# Performance Comparison of FSA Red & Apriori Algorithm's in Mutation Analysis

Mayilvaganan M<sup>#1</sup>, Hemalatha R<sup>\*2</sup>

<sup>#1</sup>Associate Professor & Dept.of Computer Science & PSG College of Arts and Science, Coimbatore, India. <sup>#</sup>Research Scholar(Ph.D) Karpagam University, Dept.of Computer Science, Coimbatore, India.

Abstract-In this paper the attempt has been made to analyze the DNA gene cancer dataset with RBC, WBC and platelet cancer data set. The basic idea behind this proposed method is comparing the 3 large nucleotide DNA dataset with with Bloom filter and discovering the matched subsequence. To validate the proposed algorithm,

association and classification rule based on the FSA red algorithm with bloom filters and apriori algorithm using hierarchical clustering are compared using data mining technique. Here this algorithm is applied to find no of sequence occurrences and mutation analysis for the 3 nucleotide DNA gene dataset. In order to evaluate the proposed methodology, Comparisons are made based on the Execution time and memory efficiency in finding frequent patterns. The extracted rules and analyzed results are graphically demonstrated. The performance is analyzed based on the different no of instances and confidence in DNA sequence data set.

Keywords: Association Rule and Classification, Zero rule, fsa red and Apriori algorithm.

#### I. INTRODUCTION

The data from the human genome project is likely to be of significant assistance in medical genetics, including diagnosis of diseases. According to Fredj Tekaia, bioinformatics is defined as " the mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information. Classification is a machine learning discipline, and is inspired by pattern recognitions, which is a branch of science. The data classification process involves learning and classification. Association rule mining is the discovery of association relationships or correlations among a set of items. Sequence alignment is a method of aligning sequence RNAs, DNAs and proteins, so, sequence alignment determines locations of similarity between sequences. These places of similarity between two or more sequences, can indicate the relationship between functional, structural, and [2]. evolutionary times between sequences Different solutions are presented for sequence alignment problems,

such as the Hidden markov model, Heuristic algorithms, and Dynamic programming. In recent decades, advances in molecular biology and required equipment for this research caused a rapid increase in sequencing the genomes of many organisms of species, in such a way that the genome sequencing projects are common projects in Bioanformatic[1].

# II. RELATED WORK

Varities of techniques are presented for sequence alignment criteria, and in this module describe these methods. There are pre planned works which use Dynamic programming, such as the Needleman -Wunsch and the Smith -Waterman. These methods have two ladder. In the first step,

the algorithm assigns a substitution matrix for comparing and scoring characters of the input succession, and in the second move, traces back the replacement matrix for estimating an answer for the problem. Order of these methods is exponential.. These methods find the answer of sequence alignment problems by combining a pair of aligned characters that start from the most similar couple and end to the most un similar couple. These methods have two steps [5]:

1) Presenting the relationship between the sequences using a supervision tree.

2) Running Multi Sequence Alignment (MSA), according to the guidance tree for adding sequences to each other. The heuristic methods are efficient to be implemented on large scales(100 - 1000), But in these methods, the answer is not assured to come together to the most favourable solution. Evaluation of this methods is very difficult and biological concepts are vague [6]. The next method is the Hidden Markov model. the Hidden Markov model is based on probabilities which can be considered

for all possible cases of arrangement of jumps, matches and mismatches to find the most likely aligned sequences or a set of associated sequences. The order of these methods is more worse than other methods[5]. In this paper, we propose an algorithm to align the sequences using Bloom filters that is more efficient and fast.

### III. PROPOSED METHODOLOGY -FSA RED ALGORITHM

The idea behind this proposed method is comparing subsequence of incoming sequences with Bloom filter and discovering the matched subsequence. The sequence alignment problem has a large number of comparisons. According to the last successful experiences of useing Bloom filter in similar issues such as web search, it seems that using Bloom filter to solving this problem can be an appropriate solution. Algorithm is used for data reduction or pre-processing to minimize the attribute to be analyzed. The goal is to make strong association rules using data mining techniques related to the data which is reduced . The data preprocessing in FSA-Red performed a few of reduction techniques such as attribute selection, row selection and feature selection. Row selection has done by deleting all signed record which related to the attribute which need to be analyzed. Feature selection will remove all the unwanted attribute, ended with attribute selection to eliminate the non value attributes which is no need to be included..

# Sequence Alignment of Data for research

The genetic code is the set of rules by which information encoded in genetic material. The data set used for proposed work in DNA sequences and it is obtained from PubChem. Pubchem is an important public, web based information source for and bioactivity information. chemical The information in DNA is stored as a code made up of A,G,C,T chemical codes. A-Adenine, G-Guanine, C-Cytosine, T-Thymine. Human DNA consists of three billion bases and more than 99 percent of those bases are the same in all people. This database contains 40000 gene sequences. BLAST(Basic Local Alignment Search Tool) are the most popular heuristic methods for sequence alignment.

Table 1 heuristic sequence alignment of DNAdataset with respect to RBC,WBC,PLATELET

AGCGAGCATCT CGAAACAAAC CTCAACT CCAAATCCTT CACTGTCCAC ACAAGTAC ACTCCATGTCT TGCTGGATGG

#### Word list of amino acids C T C T A G C A T T A G T G C A C C A

#### Heuristic method of alignment of codons

C T C T A G C A T T A G G U - G C A C C C A (insert a character U in the place of T)

In the above heuristic method of alignment the Nucleotide BLAST search was performed with one input nucleotide sequences and compared these against other nucleotides. In the above data set red colour indicate the exact match was found in the nucleotide sequence.

# FSA RED ALGORITHM Implementation

Purpose of this phase is discovering the similar subsequence's between incoming sequence and the most similar sequence in the database, using Bloom filter. For this purpose, first process of this phase creates subsequence's of incoming sequence in three groups, 2 characters, 3 characters, and 4 characters. Each created subsequence is compared with corresponding Bloom filter. If it be found in the Bloom filter, the subsequence and its final index are added to a queue. After comparison of subsequence's with specific length, queue will be rated and go to the next group. Finally, after examining all of the sequences of database, best queue will be sent to Phase II.

**Algorithm**: align (A<sub>1</sub>, A<sub>2</sub>)

**Input**: string  $A_1$  of length m, string  $A_2$  of length n

Initialize matrix D properly; for x=1 to m, for y=1 to n max = D[x-1][y] + gapScore2 (A<sub>2</sub>[y-1]) if max < D[x][y-1] + gapScore1 (A<sub>1</sub>[x-1]) max = D[x][y-1] + gapScore1 (A<sub>2</sub>[x-1]) if max < D[x-1][y-1] + matchScore (A<sub>1</sub>[x-1], A<sub>2</sub>[y-1])

 $\begin{array}{l} max = D[x-1][y-1] + matchScore \ (A_1[x-1], \\ A_2[y-1]) \\ [x][y] = max \quad endfor \ 13. \quad Endfor \ return \\ D[m][n] \end{array}$ 

# Output: value of optimal alignment

Purpose of this phase is aligning sequences based on sub sequences location in both sequences which are in the best queue. Since it is possible that location of subsequence in the two sequences has situations that are in appropriate and too far, so the nucleotide sequence, established conditions for the acceptance subsequence.

# PRE-PROCESS PHASE

while  $2 \le n \le 4$  do for  $i = 0 \rightarrow m - \text{length} + 1$  do  $CS \leftarrow \text{NextSubsequence}(i; \text{length})$ BloomFilter:Add(CS) end for Save(BloomFilter) n + + end while

Table 2. Sample Nucleonue Data Set Sequence					
А	1	-1	-1	-1	-2
С	-1	1	-1	-1	-2
Т	-1	-1	1	-1	-2
G	-1	-1	-1	1	-2
	-2	-2	-2	-2	

Single codon letters

Algorithm -Locating Subsequence In The Sequence Of Database

while  $2 \le n \le 4$  do for  $i = 0 \rightarrow m - \text{length} + 1$  do  $CS \leftarrow \text{NextSubsequence}(i; \text{length})$ if thenBloomFilter:Contain(CS) Queue:Enquqe(CS) Score = Scoring(Queue:Count()) if thenScore > BestScore BestQueu  $\leftarrow$  Queue n + + end while

#### Table 3:Nucleotide Sequence of DNA data set

string 1	А	С	А	
String 2	А	С	G	
Alignment Score	1	1	-1	1

The above table represents the two strings nucleotide DNA data set. If the single codon match is found the alignment score is calculated as 1 otherwise -1.

The following figure 1 represents the single codon and double character heuristic search with the amino acids A ,C,G and T respectively. In WBC cancer nucleotide dataset sequence double pair of amino acid sequence such as AA, AC, AG,AT,CA and CG are estimated.

lew New Open Compare Import cript ▼ ▼ Data Wo	Bi Dope Vanade + ∂ Run ed Time Smalink Layout Seport Asport 2 Oler Vanade + ∂ Report Sopport Asport 2 Oler Vanade + Dear Vanade + Dear Vanade + AdoOns +			
FILE	VARABLE CODE SMUUNK EW/ROMENT RESOURCES			
ment Folder	Common Window	Worl	distance	
D Nama a	Konstant Mitta AB2 Watch this Vision can Semanler, as used Setting Stated	X	inspect	
mðiregistry registry util	mitochondria 1x16569 33138 char wbc 1x6497 12894 char	^ iii ar ⊯ m	ns nitochondria Abc	<1x1 struct> 'GATCACAGGTCTAT 'ATCCTCAGAATTCC
win32 deploytool.bat insttype.ini	>> basecount(who) ans =			
Icdata.xsd Ticdata.utf8.xml	A: 1320	=		
matlab.bat	C: 2186			
matlab.exe	G: 1838			
mbuild.bat	T: 1103			
mcc.bat		€	п	
MemShieldStarter.bat	>> figure	Com	mmand History	
💁 mex.bat	<pre>basecount(mitochondria,'chart','bar');</pre>	Com	in and indexity	
mex.pl	>> figure		basecount (w	bc)
	dimercount (mitochondria, 'chart', 'bar')		figure	
mexet.bat			basecount (m	itochondria.'
mexet.bat mexetup.pm			fimure	
S mexet bat mesetup.pm mexetis.pm	ans =			
imexet.bat mexetup.pm mexetis.pm move_modules_installer.pm	ans =		dimensiont (	nitochondria
mexet bat messetup.pm mexufis.pm morc_modules_installer.pm morc_modules_installer.pm mexufis.pdf	ans = 38: 1604		-dimercount (	mitochondria,
Simeset bat messetup.pm mesubits.pm mevc_modules_installer.pm Sime_mpiesec.bat Si worker.bat	ang = A&: 1604 AC: 1495		-dimercount( -figure	mitochondria,
mexet.bat mexet/spm mexufis.pm merc_modules_installer.pm merc_modules_installer.pm merc_modules_installer.pm worker.bat	ans = Ah: 1604 Ac: 1495 Ab: 795		-dimercount( -figure -basecount(w	mitochondria, bc,'chart','b
G mozekbat mesetappm mesufapm mesufapm g mer, modules, installer,pm g mer, mpiper, bat g worker,bat	ans = A: 1604 AC: 1495 AS: 795 AT: 1220		-dimercount ( -figure -basecount (w -figure	mitochondria, bc,'chart','b
imesekbat mesekbap mesekbap more, modeles, installer, pm مریس repierse, bat worker, bat	azz = Ba: 1606 BC: 1485 BC: 195 AC: 1595 Ch: 1534		-dimercount ( -figure -basecount (w -figure -dimercount (	mitochondria, bc,'chart','b wbc,'chart','
© meetabit messtap pm immentables intaller pm © mor, modules intaller pm © mor, modules intaller pm © morter bat	ano = &: 1604 &: 1455 &: 795 &: 1230 C:: 1534 C:: 171		dimercount ( figure basecount (w figure dimercount ( figure	mitochondria, bc,'chart','b wbc,'chart','
G montch t messthypm messthypm merychodeler, intallerpm merychodeler, intallerpm merychotek worket.bkt	azg = X X 1604 XC: 1935 XC: 1935 XC: 120 CL: 1534 CC: 1771 (6 CC: 435	Ţ	dimercount ( figure basecount (w figure dimercount ( figure dimercount (	mitochondria, bc,'chart','b wbc,'chart','j wbc,'chart','j

Fig1 wbc cancer single codon data sequence:

# **Mutation Analysis**

A mutation occurs whenever there is a change in the genetic information of an organism, due to a variety of causes. In point mutations, it is important to remember which bases are **purines** (A/G) and which are **pyrimidines** (C/T). When a point mutation causes a purine to convert to another purine (for example, C to T), this is known as a **transition**. When a point mutation changes a purine to a pyrimidine, or vice versa, (i.e., A to T), this is known as a **transversion**.

# Mutation Algorithm

Target  $\leftarrow$  Queue:Dequeue()  $i \leftarrow$  Target:F inalIndex() for j = i - 2 > Length  $\rightarrow i$  do CS  $\leftarrow$  NextSubsequence(j; length) if Target == CS then Aligning()



Fig 2 : flowchart for mutation analysis

HOME PLOTS A	ANS .	skorrouns	2 3 4 5 6	7 8 9 D	mentation	
= 비나님 양 🔒 🕨 나나 Fample wol	ଳ) ଜା	Campad Window	Workstace			
D Name &		L No these	Name of	Velue	16.	Mari
Imra-     grap par constant, file     grap par constant, file     TH-Versmith Package, file     Using the Unsernalised Booth Rail     Songher Exceeding Booth Rail     Constant and Unsernative Booth Rail     Constant and Unsernative Booth Rail     Constant Booth Rail     Constant Booth Rail     Songher Exceeding Booth Rail     Using the Unsernative Booth Rail     Songher Exceeding Booth Rail     Visualize and Investigate Roton Face     Visualize and Investigate Roton Face     Visualize and Investigate Roton Face	les 1 Alignmen tm 1 Alignmen	>> theo Remains: The recent 219493.3 has been replaced by 179635. Remains: proceed 20949453 >>> Management proceedings of the second	Name + A A D D ND25eq ND25er12 ND25er2 ND25er2 ND25er2 Statistex Statistex Statistex a backet d of matrix of ensy of soft statistex of of of soft statistex of of of of soft statistex of of of of of of of of of of	Value           440         struct>           4470           4470           (1)           'gpreccacousurgerge.           (1)           (1)           (2)           (2)           (3)           (3)           (4)           (3)           (4)           (3)           (3)           (3)           (3)           (3)           (4)           (3)           (3)           (3)           (3)           (3)           (3)           (4)           (3)           (4)           (4)           (4)           (4)           (4)           (4)           (5)           (1)           (1)           (1)           (1)           (1)           (1)           (1)           (1)           (1)           (1)           (1)           (1)           (1)           (1) </td <td>4470 4470 4470</td> <td>Max 4470 4470 -9,7000 1</td>	4470 4470 4470	Max 4470 4470 -9,7000 1
		In getgenhank at 67	E How			
		12 1200 N. C 120 = ALCOLOTTOROMATINGTOMOTOROMOTOROMOTOROMOTOROMO 121 =	-clc -clc -clc -clc -clc -clc -clc			
		and consider the constant of	clc -rbcc -clc -rna -clc -rna -clc			
on m (11/T) /D Ender	^	۲ <u>.</u>	There			

Fig 3 Mutation analysis in RBC cancer nucleotide sequence

# IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

Figures(4), (5) and (6) shows that amount of used memory, efficiency and speed in the proposed algorithm that is less than the apriori algorithm. This is because of the Bloom filter is a compact data structure and matrix is a data structure that growths quickly. Definition of response time is the time required to produce output regardless of the output quality. The implementation results show that the fast sequence alignment algorithm can improve the quality, memory and speedup metrics. Results show that in average efficiency 71%, memory usage 93.3% and response time 0.96% as compared to apriori algorithm method are improved.



Fig 4 FSA RED algorithm efficiency in Wbc data based plots:



Fig 5 FSA RED algorithm efficiency in RBC data based plots:



Fig 6 FSA RED algorithm efficiency in PLATELET data based plots:

#### V. CONCLUSION AND FUTURE SCOPE

In this paper, the proposed fast sequence alignment algorithm using Bloom filters is used for DNA nucleotide set sequence of cancer affected RBC,WBC and Platelet. Our proposed algorithms included two phases: phase I is related to discovering the similar subsequence's and creating a queue of similar sequences. In phase II, FSA algorithm aligns the sequences using subsequence's queue of phase I. The implementation results show that the fast sequence alignment algorithm can improve the quality, memory and speedup metrics. Results show that in running time memory usage and speed are better compared to apriori algorithm. . In future the work will be extended in the following areas. I) identifying the root cause of the cancer ii) analyse the future spread over portion of the cancer affected areas.

[14] F.Braz,"A review of the association rules data mining techniques for the analysis of gene expressions"

#### References

- [1] Role of Association Rule Mining in Numerical Data Analysis Sudhir Jagtap, Kodge B. G., Shinde G. N., Devshette P. M
- [2] M.Anandavalli, M.K.Ghose ,K.Gauthaman,"Association Rule Mining in Geonomics", International journal of Computer Theory and Engineering Vol.2, No.2 April 2010.
- [3] Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G. Piatetsky-Shapiro & W. J. Frawley, eds, 'Knowledge Discovery in Databases', AAAI/MIT Press, Cambridge, MA.
- [4] Role of association rule mining in numerical data analysis, sudhir Sudhir Jagtap, Kodge B. G., Shinde G. N., Devshette P. M
- [5] Bayardo, Roberto J., Jr.; Agrawal, Rakesh; Gunopulos, Dimitrios (2000). "Constraint-based rule mining in large, dense databases". *Data Mining and Knowledge Discovery* (2): 217–240. doi:10.1023/A:1009895914772.
- [6] Webb, Geoffrey I. (2000); Efficient Search for Association Rules, in Ramakrishnan, Raghu; and Stolfo, Sal; eds.; Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000), Boston, MA, New York.
- [7] http://www.b3intelligence.com/NumericalDataMinig.html
- [8] http://en.wikipedia.org/wiki/Numerical\_analysis
- [9] http://www.saedsayad.com/zeror.html
- [10] http://www.cogsys.wiai.unibamberg.de/teaching/ss05/ml/slides/cogsysI I-6.pdf
- [11] http://www.slideshare.net/totoyou/covering-rulesbased-algorithm
- [12] M.Anandavalli , M.K.Ghose , K.Gouthaman ,"Association Rule Mining in Genomics", International journal of computer Theory and engineering ,Vol.2,No.2 April,2010.
- [13] Arun.K.Pujari'data mining techniques ",Universities Press (india) private limited.2001.

[15] Douglas Trewartha, "Investigating data mining in MATLAB ",Rhodes University 2006.