



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Exploration of DNA Sequences Using Pattern Mining

Mr. Dipak R. Kawade

Research Scholar, Department of Computer Science, Shivaji University, Kolhapur, India.

Dr. Kavita S. Oza

Assistant Professor, Department of Computer Science, Shivaji University, Kolhapur- 416 004, India.

Abstract: Easy availability of Low-priced hardware and software, biological data storages are growing rapidly in size and complexity. Bioinformatics plays an important role in development of applications using data mining techniques to solve biological problems by using these data storages. The proposed work provides new technique for analysis of DNA sequences stored in databases. In this analysis we have used statistical pattern mining technique. By using this technique we predict probability of occurrence of breast cancer (BRCA1 gene) in a given DNA sequence.

Keywords: Data mining, bioinformatics, pattern mining, DNA sequence, BRCA1 gene.

I. Introduction

In recent years Biological related data are growing speedily in size and complexity. A huge amount of data is available for extracting high level of information in the form of finding interested pattern from the database.[1] Finding information from these data requires strong computational analysis in terms of both the speed and the sophistication of data analysis. Biological data contains DNA sequences and protein sequences which represent some spectrum that have some explicit features. To identify these biological functional sites or features in a genomic sequence is an important research application [4]. In last few years research in related field is advanced due to innovated biological technology and low cost computer hardware and software. This facilitates high volume and better quality of biological data to be collected and stored automatically. Using better analytical methods these huge amount of data is processed effectively and optimally.[5]

A. Bioinformatics:-

Bioinformatics or Computational biology is new interdisciplinary branch of science which is combination of biological science, information technology and computer science that are mainly used to discover, select and prioritize targets [3]. Biological science provides complex domain whereas computer science is used for efficient storage and analysis of large volumes of complex data. Combining these two fields provides the usefulness in understanding biological systems and increasing the effectiveness of biological research.

Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting and utilizing information from biological sequences. Some of the important research area in bioinformatics are Sequence analysis, Genome annotation, Analysis of gene expression, Analysis of protein expression, Analysis of mutations in cancer, Protein structure prediction, etc.[2] Research in bioinformatics includes development of applications using data mining techniques to solve biological problems. Therefore, there is a great potential to increase the interaction between data mining and bioinformatics.

B. Knowledge discovery in databases and data mining:-

Generally most of data stored in databases are sequential in manner. Some example of sequence data are DNA sequence, Science and Engineering data, medical data, telephone calling patterns, web logs clicks etc. [6]. It is very tedious and time consuming task to find all patterns from such large database. In last two decade large numbers of algorithms are developed to retrieve patterns from large datasets.

The prediction of any diseases is one of the challenging problems for many researchers. Breast cancer is a one of major disease in world today. Cancer in general and breast cancer in specific is still a major cause of concern in the world. Predicting the occurrence of a disease is one of the most interesting and challenging tasks in data mining applications. This paper deals with prediction of breast cancer by using pattern mining technique.

II. Related Work

Identification of relevant genes whose expression is closely related to the outcome using support vector machine has been proposed by [8].

D-Miner algorithm is used for mining concepts under user defined monotonic constraints which increase the applicability of concept. This D-Miner can be used dense Boolean matrices. Here the analyst can provide monotonic constraints on both set components of desired concepts and which can be pushed into the extraction process. [9].

Artificial intelligence technique can also be used to classify cancer samples with a top—down approach. But this technique is time consuming and possible suboptimal feature selection is difficult. Also it requires validation on extended database for more cancer and control samples. [10]

CARPENTER algorithm is used for finding frequent closed patterns in long biological sequences datasets. It can perform different operations on row not on columns. This can remove drawback of high dimensionality of many biological datasets. [11]

Artificial neural networks and decision trees are most commonly used statistical method to develop the prediction models using a large dataset [5] have shown that the decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample whereas artificial neural networks have 91.2% accuracy and the logistic regression models have 89.2% accuracy.

Sequential pattern mining using DNA microarray for predicating grade of breast cancer has been proposed by [13].

III. Proposed Pattern Mining technique

Many researchers have focused on the comparison of one gene with another gene. As t only single gene can not cause any disease, but combination of many genes causes a disease. There is need to study interaction and relationship between them. This study is useful to find interesting pattern form the proteins or group of genes. Generally, laboratory studies, observational studies and clinical trials are used for investigating Breast cancer. In laboratory studies, we have to test a hypothesis under controlled conditions and then find results. It operates in controlled condition which is drawback of this test. Medical studies of humans are involved in clinical trials, which can show the cause-and-effect relationship between the variables and the outcomes [7].

The purpose of this research is to develop predictive model for the survivability in the field of breast cancer. Here a Prefix Tree Acceptor (PTA) has been developed which contains four elements A, C, G & T which are elements of DNA called as nucleotides. To translate nucleotides to amino acids, it has group of three nucleotides called as codons i.e. code for one amino acid. Here in the PTA each state represent unique sequence pattern. At final states we get unique codons. At each state we can calculate frequency of each pattern. At final execution we get regular expression of different patterns and their frequencies.

This algorithm contains five steps described as follows

- 1) DNA sequence is read from file until end of file.
- 2) Count occurrences of each character (A, C, G and T) in DNA sequence.
- 3) From given Prefix Tree Acceptor (PTA) which is Deterministic Finite Automata (DFA). Apply transactions on this PTA. At beginning value of start state is 0 i.e. state=0.
Now read first character from file and apply on PTA, move to next state say state 1. Similarly apply all transaction according to the current state and scanned character. Here it checks cases as mentioned in PTA.
- 4) Before moving to next state, state count will be incremented.
- 5) At each state it represent separate pattern. From step four get state count, which is nothing but pattern count. By adding related state count get pattern count of specific pattern.

Algorithm

Input:-DNA sequence

Output:-Different patterns with their counts.

Step 1:-start reading DNA sequence file from start.

while(FileInputStream.available()>0)

Step 2:-count occurrences of A, C, G and T.

CountA++, CountC++, CountG++

,CountT++

Step 3:-According to the Prefix Tree Acceptor (PTA), scanned letter and current state it can move to next state.

If(state==0 and ch='A' or ch='a')

State=1

//It can check such 20 states and 4 characters.

Step 4: Before move to next state we can increase state count by one.

Stcount[i]++;

Step 5: in this step algorithm count frequencies of all patterns such as AA,ACA,AA[ACGT] by adding different state counts.

stcount[15]+= stcount[31]+stcount[47]+stcount[63]+stcount[79]

// this is for gg pattern. At each state PTA represent different pattern. So adding related pattern calculates frequencies of patterns.

A. Datasets used

All experiments are carried out on Computer with Intel Core i3, 3.3 GHz processor with 2 GB RAM and Windows XP operating system. Program is written in Java 1.6. DNA sequence is stored in a file and program reads this file to find frequencies of DNA patterns.

Data sets used for this application are real life DNA sequences downloaded from the web site of National Center for Biotechnology Information [12]. DNA data set for Normal gene was extracted by using advanced search technique provided by NCBI website with

(1) search category = "Nucleotide", (2) sequence length = [1000:10000], (3) Organism = "human"] and All Field="normal Breast gene sequence". By using above search technique we have Downloaded 12 sequences of each with average 4800 character or base pair. These sequences are in the form of FASTA.

Similarly DNA data set for Breast Cancer-1 (BRCA1) gene was extracted by using advanced search technique provided by NCBI website with (1) search category = "Nucleotide", (2) sequence length = [10000:15000], (3) Organism = "human"] , (4) All Field="Breast Cancer" and (5) Gene="BRCA1".

	Number of sequences	Avg Length	Min Length	Max Length
Normal	12	4827	1372	7287
BRCA1	12	11919	11066	12750

Table-1: data sets details

From this site we have downloaded 12 sequence of each with average 12000 characters or base pair. These sequences are in the form of FASTA. These sequences are stored in file. The program gives results as patterns and their frequencies. Results are recorded in Microsoft excel for analysis purpose.

Following table summarizes result of Normal Breast gene. This table shows percentage value of each pattern.

Pattern	Seq-1	Seq-2	Seq-3	Seq-4	Seq-5	Seq-6	Seq-7	Seq-8	Seq-9	Seq-10	Seq-11	Seq-12	Total	Avg
AA[ACGT]	3.65	4.35	3.79	4.16	4.29	4.54	4.61	4.65	4.74	2.61	3.99	4.55	49.92	4.16
AC[ACGT]	1.30	1.72	1.48	1.79	1.45	1.66	1.98	1.99	1.02	2.22	1.71	1.65	19.97	1.66
AG[ACGT]	2.89	3.15	3.17	2.66	3.31	3.22	2.91	2.93	2.99	2.80	3.50	3.00	36.55	3.05
AT[ACGT]	1.65	3.15	1.63	2.42	1.60	3.36	1.91	1.90	4.74	1.89	1.59	3.41	29.26	2.44
CA[ACGT]	2.68	2.41	2.57	2.37	2.50	2.59	2.61	2.55	2.33	2.80	2.49	2.46	30.36	2.53
CC[ACGT]	1.95	1.40	1.52	1.76	1.50	1.31	1.72	1.68	1.02	1.83	0.93	1.33	17.96	1.50
CG[ACGT]	0.30	0.24	0.34	0.26	0.28	0.21	0.32	0.33	0.29	0.33	0.20	0.23	3.33	0.28
CT[ACGT]	2.03	3.06	1.65	3.00	1.63	3.27	1.57	1.57	3.21	2.54	1.01	3.25	27.79	2.32
GA[ACGT]	1.87	1.58	1.88	1.89	1.84	1.21	5.06	5.07	0.80	3.00	2.02	1.44	27.67	2.31
GC[ACGT]	1.62	1.51	1.39	1.32	1.37	1.35	1.50	1.50	1.31	1.70	1.10	1.30	16.97	1.41
GG[ACGT]	2.38	1.05	2.14	1.74	2.12	1.03	1.56	1.61	1.31	2.41	1.74	1.04	20.12	1.68
GT[ACGT]	1.38	2.06	1.22	1.95	1.19	2.19	1.81	1.81	2.19	2.41	0.95	2.21	21.37	1.78
TA[ACGT]	1.84	1.78	2.72	1.66	2.81	1.82	0.55	0.51	2.55	1.30	4.66	1.72	23.92	1.99
TC[ACGT]	1.78	1.39	1.88	1.55	1.89	1.26	2.13	2.15	0.73	1.50	2.08	1.19	19.55	1.63
TG[ACGT]	3.78	2.16	3.90	2.53	3.93	2.12	0.98	0.96	1.31	1.89	4.22	2.16	29.94	2.50
TT[ACGT]	2.24	2.30	2.06	2.26	2.05	2.17	2.09	2.10	1.53	2.09	2.00	2.20	25.09	2.09
A	31.20	33.48	33.43	30.84	33.49	34.25	34.51	34.58	37.03	28.25	36.44	34.24	401.72	33.48
C	20.84	19.42	19.43	21.13	19.35	19.24	19.03	18.77	20.19	22.11	17.44	19.09	236.05	19.67
G	22.76	22.41	22.37	23.08	22.46	22.31	22.19	22.30	20.41	25.44	21.49	22.52	269.73	22.48
T	25.20	24.69	24.77	24.95	24.71	24.20	24.27	24.35	22.38	24.20	24.64	24.15	292.49	24.37

Table-2- Percentage value of each pattern of Normal Breast gene

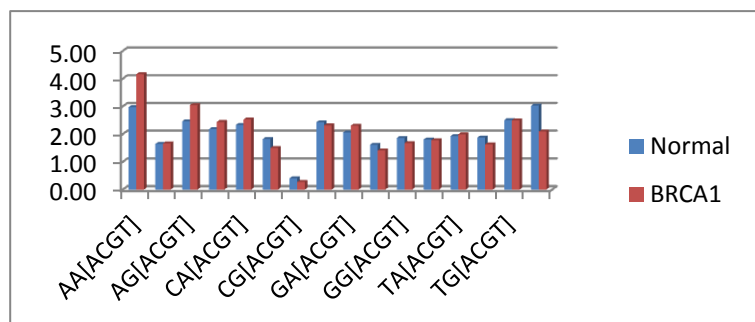
Following table summarizes result of BRCA1 gene which causes Breast cancer.

Pattern	Seq1	Seq2	Seq3	Seq4	Seq5	Seq6	Seq7	Seq8	Seq9	Seq10	Seq11	Seq12	Total	Avg
AA[ACGT]	2.86	2.59	2.78	2.74	2.88	2.78	4.03	3.50	2.82	3.36	2.86	2.47	35.68	2.97
AC[ACGT]	1.49	1.65	1.47	1.65	1.86	1.73	1.64	1.84	1.70	1.55	1.65	1.51	19.74	1.65
AG[ACGT]	2.67	2.58	2.78	2.82	2.52	2.51	1.83	2.11	2.55	2.26	2.45	2.45	29.53	2.46
AT[ACGT]	2.18	1.98	2.13	1.90	1.80	2.32	2.62	2.60	1.97	2.44	2.34	1.90	26.17	2.18
CA[ACGT]	2.22	2.43	2.16	2.15	2.38	2.53	2.32	2.36	2.35	2.31	2.17	2.60	27.98	2.33
CC[ACGT]	1.77	1.90	1.83	1.88	1.73	1.86	1.75	1.84	1.73	1.84	1.63	2.13	21.90	1.82
CG[ACGT]	0.42	0.42	0.52	0.54	0.32	0.31	0.15	0.21	0.45	0.55	0.49	0.48	4.86	0.41
CT[ACGT]	2.44	2.75	2.30	2.29	2.43	2.68	2.33	2.28	2.38	2.19	2.28	2.76	29.11	2.43
GA[ACGT]	2.11	2.18	2.16	2.27	2.01	1.87	2.41	1.99	2.04	1.70	1.89	2.11	24.74	2.06
GC[ACGT]	1.52	1.78	1.62	1.65	1.77	1.66	1.33	1.24	1.84	1.53	1.53	1.92	19.39	1.62
GG[ACGT]	1.91	2.03	2.06	2.21	2.05	2.04	1.16	1.32	2.12	1.82	1.70	1.91	22.32	1.86
GT[ACGT]	1.90	1.92	1.85	1.73	1.68	1.88	1.70	1.75	1.87	1.73	2.03	1.62	21.68	1.81
TA[ACGT]	1.93	1.79	1.89	1.76	1.91	2.03	2.31	2.25	1.63	2.23	1.93	1.48	23.16	1.93
TC[ACGT]	1.88	1.68	1.88	1.95	1.93	1.71	2.04	1.74	2.01	1.68	1.97	2.02	22.50	1.87
TG[ACGT]	2.55	2.70	2.52	2.41	2.68	2.62	2.16	2.44	2.78	2.22	2.42	2.59	30.07	2.51
TT[ACGT]	3.00	2.46	2.91	2.91	2.97	2.66	3.54	3.39	2.62	3.43	3.52	2.92	36.32	3.03
A	27.73	27.38	27.36	27.24	27.71	27.57	31.38	30.98	27.23	29.49	27.70	25.67	337.43	28.12
C	21.34	21.55	21.52	21.66	21.31	21.44	19.67	20.12	21.64	20.80	20.34	23.51	254.89	21.24
G	22.80	23.67	23.74	23.76	22.87	22.98	18.13	18.47	23.82	20.53	21.47	22.88	265.11	22.09
T	28.14	27.41	27.38	27.34	28.11	28.02	30.82	30.42	27.31	29.17	30.50	27.94	342.56	28.55

Table-3 result of BRCA1 gene which causes Breast cancer

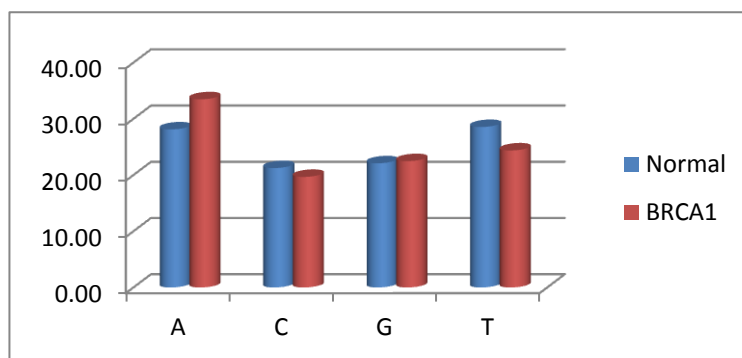
Following table shows average frequencies of each pattern for Normal and BRCA1 gene. This table also shown mutation calculated by using $Mutation = BRCA1 - Normal$.

Pattern	Normal	BRCA1	Mutation
AA[ACGT]	2.97	4.16	1.19
AC[ACGT]	1.65	1.66	0.02
AG[ACGT]	2.46	3.05	0.58
AT[ACGT]	2.18	2.44	0.26
CA[ACGT]	2.33	2.53	0.20
CC[ACGT]	1.82	1.50	-0.33
CG[ACGT]	0.41	0.28	-0.13
CT[ACGT]	2.43	2.32	-0.11
GA[ACGT]	2.06	2.31	0.24
GC[ACGT]	1.62	1.41	-0.20
GG[ACGT]	1.86	1.68	-0.18
GT[ACGT]	1.81	1.78	-0.03
TA[ACGT]	1.93	1.99	0.06
TC[ACGT]	1.87	1.63	-0.25
TG[ACGT]	2.51	2.50	-0.01
TT[ACGT]	3.03	2.09	-0.94
A	28.12	33.48	5.36
C	21.24	19.67	-1.57
G	22.09	22.48	0.39
T	28.55	24.37	-4.17



Graph-1: Comparison between different patterns of Normal and BRCA1 gene

Table-4: pattern for Normal and BRCA1 gene



Graph-2: Comparison between nucleotides of Normal and BRCA1 gene

From above table and graphs it is observed that in normal DNA sequences, frequencies of Adenine (A) and Thymine (T) and frequencies of Cytosine (C) and Guanine (G) are nearly similar. But this is not the case of BRCA1 gene. In BRCA1 gene Adenine (A) level is increased and Thymine (T) is decreased. In BRCA1 sequences frequencies of Adenine (A) grows abnormally as compared to Thymine (T). Also in BRCA1, AA[ACGT] pattern have much positive difference i.e. mutation in this pattern increases. Also we observe that, there is much difference in the pattern AA[ACGT], AG[ACGT], AT[ACGT], CA[ACGT]. From this we can say that, difference between normal and BRCA1 gene for above said pattern increases; there is chance that the person may have breast cancer.

IV. Conclusion

It is very tedious and time consuming task to find all patterns in large DNA sequences and their analysis. Analyses of DNA patterns are useful to predict diseases. Here we have compared BRCA1 gene and normal gene by using pattern mining technique. From this study we may say that if Adenine (A) level is increased and Thymine (T) level is decreased and at same time some pattern related to Adenine are increased then that DNA may be affected by BRCA1 gene i.e. there may be possibility that the person have breast cancer.

V. References

- [1]. Vladimir Brusilovskiy And John Zelezniak, Knowledge discovery and data mining in biological databases. The Knowledge Engineering Review, Vol. 14:3, 1999, 257-277
- [2]. Khalid Raza, Application Of Data Mining In Bioinformatics, Indian Journal of Computer Science and Engineering Vol 1 No 2, 114-118
- [3]. Yongliang Yang, S. James Adelstein and Amin I. Kassir, Drug Discovery Today Volume 14, Numbers 3/4 February 2009
- [4]. Huiqing Liu and Limsoon Wong, data mining tools for biological sequences, Journal of Bioinformatics and Computational Biology Vol. 1, No. 1(2003) 139-167 Imperial College Press
- [5]. Dursun Delen, Glenn Walker, Amit Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine (2004) xxx, xxx-xxx.
- [6]. Zhenglu Yang, Yitong Wang, Masaru Kitsuregawa : Effective Algorithms for Sequential Pattern Mining, DBSJ Letters Vol.5, No.1.
- [7]. The ABCs of breast cancer—types of research studies (http://www.komen.org/bci/abs/chap_01.asp).
- [8]. Huiqing Liu, Jinyan Li and Limsoon Wong, Use of extreme patient samples for outcome prediction from gene expression data Bioinformatics Vol. 21 no. 16 2005, pages 3377-3384
- [9]. Jeremy Besson, Celine Robardet, Jean-François Boulicaut and Sophie Rome, Constraint-based concept mining and its application to microarray data analysis, Intelligent Data Analysis 9(1):59-82, 2005. IOS Press.
- [10]. Lihua Li, Hong Tang, Zuobao Wu, Jianli Gong, Michael Gruidl, Jun Zou, Melvyn Tockman, Robert A. Clark, Data mining techniques for cancer detection using serum proteomic profiling, Artificial Intelligence in Medicine (2004) 32, 71-83
- [11]. Feng Pan, Gao Cong, Anthony K. H. Tung, Mohammed J. Zaki, CARPENTER: Finding Closed Patterns in Long Biological Datasets SIGKDD '03, August 427, 2003, Washington, DC, USA
- [12]. National Center for Biotechnology Information (www.ncbi.nlm.nih.gov)
- [13]. Mickael Fabregue, Sandra Bringay, Pascal Poncelet, Maguelonne Teisseire, Béatrice Orsetti, Mining microarray data to predict the histological grade of a breast cancer, Journal of Biomedical Informatics 44 (2011) S12-S16