2011. Theoretical and Empirical Quality Assessment of Transcription Factor-binding motifs. *Nucleic Acids Research*. 39, 3 (2011), 808–824.

[41] Morgante, M. and Policriti, A. 2005. Structured Motifs Search. *Journal of Computational Biology*. 12, 8 (Oct. 2005), 1065–82.

[42] Nair, S. et al. 2012. Motif mining: an assessment and perspective for amyloid fibril prediction tool. *Bioinformation*. 8, 2 (2012).

[43] Nguyen, T.T. and Androulakis, I.P. 2009. Recent Advances in the Computational Discovery of Transcription Factor Binding Sites. *Algorithms*. 2, 1 (Mar. 2009), 582–605.

[44] Pizzi, C. 2011. Motif Discovery with Compact Approaches-Design and Applications. *intechopen.com*. (2011).

[45] Platt, A. et al. 2010. The scale of population structure in Arabidopsis thaliana. *PLoS Genetics*. 6, 2 (Feb. 2010), e1000843.

[46] Pradhan, M. 2008. *Motif Discovery in Biological Sequences*. San Jose State University.

[47] Qader, N.N. and Al-khafaji, H.K. 2014. Motivation and Justification of Naturalistic Method for Bioinformatics Research. *Journal of Emerging Trends in Computing and Information Sciences*. 5, 2 (2014), 80–87.

[48] Rajasekaran, S. 2006. Algorithms For Motif Search. *Computational Biology*. (2006).

[49] Rajasekaran, S. and Dinh, H. 2011. A Speedup Technique for (l, d)-Motif Finding Algorithms. *BMC research notes*. 4, 1 (Jan. 2011), 54.

[50] Reid, J.E. et al. 2010. Variable Structure Motifs for Transcription Factor Binding Sites. *BMC genomics*. 11, 1 (Jan. 2010), 30.

[51] Saccharomyces Genome Database: *http://www.yeastgenome.org/*. Accessed: 2012-06-13.

[52] Sundar, A.S. et al. 2008. STIF: Identification of stress-upregulated transcription factor binding sites in Arabidopsis thaliana. *Bioinformation*. 2, 10 (2008), 431.

[53] TAIR - Genome Annotation:: *http://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/annotation_data.jsp*. Accessed: 2012-06-13.

[54] Tata, S. et al. 2006. *Declarative Querying for Biological Sequences*. IEEE.

[55] Tompa, M. et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology (2005)*. 23, 1 (2005), 137–144.

[56] Wang, D. 2009. GAPK: Genetic algorithms with prior knowledge for motif discovery in DNA sequences. *2009 IEEE Congress on Evolutionary Computation* (May. 2009), 277–284.

[57] Wang, J.T.L. et al. 2005. *Data Mining in Bioinformatics*. Springer.

[58] Wang, K. et al. 2004. Scalable Sequential Pattern Mining for Biological Sequences. *Proceedings of the Thirteenth ACM conference on Information and knowledge management CIKM 04 (2004)*. (2004), 178–187.

[59] Wang, X. et al. 2012. An Improved Immune Genetic Algorithm for Weak Signal Motif Detecting Problems. *International Journal of Computer Applications (0975 – 8887)*. 43, 15 (Apr. 2012), 23–27.

[60] Wang, Y. et al. 2011. dMotifGreedy: a novel tool for de novo discovery of DNA motifs with enhanced power of reporting distinct motifs. *Arxiv preprint arXiv:1102.4015*. (2011).

[61] Weng, J. et al. 2005. Functional analysis and comparative genomics of expressed sequence tags from the lycophyte Selaginella moellendorffii. *Bmc Genomics*. 6, (Jan. 2005), 85.

[62] Yamashita, R. and Sugano, S. 2012. DBTSS: DataBase of Transcriptional Start Sites progress report in 2012. *Nucleic Acids Research*. 40, November 2011 (2012), 150–154.

[63] Zaki, M. et al. 2011. 10th International Workshop on Data Mining in Bioinformatics (BIOKDD 2011). *Electrical Engineering* (2011).

[64] Zaki, M.J. 2001. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*. 42, 1 (2001), 31–60.

[65] Zhang, Y. and Zaki, M. 2006. EXMOTIF: efficient structured motif extraction. *Algorithms for Molecular Biology*. 18, (2006).

[66] Zhang, Y. and Zaki, M. 2006. SMOTIF: efficient structured pattern and profile motif search. *Algorithms for Molecular Biology*. 1, 1 (Jan. 2006), 22.

[67] Zubi, Z.S. and Emsaed, M.A. 2010. Sequence Mining in DNA Chips Data for Diagnosing Cancer Patients. *Proceedings of the 10th WSEAS international conference on Applied computer science (2010)*. 4, 4 (2010), 139–151.

Author' biography with Photo

Nooruldeen Nasih Qader received B.Sc. Engineering degrees in Engineering college from University of Baghdad in 1993. Thereafter he received MSc and PhD in Computer Science from University of Sulaimani in 2oo7 and 2013 respectivily. He is working in as researcher and professor in University of Sulaimani. These days he is working as dean of Sciences and Technology in Univrsity of Human Development.