Intl. J. Bioinformatics and Biological Sci.: (V. 6 n.1, p. 35-41): June 2018

DOI: 10.30954/2319-5169.01.2018.5

An Efficient Contiguous Pattern Mining technique to Predict Mutations in Breast Cancer for DNA Data Sequences

S. Jawahar and P. Sumathi

Department of Computer Science, Government Arts College, Coimbatore-18, Tamil Nadu, India

Corresponding author: shivamjawahar@gmail.com / sumathirajes@hotmail.com

Abstract

In data mining, one of the most important tasks is sequential pattern mining (SPM). This SPM is used to mine most interesting subsequences in a set of sequences. The various real-life applications of SPM is bioinformatics, market basket analysis, web stream analysis and many more. The development of applications using data mining techniques to solve biological problems plays an important role in bioinformatics. This paper aims to propose mining of contiguous patterns in Deoxyribonucleic Acid (DNA) to identify breast cancer disease. The CSpan (Contiguous Sequential Pattern Mining) method is used to find contiguous patterns of DNA sequence database. Instead of mining all the patterns in a given sequence only contiguous patterns are mined i.e., compact patterns. The contiguous patterns with greater homogeneity are considered as biomarker to identify breast cancer disease. The patterns frequency occurrence of normal DNA is compared with mutated patterns of breast cancer gene (BRCA1) for identifying the disease. The mutation ratio is calculated to identify the level of change in the contiguous pattern between normal and mutated patterns.

Keywords: Sequential Pattern Mining (SPM), Deoxyribonucleic Acid (DNA), Contiguous Sequential Pattern Mining (CSpan), breast cancer gene (BRCA1)

Data mining is used for extracting useful information or to mine hidden information from large databases. These extracted data can be used for understanding the data or for making useful decisions. The various fundamental tasks of data mining are (i) clustering, (ii) classification, (iii) outlier analysis and (iv) pattern mining[1, 2]. Pattern Mining (PM) is used for frequent itemset mining and association rule mining. But pattern mining is not suitable for mining sequential data. To address this problem, Sequential Pattern Mining (SPM) is used which can mine interesting patterns from large databases. The SPM can be used in many fields such as bioinformatics, text analysis, e-learning, market basket analysis and many more.

Sequential Pattern mining

Sequential Pattern Mining is used to mine subsequences or frequent sequence with various user specified constraints. The SPM is used to analyze the various user access patterns. There are 3 types of SPM namely, (1) Closed sequential patterns, (2) Maximal sequential patterns and (3) contiguous sequential patterns. In closed sequential patterns, the patterns with same minimum support value are not included in the sequential set. The maximal sequential patterns are not included in other patterns or maximal patterns are not larger than closed sequential patterns. The contiguous patterns have no subsequences have no subsequences with same minimum support value.

Bioinformatics

Bioinformatics is the combination of computational and biological science. It is an emerging area which strengths computer science, IT and mathematics to

analyze the genetic information. The bioinformatics research is used for mapping genomes of human and to estimate the diversity in population by calculating differences in genome (DNA). Nowadays it is used for mapping many other spices besides human. Biological sequence contains two kinds of sequences namely, 1. DNA sequence and amino acid sequence.

Table 1: Example of DNA sequence database

ID	Sequence
10	ATCGGT
20	CATCGTT
30	CATCG
40	TCGT
50	GCGTGATTC

The DNA (deoxyribonucleic acid) contains information of life in the form of codes and it represents the human genome. Bioinformatics is very important in DNA sequencing and sequences with huge volume of data. The various applications of bioinformatics are pattern discovery, protein folding, alignment and homology, information retrieval and data mining from biological databases, analysis of biological sequence and pattern discovery, micro-array gene expression and gene regulatory network.

Related works and Background study

In the field of biological sequential mining many researchers have made for various applications. The contiguous patterns are mined to analyze the important biological functions in larger genomic sequences. The process of sequencing DNA is to determine three million nucleotide bases (A, T, C, G) in DNA molecule. The Breast Cancer (BC) is one of the reasons for death among women worldwide. The major cause for breast cancer is geographical variation [3]. The two major susceptibility breast cancer genes are BRCA1 and BRCA1 respectively. The BRCA 1 and BRCA2 genes are hereditary breast cancers in most cases with family history of cancer. The mutation in these genes also leads to ovarian cancer, colon, prostate, gastric cancers. There are 2 types of cancer (1) benign cancer and (2) malignant cancer. The breast cancer is malignant type of cancer.

The genetic algorithm and K-means are used to identify cancer patients through DNA micro-array[7]. Seong, Cho *et al.* BRCA1 and BRCA2 genes are more susceptibility genes for breast cancer. The mutation in BRCA1 and BRCA2 genes causes increased risk for early-onset breast cancer development.

Rashid *et al.*[5] proposed efficient approach for mining contiguous patterns from DNA sequences by constructing fixed length spanning tree and uses threshold value which reduces the number of candidate pattern from the sequence.

Zerin *et al.*[6] proposed a fast position-based method to mine contiguous patterns from given sequences which needs only one database scan to construct fixed length spanning tree.

The maximal contiguous pattern mining is used to find the maximal contiguous patterns from sequences[4, 7, 8]. The important problem in bioinformatics is to find maximal contiguous patterns. This research is used to propose method of contiguous pattern mining to identify cancer disease pattern through DNA sequence method.

Concepts and Definitions

In this section, the problem of mining contiguous pattern is defined first and then some basic knowledge of the algorithm is discussed.

Let $\Sigma = \{A, C, G, T\}$ be a set of DNA alphabets where A, C, G, and T are called DNA characters or four bases; A stands for *Adenine*, C for *Cytosine*, G for *Guanine*, and T for *Thiamine*. A DNA sequence S is an ordered list of DNA characters. S is denoted by {S1, S2,…..Sn}, where $Sn \in \Sigma$ and $|S|$ DNA sequence S. A sequence database SD contains <Sid, S> where Sid represents sequence ID and S is a sequence.

Table 2: Sequence Database

Sequence ID	Sequence
1	CAAGC
2	AGCGT
3	CACG
4	AGGCA

For example the above table contains 2 fields sequence ID and Sequence. There are 4 sequences with respective sequence ID for each sequence in the table. The sequence <CAAGC> is 5-sequence since the length of the sequence is '5'. When min_sup = 2 the sequence <AGC> is contiguous frequent sub-sequence because the sequence is included in 3 sequence ID namely 1,2 & 3. Also sequence <AG> is contiguous frequent sub-sequence because it is included in sequence ID 1,2,3 &4. The sequence <CAGC> is one of the contiguous frequent sub-sequence because there is no contiguous super-sequence of <CAGC> with minimum support = 2 respectively.

Definition 1: Given a pattern P and sequence S, the number of occurrences of P in S is the support of pattern P in sequence S. It is represented as support(P, Si). In DNA sequence database SD, the support of P in SD is given as, support (P, SD) = $\sum_{i=1}^{n} \sup port(P,S)$.

Definition 2: In a pattern P =P1, P2,……Pn and a DNA sequence database SD, the confidence of P1, P2…. Pn is defined as confidence (P1, P2) = support (P1, P2, SD)/support (P1, SD)

Definition 3: A Pattern is a contiguous pattern with certain contiguous sub-sequence of DNA sequence S from Σ = {A, C, G, T} A sequence a=<a1, a2,….an> is called as a contiguous sub-sequence of another sequence b=<b1,b2,…..bn> where 'a' is subsequence and 'b' is the super sequence of 'a'.

Proposed contiguous pattern mining technique

In this section, an algorithm Contiguous Sequential Pattern Mining (CSpan) is used for finding contiguous patterns for biological data sequence which can mine the contiguous patterns more efficiently.

Algorithm: 1 Contiguous Sequential Pattern Mining (CSpan)

Input: Sequence Database SD, support sup and minimum support min_sup

Output: Contiguous sequential patterns for user specified minimum support, min_sup

Step 1: Read DNA sequence file.

While (FileInputStream.available()>0)

Step 2: By using n-gram model split the sequence database, SD

Step 3: A, T, C & G occurrence value in sequence are counted. The corresponding values are incremented as count A++, Count T++, Count C++ and Count G++.

Step 4: Check the scanned letter and current letter

If (option==0 & ch= 'A' or ch= 'a')

Call Step 3

Step 5: Count the number of sub-sequences for particular input sequence.

The input parameters for the above algorithm are a sequence database SD, minimum support min_sup. The n-gram model is used to split the sequence database. The backscan pruning is used to remove the unnecessary sequence i.e. no forward or backward extensions. The contiguous is called when the given sequence is frequent with the sequence database SD.

Comparison of normal gene and breast cancer gene / Experimental Results

The experiments for normal and breast cancer gene BRCA1 are carried out on computer Intel Core i5, 3.5 GHz processor with 4 GB RAM and Windows 10 operating system. The program is written in java and net beans are used to run it. The input DNA sequence is stored in text file in FASTA format and program reads the input file to find all contiguous patterns. The data sets used are real life DNA sequences which are downloaded from the website of National Center for Biotechnology Information (NCBI)[4]. The input DNA sequence is 11th exon from 5 humans from NCBI website. By using the advanced search technique in NCBI the normal gene can be downloaded as:

- Search Category = "Nucleotide", (b) Organism = "human" and (c) All Fields = "normal breast gene".

By using the above searching method 5 normal human sequences are downloaded with average of 95 character or Base Pair (BP) in FASTA format.

The Breast Cancer-1 (BRCA1) DNA data set was downloaded using advanced search technique with following parameters,

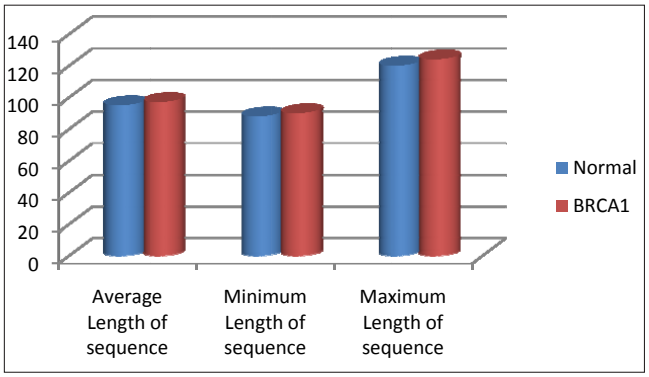
- Search Category = "Nucleotide", (b) Organism = "human" and (c) gene = "BRCA 1 gene".
- All Fields = "breast cancer gene".

By using the above searching method 5 BRCA1 gene for human sequences are downloaded with average of 97 character or Base Pair (BP) in FASTA format.

Table 3: The various details of dataset

Gene Name	Total no. of Sequences	Average length of Sequence	Minimum Length of Sequence	Maximum Length of Sequence
Normal	5	95	88	120
BRCA1	5	97	90	124

The graph 1 represents the variation between the normal gene and BRCA1 gene with various parameters namely average length of sequence, minimum length of sequence and maximum length of sequence. The average length of normal gene sequence is 95 and BRCA 1 gene is 97 respectively. The total number of sequences in normal gene and BRCA 1 gene used is 5.



Graph 1: Various datasets with parameter values

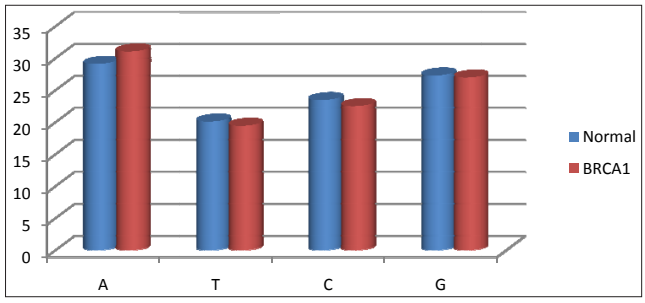
The Table 4 summarizes the percentage value of 5 normal gene patterns with total and average value of individual patterns.

The Table 5 summarizes the percentage value of 5 BRCA1 gene patterns with total and average value of individual patterns.

The Table 6 is used to calculate the mutation value for the various patterns in the 5 sequences. The mutation value is calculated by,

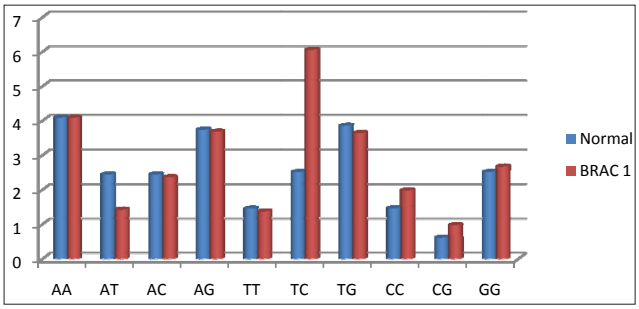
Mutation value = – BRCA1 gene – Normal gene

In the Table 7 represents no change in the pattern, positive value represents increase in particular pattern and negative value represents decrease in particular pattern value.



Graph 2: Normal gene VS BRCA1 for A, T, C & G

From the table 8 and graph 2 it is observed that in normal gene sequence Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) sequences are similar in their values. But BRCA1 gene Adenine (A) value is increased and Guanine (G) value is decreased. The Adenine (A) is increased higher than Guanine (G) in BRCA1 gene. This mutation increase in value may lead to the cancer in the particular patients.



Graph 3: Normal gene VS BRCA1 for other patterns

In the graph 3 various other contiguous pattern

Table 4: Percentage value of each pattern of Normal gene

Patterns	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5	Total	Average
A	28.74	28.34	29.20	30.10	29.10	145.48	29.10
T	20.33	20.73	19.66	19.70	20.00	100.42	20.10
C	23.90	23.10	23.70	24.10	22.50	117.3	23.46
G	27.04	27.34	27.44	26.10	28.40	136.32	27.30
AA	3.05	4.26	3.92	4.48	4.76	20.47	4.09
AT	1.78	2.94	1.94	2.38	3.18	12.22	2.44
AC	2.34	2.17	2.90	2.56	2.23	12.2	2.44
AG	3.32	4.51	3.82	3.90	3.19	18.74	3.74
TT	1.95	1.32	1.62	1.11	1.29	7.29	1.45
TC	2.31	2.28	2.87	2.91	2.21	12.58	2.51
TG	3.12	4.20	3.98	3.43	4.56	19.29	3.85
CC	1.60	1.31	1.21	1.78	1.40	7.3	1.46
CG	0.98	0.63	0.45	0.39	0.62	3.07	0.61
GG	2.67	1.39	2.86	1.84	3.83	12.59	2.51

Table 5: Percentage value of each pattern of BRCA1 gene

Patterns	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5	Total	Average
A	30.84	32.24	29.10	31.00	32.10	155.28	31.05
T	20.23	20.73	19.56	18.70	18.00	97.22	19.44
C	21.90	19.60	22.90	25.10	23.00	112.5	22.5
G	27.04	27.44	28.44	24.10	27.90	134.92	26.98
AA	4.15	4.76	2.92	4.88	3.76	20.47	4.09
AT	0.78	1.74	1.24	1.18	2.18	7.12	1.42
AC	2.14	2.07	2.40	2.36	2.83	11.8	2.36
AG	2.32	3.50	4.82	4.70	3.09	18.43	3.68
TT	0.94	1.02	1.42	1.81	1.69	6.88	1.37
TC	1.31	2.18	2.67	2.01	2.11	30.28	6.05
TG	3.02	4.10	3.48	3.13	4.46	18.19	3.64
CC	1.90	1.61	1.31	2.58	2.50	9.90	1.98
CG	1.98	0.93	0.25	0.89	0.82	4.87	0.97
GG	3.67	2.39	1.66	1.74	3.93	13.39	2.67

frequencies are illustrated. We observed that the pattern TC has more 3.54 positive value i.e., the mutation value is higher when compared to other mutation pattern values. Also the pattern AA has neutral value which means there is no change in that pattern in both normal gene and BRCA1 gene. The pattern CC has 1.52 positive value i.e., the mutation value is slightly higher when compared to TC pattern value. By using these positive values of patterns between normal gene and BRCA1 gene we can say that the person may have breast cancer.

The Table 7 represents the contiguous patterns for normal gene. The minimum support value ranges from 2 to 10. When the minimum support value increases the discovered contiguous pattern is decreased. The

Table 6: Comparison of Normal gene and BRCA1 gene values

Patterns	Normal gene	BRCA1 gene	Mutation value
A	29.10	31.05	1.95
T	20.10	19.44	-0.66
C	23.46	22.5	-0.96
G	27.30	26.98	-0.32
AA	4.09	4.09	0
AT	2.44	1.42	-1.02
AC	2.44	2.36	-0.08
AG	3.74	3.68	-0.06
TT	1.45	1.37	-0.08
TC	2.51	6.05	3.54
TG	3.85	3.64	-0.21
CC	1.46	1.98	1.52
CG	0.61	0.97	0.36
GG	2.51	2.67	0.16

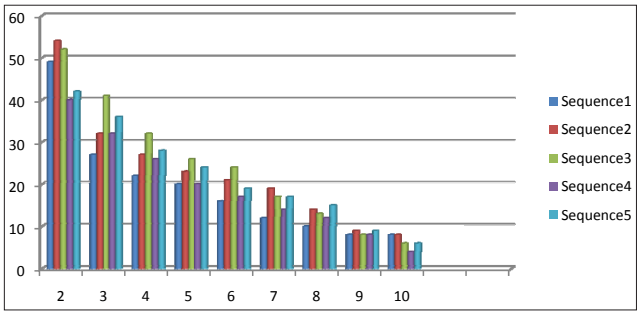
Table 7: Contiguous Patterns for normal gene

Min_Sup	Contiguous Patterns for normal gene				
	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5
2	49	54	52	40	42
3	27	32	41	32	36
4	22	27	32	26	28
5	20	23	26	20	24
6	16	21	24	17	19
7	12	19	17	14	17
8	10	14	13	12	15
9	8	9	8	8	9
10	8	8	6	4	6

Table 8: Contiguous Patterns for BRCA1 gene

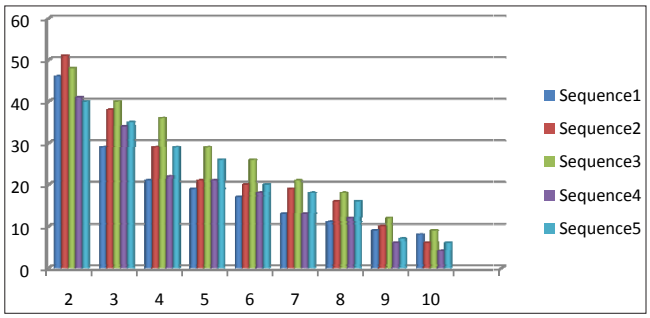
Min_Sup	Contiguous Patterns for BRCA1 gene				
	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5
2	46	51	48	41	40
3	29	38	40	34	35
4	21	29	36	22	29
5	19	21	29	21	26
6	17	20	26	18	20
7	13	19	21	13	18
8	11	16	18	12	16
9	9	10	12	6	7
10	8	6	9	4	6

number of sequences used for evaluation is 5 normal gene sequences. The highest number of contiguous pattern is 54 and least contiguous pattern is 4 for 5 sequences. The graph represents the differentiation in various patterns corresponding to minimum support values.



Graph 4: Normal gene patterns

The Table 8 represents the contiguous patterns for BRCA1 gene. The minimum support value ranges from 2 to 10. When the minimum support value increases the discovered contiguous pattern is decreased. The number of sequences used for evaluation is 5 BRCA1 gene sequences. The highest number of contiguous pattern is 51 and least contiguous pattern is 4 for 5 sequences. The graph represents the differentiation in various patterns corresponding to minimum support values. From this study we may say that Adenine (A) is increased in value and Cytosine (C) is decreased in value.



Graph 5: BRCA1 gene patterns

CONCLUSION

The contiguous pattern discovery is a sequential pattern discovery method in data mining which is used to find the patterns on DNA database. In this finding breast cancer gene BRCA1 and normal gene are used to evaluate the occurrence of breast cancer patterns. To find all patterns in DNA sequence is time consuming task. But analyses of these patterns are more useful in predicting the disease causing patterns and their occurrences. The contiguous pattern mining technique is more efficient than other mining process. It mines only compact contiguous patterns which help to identify and predict the disease causing patterns. Here we have compared the normal gene with breast cancer gene BRCA1 for finding the mutation rate. Also pattern TC is increased abnormally in mutation ratio when compared to normal gene and pattern AA has neutral value i.e. no change in this sequence between normal and BRCA1 gene. Due to this abnormal increase in mutation ratio there may be possibility that the person have breast cancer possibility. In future the BRCA2 gene can be included in the research for finding the disease causing pattern in that gene.

REFERENCES

[1] Aggarwal, C.C. 2015. Data mining: the textbook, Heidelberg: Springer.

[2] J. Han, J. Pei, and M. Kamber, 2011. Data mining: concepts and techniques, Amsterdam: Elsevier.

[3] Juwle A, Saranath D. 2012. BRCA1/BRCA2 gene mutations/SNPs and BRCA1 haplotypes in earlyonset breast cancer patients of Indian ethnicity. *Medical oncology* (Northwood, London, England) **29**(5): 3272-81.

[4] National Center for Biotechnology Information (www.ncbi.nlm.nih.gov)

[5] Rashid MM, Karim MR, Hossain MA, Jeong BS. 2011. An efficient approach for mining significant contiguous frequent patterns in biological sequences; Proceeding of 3rd International Conference on Emerging Databases (EBD'11), Aug 25-27; Incheon.

[6] Zerin SF, Ahmed CF, Tanbeer SK, Jeong BS. 2010. A fast indexed-based contiguous sequential pattern mining technique in biological data sequences; Proceeding of 2nd International Conference on Emerging Databases (EBD'10); Aug 30-31; Jeju.

[7] Zubi, Z.S. and Emsaed, M.A. 2013. Identifying Cancer Patients Using DNA Micro-Arrayy Data in Data Mining Environment. *Journal of Science and Engineering*, **3**: 63-75.