ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

FREQUENT CONTIGUOUS PATTERN (FCP) MINING IN GENOMIC SEQUENCE ANALYSIS AND PATTERN DISCOVERY

***S.Rajasekaran[1], L.Arockiam[2]**
[1] R & D Centre, Bharathiar University, Coimbatore, India
[2] Department of Computer Science, St. Joseph College, Trichy, India
***Corresponding author: E.Mail: sraja2911@gmail.com**

ABSTRACT

In Biology, all living organisms are related by evolution. This implies that there exist more similarities in nucleotides and protein sequences of spices. Sequence similarity helps to determine relative position of multiple spices in an evolution tree called phylogenetic tree. A collection of biological sequences is called as Sequence Database (SDB). Frequent Patterns are patterns that are repeatedly occurring in a database. The adjacent repeated patterns are called as Frequent Contiguous Patterns (FCP). As, biological sequences normally consist of millions of amino acid bases, the efficient and accurate discovery of long frequent contiguous sequences with a minimal time is a really challenging task to data scientists and life scientists. Various genomic sequence analysis and pattern discovery tools development such as findings of Repetitive DNA in Human Genome, Cis-regulatory modules (CRM) findings, Sequence motif, Transcription Factor, Promoter of gene and Single Nucleotide Polymorphism (SNP) etc require the widely spread short length repeated contiguous sequences i.e. FCP in genomic sequences. So FCP identification becomes one of the primary process in these genomic sequence analysis and pattern discovery tools development. FCP identification is the primary process in DNA-Drug design also. In this paper, we discuss role of FCP identification in these genomic sequence analysis tools development. Also, we discuss a computational method called Positional_nFCP for FCP identification.

Key words: Frequent Contiguous Pattern (FCP), Repetitive DNA, Cis-regulatory modules (CRM), Transcription Factor, Promoter of gene, Single Nucleotide Polymorphism (SNP)

1. INTRODUCTION

In Biology, all living organisms are related by evolution. This implies that there exist more similarities in nucleotides and protein sequences of spices. Sequence similarity helps to determine relative position of multiple spices in an evolution tree called phylogenetic tree. An alignment is a process of identifying similar sequences between two or more biological sequences and it lines up sequences to achieve highest level of similarities between the sequences. Highest degree of similarity is known as homology. Thus homology is an important property which can be obtained from sequence alignment process and so homology helps to place the spices in the phylogenetic tree in order.

A collection of biological sequences is called as Sequence Database (SDB). These sequences contain repeatedly occurring of fixed number of data items. For instance, a DNA_SDB contains sequence made of repeatedly occurrence of only four fixed characters A, G, T and C in any order. Frequent Patterns are patterns that are repeatedly occurring in a database. The adjacent repeated patterns are called as Frequent Contiguous Patterns (FCP).

Genomic sequence databases are huge in size and large in volume. As, biological sequences normally consist of millions of amino acid bases, the efficient and accurate discovery of long frequent contiguous sequences with a minimal time is a really challenging task to data scientists and life scientists. Computer algorithms and methods are developed to manage and analyse the huge volume of biological data. Computer alignment algorithms are important to compare and align biological sequences and discover bio-sequence patterns on SDB.

Various genomic sequence analysis and pattern discovery tools development such as findings of Repetitive DNA in Human Genome, Cis-regulatory modules (CRM) findings, Sequence motif, Transcription Factor, Promoter of gene and Single Nucleotide Polymorphism (SNP) etc require the widely spread short length repeated contiguous sequences i.e. FCP in genomic sequences. So FCP identification becomes one of the primary processes in these genomic sequence analysis and pattern discovery tools developments. Biological sequence analysis, which is the most widely used bioinformatics application, has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biomedicine. Biological sequence analysis is an important task for the biological processes like sequence alignment, sequence database searching, motif and pattern discovery, reconstruction of evolutionary relationships and genome assembly and comparison. There are many applications like knowledge–based drug design, forensic DNA analysis and agricultural bio-technology are using biological sequence analysis (Meng & Chaudhary, 2010).

In this paper, we discuss the necessity of FCP identification in various genomic sequence analysis and pattern discovery tools development. Also, we discuss a computational method called Positional_nFCP (Rajasekaran & Arockiam, 2014, Vol 9, Number 24) for FCP identification which helps to develop genomic sequence analysis tools.

This paper is organized as follows. Section 2 discusses about FCP identification method called Positional_nFCP. Section 3 discusses about FCP identification in genomic sequence analysis and pattern discovery. Section 4 discusses the conclusion.

2. POSITIONAL_NFCP METHOD – FCP IDENTIFICATION METHOD

As the DNA sequences consist of only four characters namely A, G, T and C, so the required n-length FCP should start from any one of the above 4 characters alone. Based on this principle the computational method Positional_nFCP creates 256 positional list to store starting positions of all possible patterns. As per number theory, permutations with repetitions for length-4 among 4 characters is PR (4,4) = 4^4 (= 256), hence there are 256 possible length_4 base patterns existing in the DNA sequence database.

The computational method Positional_nFCP consist of two sub methods namely positional_ptr_construct and positional_ptr_mapping. Positional_ptr_construct splits each sequence of the main DNA SDB into consecutive base patterns of length_4 and stores starting positions of base patterns into the 256 possible positional lists (positional _lists). The names of the positional lists are the possible 256 length_4 base patterns (Eg. AAAA, AAAT, TTTT). The generation of positional pointers list from the mainDB is a onetime effort. The remaining part of this computational method is based primarily on this positional information of the 256 positional lists.

It can be inferred from above two sub methods that the computational method Positional_nFCP selects minimum number of positional _lists' rather than the entire mainDB. And also it is to be noted that positional lists' have only positional information, not the sequences. Thus by transferring only the positional information of patterns to computer memory than the whole/part of main DB, drastically reduces volume of data to be handled by the computer memory. Hence the FCP identification will be much faster and accurate than the existing computational methods.

3. FCP IDENTIFICATION IN GENOMIC SEQUENCE ANALYSIS AND PATTERN DISCOVERY TOOLS DEVELOPMENT

Science is about building causal relations between natural phenomena (for instance, between a mutation in a gene and a disease). The distance between the reality and the observation (through the instrument) needs to be accounted (Reese & Guigo, 2006, 7). The primary issue in Genome Biology is calibrating computer programs to identify human genes in the sequence of the human genome.

Analysis of complete genome sequence is necessary to the following genomics functionalities:

- The complete genome sequence provides the basis for discovering all the genes that are encoded in a genome.
- The comparative genomic sequence shows the structural and regulatory elements associated with genes.
- It provides the basis to assess the molecular evolution of a species as well as the extent of its variations between individuals, populations, and other species.
- It provides a set of tools for future experimentation (Pevsner, 2009).

Humans are estimated to have approximately 20,000 protein-coding genes (collectively known as the exome), which account for only about 1.5% of DNA in the human genome. The primary goal of the ENCODE project is to determine the role of the remaining component of the genome, much of which was traditionally regarded as "junk" (i.e. DNA that is not transcribed). Approximately 90% of single-nucleotide polymorphisms in the human genome (that have been linked to various diseases by genome-wide association studies) are found outside of protein-coding regions (Maher, 2012, Vol 489, Issue 7414).

3.1. FCP Identification in Repetitive DNA findings in Human Genome

Human genomes contain a smaller proportion of protein-coding genes and large amounts of noncoding DNA. This noncoding material includes repetitive DNA, genes encoding RNAs that have regulatory functions, and introns that interrupt exons and are spliced from mature RNA transcripts. Repetitive DNA is a DNA sequence of varying length that occur in multiple copies in genome. It represents much of human genome. There are five main classes of repetitive DNA in human genome. They are:

- **Interspersed repeats** are repeats which constitute about 45% of the human genome. These repeats can be generated by elements that copy RNA intermediates (retroelements) or DNA intermediate (DNA transposons).
- **Processed pseudogenes** are genes that are not actively transcribed or translated (Harrison & Gerstein, 2002 (318)). These represent genes that were once functional, but they are defined by their lack of protein product. They can be recognized because of the presence of a stop codon or frameshift that interrupts an open reading frame.
- **Simple sequence repeats** are short sequences contain microsatellite (with a 1 to 6 base pair length) and mini-satellite (with a 12 to 500 base pair length) sequences. The density and length of microsatellites are considerably greater in human genome. In humans simple sequence repeats are of particular interest because they are highly polymorphic between individuals and thus serve as useful of genetic markers. Also, the expansion of triplet repeats such as CAG is associated with over a dozen diseases, including Huntington disease (Cummings & Zoghbi, 2000 (1)).
- **Segmental Duplications or low copy repeats** are often defined as two genomic regions sharing at least 90% of nucleotide identity over a span of one kilobase, although they sometimes consist of blocks of 200 and 300 kilobases in length (Bailey, Yavor, Massa, Trask, & Eichler, 2001, 11). These duplications occur both within and between chromosomes. The euchromatic portion of of the human genome consists of about 5.3% duplicated regions (She, et al., 2004, (431)). This includes 150 megabases. These segmental duplications may cause genes to become deleted, duplicated or inverted.
- **Blocks of Tandemly Repeated Sequences** are the short repeated sequences and the repetitions are directly adjacent to each other and they occur in primary structure of DNA. These repeats serve as an attachment in chromosomal segregations during mitotic and meiotic cell divisions.

All these repeats are short length DNA sequences. FCP identification method will help in identifying all these repeats.

3.2. FCP identification in Cis-regulatory modules (CRM) findings

Some of the functionally useful genomic short length repeated sequences are Promotors, enhancers, silencers, insulators and locus control regions. These regulatory elements are sometimes called Cis-regulatory modules (CRM).

CpG islands represent an example of a regulatory element. Promoter is a region of DNA that initiates transcription of a particular gene. Promoters are located near the transcription start sites of genes. Promoters are formed by high density of GC content. This high density GC content region > 50% is called as CpG islands. Thus CpG islands help to identify promoter regions. Motif is a short conserved region in a protein sequence. Motifs are frequently highly conserved parts of domain. Open Reading Frame (ORF) is a sequence of DNA or RNA located between start code sequence and stop code sequences. Regulatory regions are DNA base sequences that control gene expression.

In molecular biology, a CCAAT box (also sometimes abbreviated a CAAT box or CAT box) is a distinct pattern of nucleotides with GGCCAATCT consensus sequence that occur upstream by 60-100 bases to the initial transcription site. The CAAT box signals the binding site for the RNA transcription factor, and is typically accompanied by a conserved consensus sequence. The computational method Positional_nFCP combines GGCC, GCCA and ATCT pattern tables and produces the results.

Sequence-Tagged Site (STS) is a short (200 to 500 base pairs) DNA sequence that has single occurrence in the human genome and whose location and base sequences are known. It is useful for localizing and mapping the reported gene sequence data. Expressed sequence Tag (EST) is a short DNA sequence act as an identifier of gene. Transcription Factor is a protein or codon-region of a DNA that binds to the regulatory regions and helps to control gene expressions. The primary process to develop all these genomic sequence analysis tools are to find repeated contiguous short-length DNA sequence patterns, i.e. to find FCP on large DNA SDB. The computational method Positional_nFCP method can be widely used for the identification of these short length genomic sequences.

3.3. FCP identification in DNA-Drug Design Process

Protein is the functional molecule of life determined by the sequences of amino acid. The double helix structure of DNA contains wider length major groove and smaller length minor groove. The major groove contains more functional groups than the minor groove DNA binding ligands (Dervan, 1986, 232). Codon regions are the functional part of the major groove and the minor groove. These codon regions combine and form the proteins. In this process transcription factors in the DNA binding proteins interact with major groove of B-DNA (Zimmer & Wahnert, 1986, 47 (1)) plays major role in forming protein.

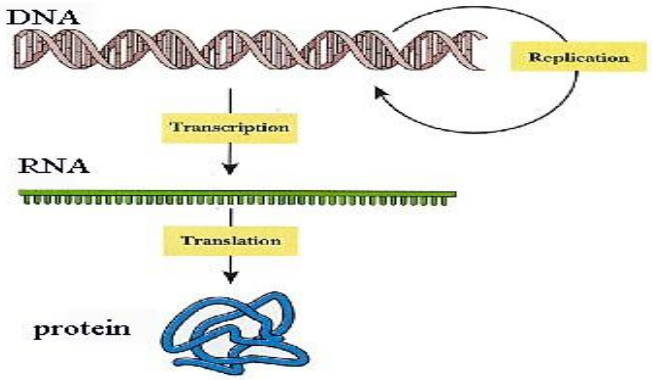

Fig 1: DNA to Protein work flow

The Fig.1 shows that DNA is replicated (copied) and transcripted to RNA. Then codon part of DNA in RNA is translated into functional protein. The mutations (errors) in transcription and translation lead to genetic disorders. Currently, a total of ~4,000 genetic disorders are known. Some of the genetic disorder mutations are:

- Point mutation: Single base substitution, is a type of mutation that causes the replacement of a single base nucleotide with another nucleotide of the genetic material, DNA or RNA.
- Deletion: A loss of part of DNA from a chromosome. It can lead to a disease or abnormality.
- Insertion: A chromosome abnormality, in which a piece of DNA is incorporated into a gene and thereby disturbs the gene normal function.
- Polygenetic disorder: While disorders are inherited such as heart disease, diabetes and some cancer, they depend on combined action of more than one gene. It is a complex hereditary pattern.
- Single Nucleotide Polymorphism (SNP): DNA sequence variation that occur when a single nucleotide (A,T,C,G) in the genome sequence is altered.

These genetic disorders can be cured by designing DNA drug, which binds to the mRNA and stops the mRNA not to translate disease causing proteins. DNA drugs are DNA sequences that match with sequences of mRNA. The DNA drug design and development process is expensive and time consuming. Computer modelling of drugs is not only for financial savings but also saving the lives of millions of people (Gibas & Beck, 2010). One of the main functions in the DNA drug design is, finding the sequence homology with mRNA sequences. The sequence homology is used in the drug design so as not to affect the important functional part of molecules. Many DNA/protein sequence analysis processes like motif finding, DNA binding sites, active sites, regulatory regions require sequence patterns matching and sequence homology/similarity.

FCP Identification is the primary process for finding sequence homology and sequence pattern matching in DNA/mRNA sequences. Thus computational method Positional_nFCP helps to design DNA-Drug.

4. CONCLUSION

In Biological SDB applications, FCP identification is useful to find Repetitive DNA, Cis-regulatory modules (CRM)**,** Motifs and commonly Repeated Patterns of Nucleotide in DNA/Protein sequence etc. Mining Bio Frequent Contiguous Patterns in the large Bio Sequences (DNA or Protein Sequences) Databases is providing solutions to the Biological functions.

- Identifying Sequence motif, a sequence pattern of nucleotides in a DNA sequence or amino acids in a protein. In genetics, a sequence motif is a nucleotide or amino acid sequence pattern that is widespread and frequently and has a biological significance.
- Finding a transcription factor, a process to identify a protein that binds to specific DNA sequences, thereby controlling the rate of transcription of genetic information from DNA to messenger RNA
- Identifying a promoter which is a region of DNA that initiates transcription of a particular gene. Promoters are located near the transcription start sites of genes.
- The polymerase chain reaction (PCR) is a technology in molecular biology used to amplify a single copy or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.
- DNA Drugs Design, a short DNA sequence that matches the sequence of mRNA that is transcribed from the mutated gene (which causes diseases)

Thus FCP identification has become an essential part of genomics, proteomics, functional genomics and bio-medical research.

REFERENCES

Bailey, J., Yavor, A., Massa, H., Trask, B., & Eichler, E. (2001, 11). Segmental duplications: Orgnization and impact within the current human genome project assembly. *Genome Res.*, 1005-1017.

Cummings, C., & Zoghbi, H. (2000 (1)). Trinucleotide repeats: Mechanisms and pathophysiology. *Annu. Rev. genomics. Hum. Genet.*, 281-328.

Dervan, P. (1986, 232). Design of sequence-specifi DNA-binding molecules. *Science*, 464-471.

Gibas, C., & Beck, P. (2010). *Developing Bioinformatics Computer Skill: An introduction to Software Tools for Biological Application.* Mumbai: Shroff.

Harrison, P., & Gerstein, M. (2002 (318)). Studying genomes through the aeons: Protein families, pseudogenes and proteome evolution. *Journal of Molecular Biology*, 1155-1174.

Maher, B. (2012, Vol 489, Issue 7414). ENCODE: The human encyclopaedia. *Nature*.

Meng, X., & Chaudhary, V. (2010). A High-Performance Heterogeneous Computing Platform for Biological Sequence Analysis. *IEEE transaction on Parallel and Distributed Systems*.

Pevsner, J. (2009). *Bioinformatics and Functional Genomics.* New Delhi: Wiley Indian Pvt Ltd.

Rajasekaran, S., & Arockiam, L. (2014, Vol 9, Number 24). Positional_nFCP: Positions based Big Data algorithm to Identify n_length Frequent Contiguous Patterns (FCP) in a Large Human Genome Sequence Database. *International Journal of Applied Engineering Research*, 23771-23780.

Reese, M. G., & Guigo, R. (2006, 7). EGASP: Introduction. *Genome Biology*.

She, X., Jiang, Z., Clark, R., Liu, G., Cheng, Z., Tuzun, E., et al. (2004, (431)). Shortgun sequence assembly and recent segmental duplications within the human genome. *Nature*, 927-930.

Zimmer, C., & Wahnert, U. (1986, 47 (1)). Nonintercalating DNA-binding ligands: specificity of the interaction of and their use as tools in biophysical, biochemical and biological investigations of the genetic material. *Prog. Biophys. Mol. Biol*, 31-112.