International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Frequent Sequential Patterns (FSP) Algorithm for Finding Mutations in BRCA2 Gene

Jawahar. S, Reshmi. S, Ahamed Johnsha Ali. S

Abstract —The Sequential Pattern Mining (SPM) is a fundamental task in data mining. The SPM mines subsequences from given sequence which can be used for various analyses. This paper aims to propose an efficient method for mining frequent sequential patterns in biological data. It also includes the k-mer for decomposing the sequence according to the user defined threshold value. The input data used is breast cancer gene BRCA2 normal and mutated BRCA2 gene. The parameters used for analyses are suffix, candidate pattern and frequent pattern. The suffix value is increased for mono-,di and tri-nucleotide in mutated gene and in frequent pattern tri-nucleotide has increased nucleotide in mutated gene. So this abnormal increase in pattern may leads to cancer in the human

Keywords - Sequential Pattern Mining (SPM), k-mer, BRCA2, Suffix

I. INTRODUCTION

A subsequence in a sequence is frequent if the subsequence appears frequently for not lesser than user specified minimum support threshold (Min_Sup) [4]. In bioinformatics computing k-mers is one of the fundamental sequence analyses. The eukaryote contains 30,000 to 40,000 genes which account 2 to 5 % of the genome in the gene coding region. The bioinformatics aims for extracting information from 3 billion bases of human DNA for studying information flow and information content in biological process[1].

There are many applications of bioinformatics namely, pattern discovery, protein folding, alignment and homology, orthologs and paralogs, information retrievaland data mining from biological databases, analysis of biological sequence and pattern discovery, micro-array and differential gene expressions, gene regulatory network and many more [1].

II. The BRCA2 GENE AND SEQUENCING

The breast cancer is one of most deadly disease among women. There are various genes which cause breast cancer and one mostly affected is Breast Cancer Gene BRCA1 and BRCA2.

Manuscript published on 30 September 2019
* Correspondence Author
Jawahar. S*, Assistant Professor,Department of Computer Science and Applications,Sri Krishna Arts and Science College,Coimbatore-641008
shivamjawahar@gmail.com
Reshmi. S, Assistant Professor,Department of Computer Science and Applications,Sri Krishna Arts and Science College,Coimbatore- 641008
reshmismca@gmail.com
Ahamed Johnsha Ali. S, Assistant Professor,Department of IT and CA Sri Krishna Adithya College of Arts and ScienceCoimbatore- 641008
ahamed.doit@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

The BRCA2 gene is located in Chromosome 13 and contains about 27 exons in human. BRCA2 gene was discovered in the year 1994. mutation in exon 11 is occurs frequently which may cause cancer. There are various types of mutations insertion, deletion, substitution, frame shift and many more [2].

III. PROPOSED METHODOLOGY

In the proposed method profile based searching technique is used for biological sequences which represent relative frequency for the sequence. The k-mer identifies the distribution of nucleotides among two sequences and captures their frequency profile.In this section an algorithm Frequent Sequential Patterns (FSP) in Breast Cancer is used for mining sequential patterns. This FSP algorithm contains k-mer which is used for decomposing the biological data into mono-, di- and tri-nucleotides.

Proposed Algorithm: Frequent Sequential Patterns (FSP) in Breast Cancer

Input: BRCA2 Sequence, Minimum Support (Min_Sup) and k-mer value

Output: Frequent Sequential Patterns for user specified Min_Sup.

Step 1: Read BRCA2 sequence file

Step 2: The k-mer decomposed the input sequence (mono-,di and tri-)

Step 3: The occurrence value for A, T, G & C in the sequence are counted and corresponding subsequence are also counted.

Step 4: The Suffix, Candidate and frequent patterns are calculated according to user defined Min_Sup value

Step 5: Display the patterns

The proposed algorithm input is stored in text file with FASTA format which mines the frequent sequential patterns. The input datasets used in the process contains real life DNA sequences downloaded from National Center for Biotechnology Information [3] website.By using the advanced search technique in NCBI the normal gene can be downloaded as,

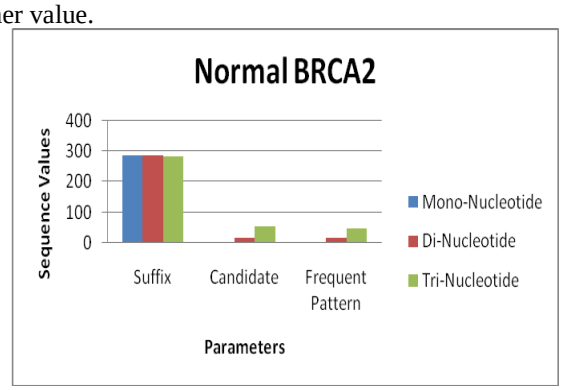
(a) Search Category=”Nucleotide”, (b) Organism =”human” and (c) All Fields=”normal BRCA2”.

(b) The Breast Cancer-2 (BRCA2) DNA data set was downloaded using advanced search technique with following parameters, (a) Search Category=”Nucleotide”, (b) Organism=”human” and (c) gene=”BRCA2 gene” (d) All Fields=” breast cancer gene”.

TABLE I. Normal BRCA2 exon 11 gene Sequence

K-mer	Suffix	Candidate	Frequent Pattern
Mono-Nucleotide	287	4	4
Di-Nucleotide	286	16	16
Tri-Nucleotide	285	56	49

The table 1 represents the k-mer for Mono-,di and tri-nucleotide sequences for normal BRCA2 gene. The suffix, candidate and frequent patterns are calculated. The frequent patterns varies according to 4^n where 'n' is the user defined k-mer value.



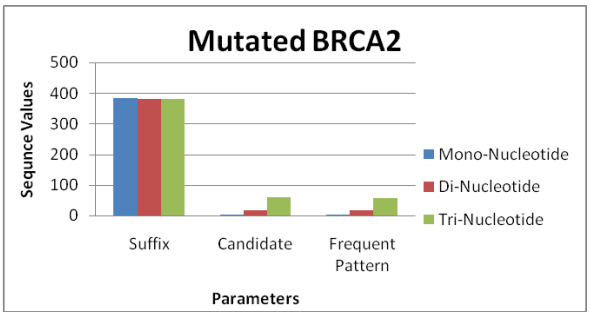
Graph 1: Normal Parameters vs Sequence value

The graph 1 represents the various parameter values for sequence in normal BRCA2 gene with mono-, di- and tri-nucleotide values.

TABLE II. BRCA2 exon 11 gene (Mutated) Sequence

K-mer	Suffix	Candidate	Frequent Pattern
Mono-Nucleotide	383	4	4
Di-Nucleotide	382	16	16
Tri-Nucleotide	381	61	56

The table 2 represents the k-mer for Mono-,di and tri-nucleotide sequences for mutated BRCA2 gene. The suffix, candidate and frequent patterns are calculated. The frequent patterns varies according to 4^n where 'n' is the user defined k-mer value.



Graph 2 Mutated BRCA2 Parameters vs Sequence value

The graph 2 represents the various parameter values for sequence in Mutated BRCA2 gene with mono-, di- and tri-nucleotide values.

IV. CONCLUSION

The frequent pattern mining is used to discover useful patterns in bio-sequences. The mutation in breast cancer gene BRCA2 is used to evaluate occurrence of breast cancer. The k-mer method is used to decompose the input DNA sequence which is more compact to analyze the sequences. These frequent sequences helps to predict and detect the disease causing mutations. The suffix value increased abnormally when compared to normal BRCA2 gene which indicates more number of insertion mutation in the gene. Also frequent pattern in Tri-nucleotide increased which may cause cancer in the human.

REFERENCES

1. Gautam B. SinghFundamentals of Bioinformatics and Computational Biology Methods and Exercises in MATLAB 2015 ISSN 2196-7326 ISSN 2196-7334 (electronic) ISBN 978-3-319-11402-6 ISBN 978-3-319-11403-3 (eBook).
2. Mehrgou A, Akouchekian M. The importance of BRCA1 and BRCA2 genes mutations in breast cancer development. Med J Islam Repub Iran 2016 Vol. 30:369.
3. National Center for Biotechnology Information (www.ncbi.nlm.nih.gov)
4. Agrawal, R., &Srikant, R. (n.d.). Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering*.doi:10.1109/icde.1995.380.