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

DNA-DRUG_FCP: AN EFFICIENT COMPUTATIONAL METHOD FOR DNA-DRUG DESIGN USING FREQUENTLY REPEATED PATTERNS IN A LARGE HUMAN GENOME DATABASE

***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

DNA-Drug design saves lives of millions of people by stopping mutated genes to code nonfunctional proteins. The proteins coded by mutated genes cause various genetic disorders like heart disease, hypertension, low IQ and diabetes. The function of nonfunctional proteins must be stopped to cure diseases. DNA-Drug is a DNA sequence which is made to bind with mRNA to stop the function of non-functional proteins. At the same time, the bonded/attached DNA-Drug should not affect the important part of molecules. This can be done by sequence homology. Many DNA/protein sequence analysis processes like motif finding, DNA binding sites, active sites, regulatory regions require sequence patterns matching and sequence homology/similarity. The main objective of sequence homology process is to find frequently occurred repeated contiguous patterns called FCP in the DNA/mRNA sequences. These bio-sequences (DNA/Protein sequences) data are huge in size and large in volume and having widely spread frequently repeated short length patterns. This paper proposes an efficient computational method to find DNA-Drug FCP with minimum execution time and computer memory usage for very large DNA/mRNA sequences. This method selects minimum number of positionally related corresponding patterns to search patterns instead of whole DNA/mRNA sequences. This method can be used in computational modelling like Hidden Markov Model (HMM) to predict DNA-Drugs.

Key words: DNA-Drug, DNA/mRNA sequences, Frequent Contiguous Pattern, HMM

1. INTRODUCTION

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 PB, 1986). 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 C; Wähnert U, 1986) plays major role in forming protein. The mutations in transcription and translation lead to genetic disorders. The proteins produced by the mutated genes cannot function properly which leads to many diseases like heart disease, hypertension, diabetes, cancer and low IQ. 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 process of DNA drug design and development 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 C; Beck PJ, 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. The bio-sequences (DNA/Protein sequences) data are huge in size and large in volume. These DNA/protein sequences are lengthy sequences having widely spread frequently repeated short length patterns. These frequently repeated short length patterns are called as Frequent Contiguous Patterns (FCP).

This paper proposes an efficient computational method for finding FCP for DNA-Drug design called DNA-Drug_FCP. This paper is organized as follows. Section 2 discusses the processes in DNA-Drug_FCP and the related works. Section 3 describes computational method of DNA-Drug_FCP. Section 4 discusses the results and discussion. Section 5 discusses the conclusion and future research direction.

2. PROCESSES IN DNA-DRUG_FCP AND RELATED WORKS

Processes in DNA-Drug_FCP: The objective of the proposed DNA-Drug_FCP method is to find the frequently occurring repeated contiguous patterns of length-n called FCP in DNA/mRNA sequences. If D is the sequence database contains DNA/mRNA sequences with their sequence IDs. And m is the length of each sequence in D. This process for a sample DNA sequence is illustrated in Figure 1.

Fig.1 is a sample DNA Sequences database 'D' contains five DNA sequences. ATGC is one of the length_4 DNA-Drug_FCP for the above sample DNA sequence database (DSDb).

Related works: There are lots of sequence pattern mining methods are used in the research of identifying repeated contiguous patterns. Apriori (Agrawal R; Srikant R, 1994 and Srikant R; Agrawal R, 1996) method is the base for the most of these pattern mining methods. The property of Apriori method is "All Non-empty subsets of a frequent item set must also be frequent". This method is accurate but it requires repeated scanning of the database to generate new patterns from the existing patterns. This requires lot of repeated computations which leads to high execution time to get the results.

The MacosVSpan method constructs the spanning tree containing both frequent and non-frequent patterns (Pan J; Wang P; Wang W; Shi B et al., 2005). The result is obtained by searching the spanning tree. Kang et al (Kang TH; Yoo J.S and Kim H.Y., 2007) method is better than MacosVSpan. Another method called Prefixspan (Pei J; Han J; Mortazavi-Asl B et al., 2001) is based on projected database. Segment-Pattern Index (SP-Index) (Wang K., Xu Y. and Yu J. X, 2004) method generates SP Index tree for mapping frequent segment into its corresponding Index list. High computation and memory are needed for this method to construct SP-index tree for all possible existing patterns.

Surprising contiguous pattern mining method (Rashid Md. M., Karim Md. Rezaul, Jeong B. et al., 2012), location based FCP method (Tanvee M. M., Kabeer S. J., Chowdhury T. M., Sarja A. A. and Shuvo Md., 2013) and position based Fast Contiguous FCP methods (Zerin S.F. and Jeong B.S., 2011) are all constructing spanning tree for patterns. The results are obtained by scanning the spanning tree. To improve execution time for scanning they uses different computational methods like Hashing method, binary search method, quick union and sorting methods.

As all these methods need to construct trees, tree traversal and lot of intermediate tables to produce the results which require high time and database storage space.

3. METHODS OF DNA-DRUG FCP

The DNA/mRNA sequences are made of four characters namely A, G, T/U and C. The required search patterns in these sequences must start with any one of these four characters only. If the starting character of the search pattern is matched with these sequences, then consecutive positions of both sequences can be compared and obtained the results, if the match occurs.

The computational method DNA-Drug_FCP contains two major methods namely PositionPtr_Construct and PositionPtr_Map. The first method PositionPtr_Construct splits each DNA sequences into consecutive patterns of length-4 and stores their starting positions of these base patterns into 256 possible pointers list. As per number theory, there will be only 256 possible length-4 patterns occur for any DNA/mRNA sequence. In other words permutations with repetitions for length-4 among four characters is defined as PR (4, 4) = 4^4 (= 256). The second method called PositionPtr_Map selects the corresponding length-4 patterns pointers list which matches the search patterns and produce the results. The consecutiveness is verified by the positional information stored in the 256 pointers list. This process is illustrated in Table I and Table II.

The method PositionPtr_Construct splits the DNA sequence of ID - 10 in the Table I into consecutive base patterns of length_4 as ACGT, CGTT, GTTA … TTAC. The starting positions of these patterns are (0, 1, 2… 10), which are stored at (ACGT_ptr, CGTT_ptr, GTTA_ptr……. TTAC_ptr) respectively. Table II illustrates the functionality of PositionalPtr_Construct method for the sequence in Table I.

If TTAC is the search pattern then the Position Ptr_Map selects TTAC_ptr and produce the results as sequence ID-10 contains the pattern TTAC at two places namely 3rd and 10th positions.

This DNA-Drug FCP need not to use entire DNA sequence to find the search pattern. Instead it selects only the matched part to obtain the results. Hence it reduces the execution time and disk storage widely. This is an apt method for finding patterns in large sequences like DNA/mRNA sequences.

Table.1.Example Sequences

Seq_ID	Sequence
10	Positions:0 1 2 3 4 5 6 7 8 9 10 11 12 13 A C G T T A C A C G T T A C

Table.2.Illustration of PositionPtr_Construct

Sl. No.	Consecutive Positions (i1, i2)	Position ptr_lists	Position Value (Start Position)
1	{ (i1=0, i2=3), (i1=7, i2=10) }	ACGT_ptr	0,7
2	{ (i1=1, i2=4), (i1=8, i2=11) }	CGTT_ptr	1,8
3	{ (i1=2, i2=5), (i1=9, i2=12) }	GTTA_ptr	2, 9
4	{ (i1=3, i2=6), (i1=10, i2=13) }	TTAC_ptr	3,10
5	{ (i1=4, i2=7) }	TACA_ptr	4
6	{ (i1=5, i2=8) }	ACAC_ptr	5
7	{ (i1=6, i2=9) }	CACG_ptr	6

4. RESULTS AND DISCUSSION

Human genome DNA sequences (Homo sapiens DNA Chromosome) are downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/nuccore/) for the analysis of the proposed DNA-Drug_FCP method. The selected human genome DNA database contains 35,000 sequences with each sequence of length 70.

The DNA-Drug_FCP method and test drivers are written in Java 1.7.0 and the MySQL 5.5.38-0 is used as the database. Ubuntu 12.04.01 is the base OS with Intel Core i7 CPU @ 2.00 GHz x 4 with 2 GB main memory and 25 GB hard disk.

The table III provides the execution time results of DNA-Drug_FCP method, for min_support is approximately 14,850, which is 42% of total sequences of the mainDB. Min_support means the minimum number of occurrence of pattern in the DNA/mRNA sequence database. The graphical results for the execution time and the computer memory usage of DNA-Drug_FCP method are shown in Fig.2 and Fig.3.

The execution time required for finding repeated patterns in DNA-Drug_FCP is between 170 seconds with memory usage as 32 MB for search pattern length-4 to 8 and 260 seconds with the memory usage as 49 MB for search pattern length-9 to 12.

Table.3.Execution time results of DNA-Drug_FCP method for min_support=42%

Pattern Length	Pattern	Buffer (in MB)	Time (sec)
5	ATTTA	31.80	166.50
6	CACGTA	32.60	170.20
7	TTACGAA	32.50	178.00
8	TGACTAGC	31.40	165.30
9	TGACTAGCA	48.30	258.50
10	TGACTAGCAC	49.50	263.40
11	TGACTAGCACT	48.20	257.00
12	GCACAGTCACTA	49.40	261.10

If D =
{ (1, GTG **ATGC** ACT **ATGC**),
(2, **ATGC** CTTC **ATGC**)
(3, CTGAAGT**ATGC**)
(4, ATTGTCGTG)
(5, GTGGCATTC) }, $1 \leq i \leq 5$

Figure.1. DNA-Drug_FCP process

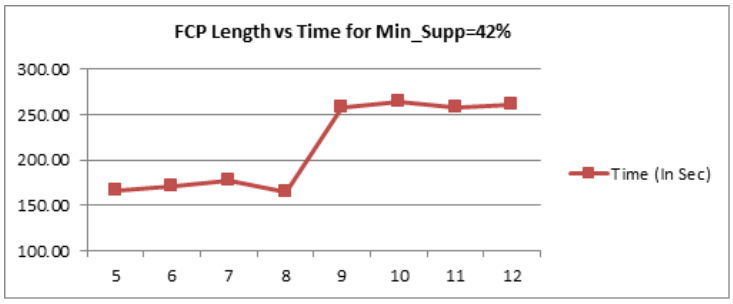

Figure.2.Execution time of DNA-Drug_FCP

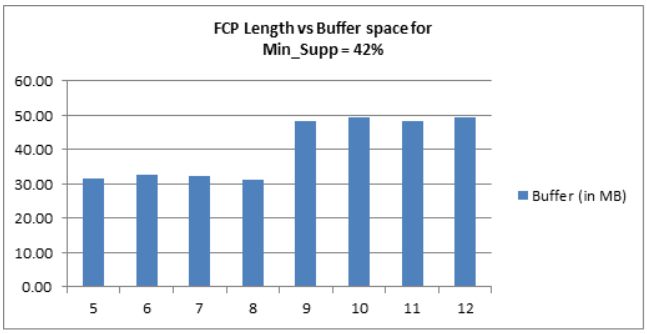

Figure.3.Computer memory usage by DNA-Drug_FCP

5. CONCLUSION AND FUTURE RESEARCH DIRECTION

DNA-Drug design saves lives of millions of people by stopping mutated genes to code nonfunctional proteins. DNA-Drug design primarily depends on sequence homology and sequence pattern matching of DNA/mRNA sequences. The method DNA-Drug_FCP proposed an efficient technique to find frequently repeated contiguous patterns called FCP. Identification of these FCPs is the base process for finding sequence homology and sequence pattern matching in DNA/mRNA sequences. The implementation of DNA-Drug_FCP shows that this method requires less execution time and computer memory for high minimum support to a very large DNA/mRNA sequences.

This DNA-Drug_FCP method can be extended to predict and design the subsequent DNA-Drug sequences as per popular saying "Nature is tinker, not an inventor" by using Hidden Markov Method (HMM).

REFERENCES

Agrawal R, Srikant R, Fast Algorithms for Mining Association Rules. Proceedings of the 20th VLDB conference, 1994, Santiago.

Dervan PB, Design of sequence-specific DNA-binding molecules, Science 232 (4749), 1986, 464–71.

Gibas C, Beck PJ, Developing Bioinformatics Computer Skill: An Introduction to Software Tools for Biological Application, Ed. Mumbai, India: Shroff, 2010.

Kang TH., Yoo J.S. and Kim H.Y., Mining frequent contiguous sequence patterns in biological sequences. In Proceedings of 7th IEEE International Conference on Bioinformatics and Bioengineering, 2007, 723-8.

Pan J, Wang P, Wang W, Shi B and Yang G, Efficient algorithms for mining maximal frequent concatenate sequences in biological datasets. In Proceedings of the Fifth International Conference on Computer And Information Technology (CTT), 2005, 98-104.

Pei J, Han J, Mortazavi-Asl B, Chen Q, Dayal U and Hsu M., PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth, ICDE, 2001.

Rashid Md. M., Karim Md. Rezaul, Jeong B. and Choi H, Efficient Mining of Interesting Patterns in Large Biological Sequences. Genomics & Informatics, 10 (1), 2012, 44-50.

Srikant R., Agrawal R, Mining sequential patterns: Generalizations and performance improvements, 5th International Conference on Extending Database Technology, 1996, Avignon, France.

Wang K., Xu Y. and Yu J. X., Scalable Sequential Pattern Mining for Biological Sequences. CIKM'04. Proceedings of the thirteenth ACM international conference on Information and knowledge management, 2004, 178-187.

Zerin S.F. and Jeong B.S., A Fast Contiguous Sequential Pattern Mining Technique in DNA Sequences Using Position Information, IETE Technical Review, 2011, 28.

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