INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

ROLE OF DATA MINING IN NUCLEOTIDE SEQUENCE OF NORMAL AND CANCER AFFECTED LIVER CELLS

M.Mayilvaganan[1], Rajamani R

[1]Associate Professor, Dept. of Computer Science, PSG College of arts and science, Coimbatore, TamilNadu, India *dev7aki@gmail.com*
[2]Assistant Professor, Dept. of Computer Science, PSG College of arts and science, Coimbatore, TamilNadu, India

Abstract

The Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Clustering algorithm used to find groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. This paper comprises of two databases such as normal liver cells and cancer affected cells. Each character variables are assigned numeric number and its corresponding pair combination of sequence are represented in a graph. The performance is analysed based on the different no of instances and confidence in gene data set. The occurrences for modified data and original data are compared together to find cluster structure.

Keywords—Cluster algorithm, data base, liver cells, gene data.

1. INTRODUCTION

Clustering can be considered the most important *unsupervised learning* technique; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabelled data. Clustering is "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The dendrogram is a visual representation of the spot correlation data.

The individual spots are arranged along the bottom of the dendrogram and referred to as leaf nodes. Spot clusters are formed by joining individual spots or existing spot clusters with the join point referred to as a node. This can be seen in the diagram above. At each dendrogram node we have a right and left sub-branch of clustered spots. In the following discussion, spot clusters can refer to a single spot of a group of spots. The vertical axis is labelled distance and refers to a distance measure between spots or spot clusters. The height of the node can be thought of as the distance value between the right and left sub-branch clusters.

CLUSTERING ALGORITHM

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modelling puts clustering in ahistorical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis. [2].

DATA FOR RESEARCH

This data set includes descriptions of DEFINITION Homo sapiens occludin (OCLN), transcript variant 1, mRNA.
ACCESSION NM_002538 XM_003118543 XM_936894 VERSION NM_002538.3 GI: 327478412 KEYWORDS.SOURCE Homo sapiens (human) ORGANISM Homo sapiens Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.REFERENCE 1 (bases 1 to 6451) AUTHORS Al-Sadi,R., Khatib,K., Guo,S., Ye,D., Youssef,M. and Ma,T. TITLE Occludin regulates macromolecule flux across the intestinal epithelial tight junction barrier JOURNAL Am. J. Physiol. Gastrointest. Liver Physiol. 300 (6), G1054-G1064 (2011) PUBMED 21415414 REMARK GeneRIF: Suggest occludin plays a crucial role in the maintenance of tight junction barrier through the large-channel TJ pathway, the pathway responsible for the macromolecule.

Normal liver cells Data for Research and Cancer Affected Liver Cells

gcctctctcc atcagacacc ccaaggttcc atccgaagca ggcggagcac cgaacg ccggggtggt cagggacccc catccgtggccccctagg agcccgcgcc tctcctctgc 121 gccccgcctc tcgggccgca acgtcgcgcg gttcctttaacagcgcgctg gcagggtgtg ggaagcaggaccgcgtcctc ccgccccctc ccatccgagt ttcaggtgaa ttggtcaccg 241 agggaggagg ccgacacacc acacctacac tcccgcgtcc acctctccct ccctgcttcc 301 tctggcggag gcggcaggaa ccgagagcca ggtccagagc gccgaggagc cggtctagga 361 cgcagcagat tggtttatct tggaagctaa agggcattgc tcatcctgaa gatcagctga421 ccattgacaa tcagccatgt catccaggcc tcttgaaagt ccacctcctt acaggcctga 481 tgaattcaaa ccgaatcatt atgcaccaag caatgacata tatggtggag agatgcatgt 541 tcgaccaatg ctctctcagc cagcctactc tttttaccca gaagatgaaa ttcttcactt 601 ctacaaatgg acctctcctc caggagtgat tcggatcctg tctatgctca ttattgtgat 661 gtgcattgcc atctttgcct gtgtggcctc cacgcttgcc tgggacagag gctatggaac ggattgtcag agaacagtgc ctatcgctgg accgttgtga aaacagagga gggaaggcaa 2341 gctctggagc cgctccctca gggcatccag gagtctctaa acaactcttc ccctggggat 2401 ttagaggaag ttgtcaagat ggaacctgaa gatgctacag aggaaatcag tggatttctt 2461 tgagctagga gaataagagt ctggagactg ggagccttca cttcggcctc cgattggtgg 2521 cgcatagggt gtaaccaata ggaaacccct aaagggtact taaaccccag attttgcaac 2581 tggggctctt gagcagcttg ctttagcctg ctcccactct gtggaatata cttttgcttc 2641 aataaatctg tgcttttatt gcttcattgt ttcattgaaa aaaaaaaaaa aaaaa

II. METHODOLOGY

The proposed methodology is using gene dataset for mining. The proposed data and outputs are taken to find the occurrences. The occurrences are applied in dendrogram to get a structure for the output.. Cluster analysis is an exploratory data analysis tool for solving classification problems. Its object is to sort cases (people, things, events, etc.) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong; and this description may be abstracted through use from the particular to the general class or type.

III. IMPLEMENTATION

Implementation is a stage, which is crucial in the life cycle of the new system designed. It is the process of changing from the old system to new one. In the existing research work association rule mining is performed in Gene databases. But in proposed clustering algorithm is used based on dendrogram method.. Pre-processing is nothing but data cleaning.

The unnecessary information is removed or reconfigures the data to ensure a consistent format. Data can be modified or changed into different formats. Cluster analysis is thus a tool of discovery. D may reveal associations and structure in data which, though not previously evident, nevertheless are sensible and useful once found. The results of cluster analysis may contribute to the definition of a formal classification scheme, such as a taxonomy for related animals, insects or plants; or suggest statistical models with which to describe populations; or indicate rules for assigning new cases to classes for identification and diagnostic purposes; or provide measures of definition, size and change in what previously were only broad concepts; or find exemplars to represent classes.

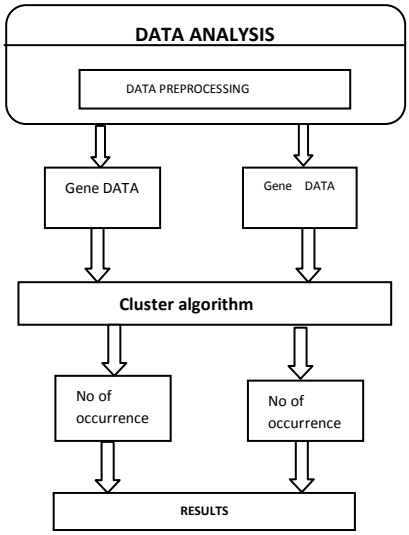


Figure1 Process Flow Diagram

Figure 1 shows the process flow of the proposed system. The performance of proposed work is measured with the existing techniques. Apriori algorithm is applied using association rule mining technique. The Count and position of gene sequences are retrieved using Apriori algorithm. This algorithm is applied separately in string and numerical data. Memory efficiency is calculated and comparisons are made based on it.

III. RESULTS AND DISCUSSION

The figure 2 & 3 shows the combination pair of sequence for normal cells and cancer affected cells. The following Figure4 shows the memory occupied by the string data and numerical data based on Apriori algorithm which we discussed above.

Vol.2 Issue.6, Pg.: 77-82
June 2014

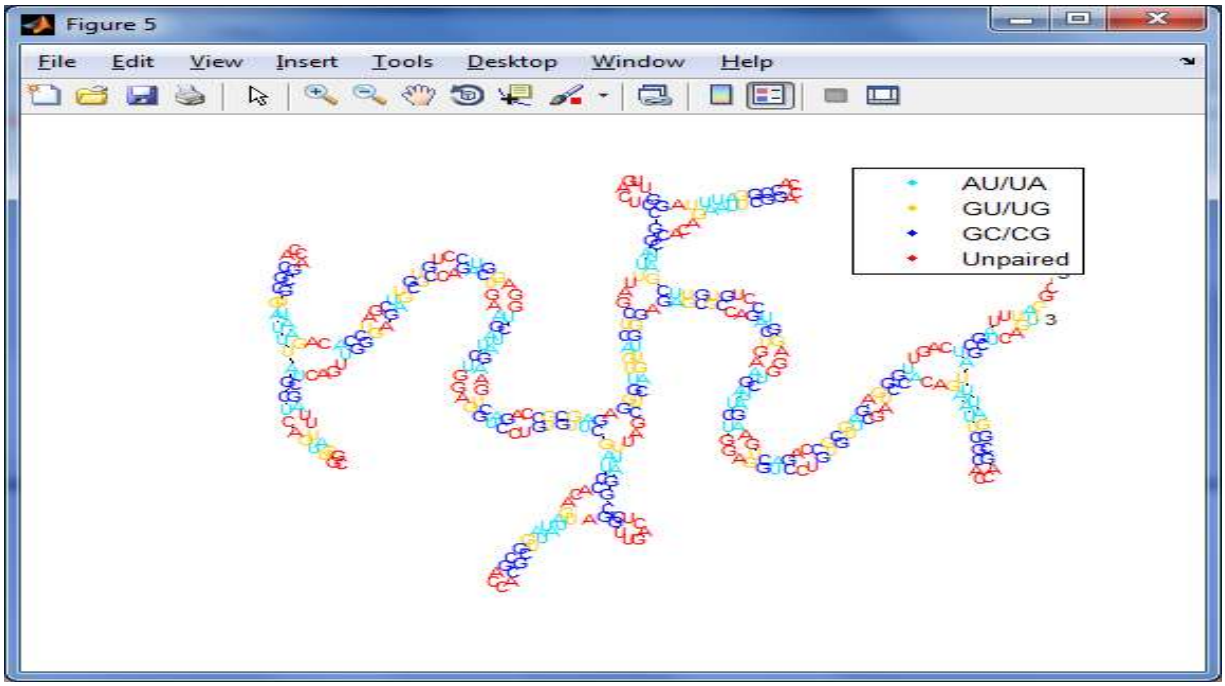
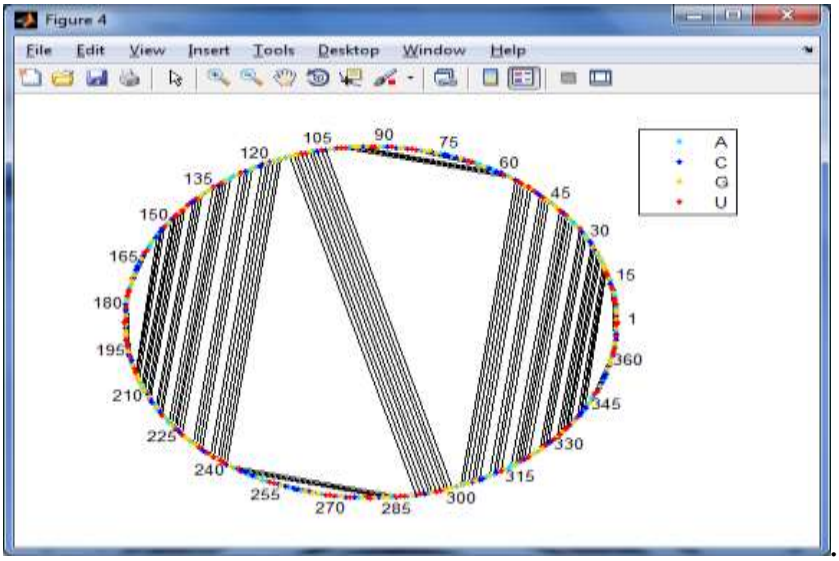


Figure 2 Sequence Pair Combination of Sequence Normal Cells

There are various methods of sequence alignment. These methods differ in the approach, computational complexity and accuracy of results. Dot matrix method is very useful for simple alignments. This method utilizes the graphical methods and it is easy to understand and apply.

Graphic similarity comparisons use the power of the computer to present relationships between sequences in such a graphic form that enables the humans to discern patterns in the data. The following fig 5 represents the double character search in cancer affected liver cells. The following fig 3 represents the triple character search in cancer affected liver cells.



Vol.2 Issue.6, Pg.: 77-82
June 2014

Figure 3 Pair Combination of Sequence Cancer Affected Cells

In fig 4 the blue line indicates the memory efficiency level of numerical data and the red line indicates the timing efficiency level of string data. Finally the comparisons are made and based on the comparison we can conclude that the memory level of Numerical data is very less based on memory for various no of instances.

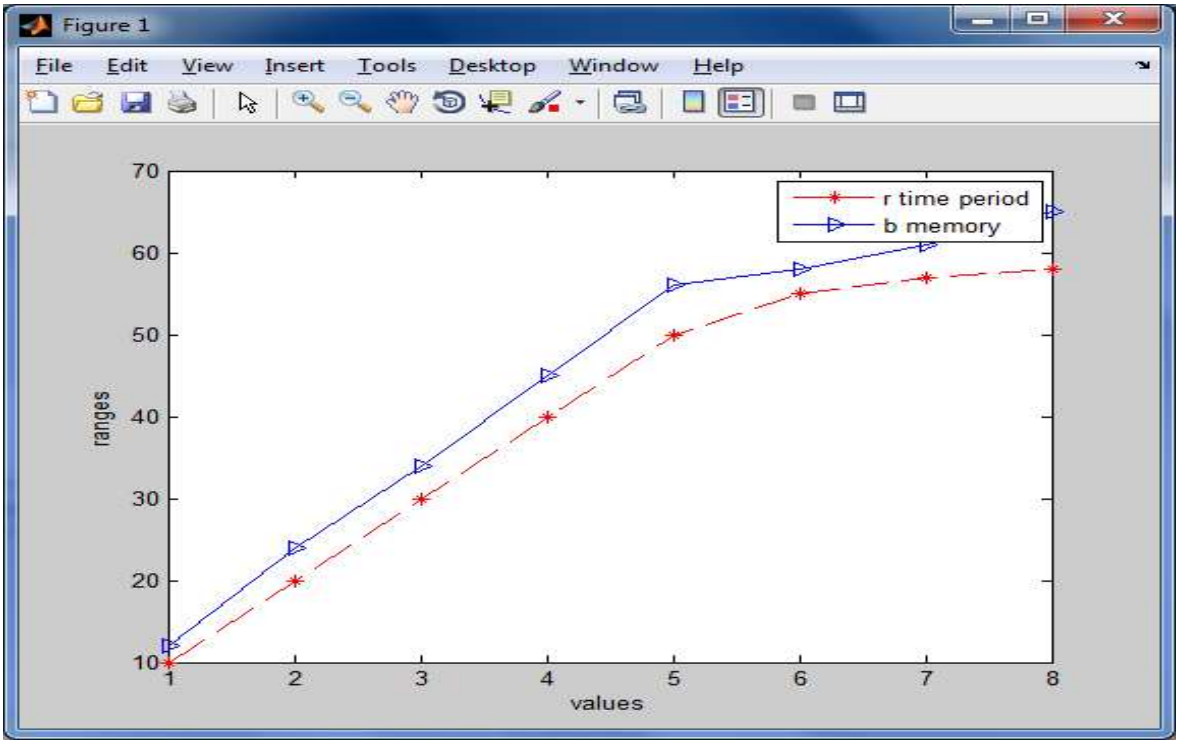


Fig 4 Time A Memory Efficiency For Numerical Data

In fig 5 and 6 shows the double and triple character search sequence in liver cancer nucleotide sequence.

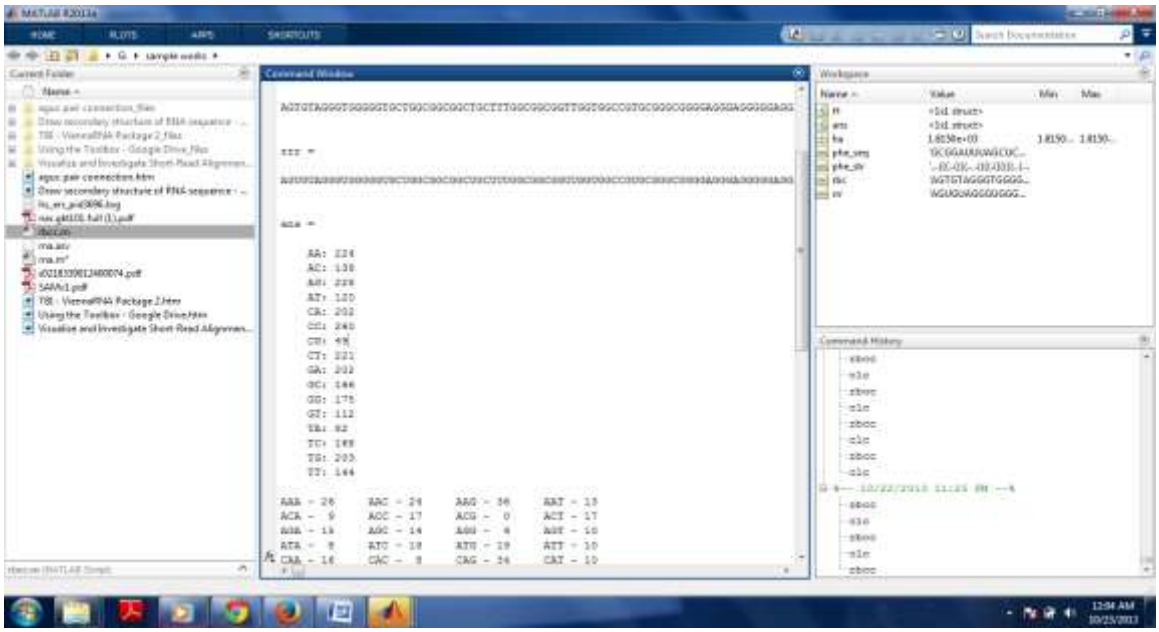


Fig 5 Double Character Search Sequence in cancer affected liver cells

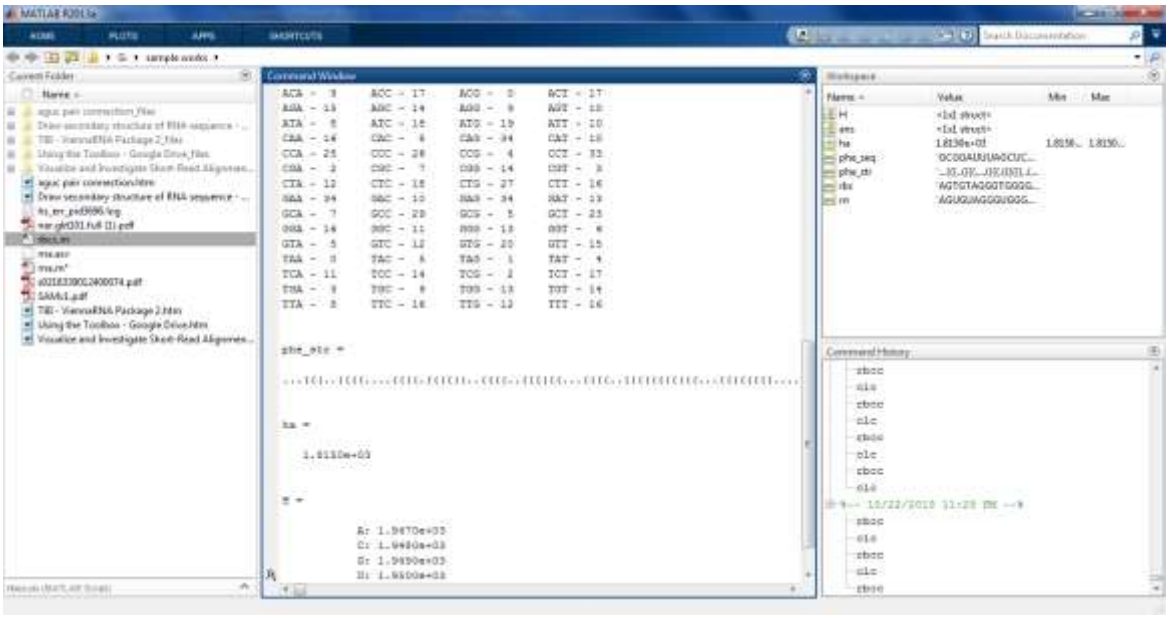


Fig 6 Triple Character Search Sequence in cancer affceted liver cells

CONCLUSION

In this paper ,the clustering algorithm can be applied both on cancer affected liver data sets and normal liver dataset. Using the clustering sequence alignment technique in data mining, the pair of sequence alignment within nucleotide, single and double character search alignment and combination pair of cells are analysed with running time and memory efficiency. In future, the research work will be extended into following direction. 1. Using the Dot Matrix Method graphical representations of nucleotide sequence are analysed. 2. Using the Hidden Markovo Model, the finite state machine model will be generated and analysed for nucleotide sequence for cancer affected liver cells and normal liver cells.

REFERENCES

[1] Survey of clustering data mining techniques. Pavel Berkhin.
[2] Comparison between clustering algorithm. Osama abu abbas.
[3] Supervised clustering algorithm and benefits. ChristopF.Eick.
[4] 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.
[5] 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.
[6] http://www.b3intelligence.com/NumericalDataMinig.html
[7] http://en.wikipedia.org/wiki/Numerical_analysis
[8] http://www.saedsayad.com/zeror.html
[9] http://www.cogsys.wiai.unibamberg.de/teaching/ss05/ml/slides/cogsysII-6.pdf
[10] http://www.slideshare.net/totoyou/covering-rulesbased-algorithm
[11] 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.
[12] Arun.K.Pujari"data mining techniques ", Universities Press (India) private limited.2001.
[13] F.Braz,"A review of the association rules data mining techniques for the analysis of gene expressions"
[14] Douglas Trewartha,"Investigating data mining in MATLAB ", Rhodes University 2006.