A Survey on Frequent Pattern Mining on DNA Sequence

Hole Vishakha G.[1] Londhe Komal R.[2] Nimbalkar Deepika S.[3] Shaikh Sajida H.[4] Prof. R. M. Kedar[5]

[1,2,3,4,5]Department of Computer Engineering

[1,2,3,4,5]KJCOEMR, SPPU, Pune, India

Abstract— Genes are an integral part of any organism's body. It contains information about the various different processes and materials found inside the body. The Human Genome is like a blueprint of the human body, the Genes contain valuable information that is utilized by the cells in replicating themselves. Due to the fact that the genes should be studied as they can unlock a lot of information about our own bodies. Therefore, frequent itemset mining is the most eligible candidate for this application. This is due to the fact that the frequent itemsets are generated when frequently occurring genes or the defective genes are being encountered together. The gene data is particularly huge and cannot be processed with traditional techniques. therefore, this survey paper studies the existing works on frequent itemset mining and try to evaluate their limitations, thereby practicing to introduce a new model of effective frequent itemset mining for the DNA Sequence.

Keywords: DNA, Frequent itemset Mining, DNA patterns, Pattern Mining

I. INTRODUCTION

A frequent itemset is a design that is encountered frequently within a dataset. A recurring itemset, which is made from one of these designs, is the reason why often mining is alternatively often called frequent itemset mining.

Frequent itemset mining is most likely explained by presenting market basket analysis, a typical usage for which it is widely known. Market basket analysis tries to locate associations or interrelations between the various items that have been selected by a specific shopper and placed in their market basket and assigns support and confidence measures for their differences.

The generalization of market basket analysis is frequent itemset mining and is actually quite equal to classification except that any entity or mixture of entities can be predicted in an organization. As organizations do not require the labeling of classes this belongs to the paradigm of unsupervised learning.

An important area of DNA research is in genetics and medical research. Due to our discovery of DNA, our ability to diagnose diseases really early has greatly improved. Additionally, geneticists are able to better assess a person's genetic susceptibility to specific diseases. For diseases that were previously considered fatal and where treatment was non-existent or largely unsuccessful, DNA discovery has inevitably led to medications and treatment for patients with severe illnesses.

DNA has been quite important in the field of forensic science. It is also important that victims can be identified, especially in cases where the victim's status is unfamiliar to family or friends. In this sense, DNA has been instrumental in bringing a revolution in the whole field of forensic science. This effect is felt within the criminal justice system and contributes to accuracy for the protection of society.

Microarray gene expression is the conversion of the DNA sequences into mRNA sequences by transcription then translated into amino acid sequences called proteins. Microarray technologies provide opportunities to compute the expression level of tens of thousands of genes in cells simultaneously. The data collected from microarray experiments is commonly in the form of an MXN matrix of expression level when representing columns. The total gene expression data can be valuable in understanding genes, cellular states, and biological networks. Analysis of this genomic data has two important goals. The first is to determine how expression of any particular gene might affect the expression of other genes. The second is to determine what genes are expressed as a result of certain cellular conditions. What genes are expressed in diseased cells that are not expressed in healthy cells? A Microarray database is a repository containing microarray gene expression data. The Key uses of a microarray database are to store the measurement data, manage searchable index and make the data available to other applications for analysis and interpretation.

This paper dedicates section II for analysis of existing work as literature survey, section III gives proposed system and section IV concludes the paper with feasible statement of the literature study.

II. LITERATURE SURVEY

This section of the literature survey eventually reveals some facts based on thoughtful analysis of many authors work as follows.

M. Zhang [1] explains that the DNA order and structure outline is very important for DNA nano applications. An exchange of perspectives in DNA sequence and structure design software tool is done by the proposed DNA shop and executed. Imagination tools can create DNA structures by specifying, choosing, and moving DNA sequences around and displaying the corresponding structures. A software tool called DNA shop, which is a point for interactive DNA sequence and structure design is detailed in this paper. The main aim of the proposed paper is to upgrade DNA structure design efficiency and have a finer idea about the designed DNA sequences.

X. Ma [2] estimates the identification of DNA-binding remnant in protein which makes it hard to understand the DNA interactions and other parameters for drug design. Thus, in the proposed paper, the prediction of DNABR (DNA Binding Residues) is offered for predicting DNA-binding residues by using random forest classification in combination with support sequence-based features. There are two types of novel sequence features first one is throwback information about the communication of physicochemical features of the amino acids and the second

is the correlation of amino acids between different sequence positions in terms of physicochemical properties. After the DNABR method, the upgraded RNA-binding residues in proteins can be also forecast with high accuracy.

K. Aeling [3] introduces that the lives of cells are developed by the interactions of the proteins and DNA. Gene expression, replication of DNA, DNA repair, and other vital cellular processes is managed by the Proteins that bind to specific locations on genomic DNA. Classifiers that were unskilled using the deformation energy are very important. Both direct and indirect recognition mechanisms sited are used to detect the proteins. The output of the proposed paper supports DNA deformation energy as an indirect recognition mechanism. Deformation energy predicts capacity for the underlying catalytic mechanism of DNA-binding enzymes.

Z. Elyazghi [4] elaborates exact order of nucleotides within a DNA molecule is determined by using DNA Sequencing. In basic biological research, the Knowledge of DNA sequences has become mandatory and there are various fields in the medical paradigm such as diagnosis, biotechnology, forensic biology, virology, genomics and biological systems where the knowledge of DNA is required. A four-color chromatogram is generated by sequencing and shows the result in sequencing run. Automated DNA sequencers build chromatogram files in ABI format in proposed paper.

B. Yimwadsana [5] narrates DNA computing is a method of computing using biological DNA it uses synthesized DNA sequences to form DNA structures that can carry out Boolean logic computation which is the primary form of computation for conventional computers. DNA logic circuit is a famous research matter in biological computing. In beginning there were simple DNA logic gates which were improved introduced later as DNA logic circuits. Building efficient logic gates from DNA materials is a very critical task because of the features of DNA, which consists of 4 different types of nucleotides and the chemical bonding possibilities between different sets of DNA nucleotides. The results of this paper help to cut the costs by running experiments in vitro.

P. Eaton [6] introduces nano diagnostics which is defined as the application of nanotechnology in materials, devices, or systems for diagnostics purposes. It is a rapidly growing field that has gained the interest of the people regarding these techniques. In this Single analyte molecules can be retrieved through one-to-one interrelation such as noble metal nanoparticles. Nucleic acid hybridization and recognition of complementary sequences are of interest and are broadly applied to nanoparticle-based detection systems. DNA was directly analyzed by atomic force microscopy (AFM). About 70% of the cases that are observed revealed that the Au nanoprobes had bound at the end of the AFM. Hybridization experiments of Au nanoprobes contain the AFM images efficiently.

A. Agrawal [7] discusses sequence alignment as one of the most frequent tasks in bioinformatics. Using substitute matrix protein sequence alignment can be easily used for DNA sequence alignment. As there 20 amino acids, therefore, the alphabet size of the protein sequence is 20 and the same for DNA sequence as there are four nucleotides. In the proposed paper DNA pairwise alignment which is a modified iterative method is implemented using multiple framework pairs with sequence-specific transition-transversion ratio. The proposed method has very good potential indicated by the preliminary experiments on a small synthetic dataset. Further experimentation with multiple parameter sets and much larger databases and comparing it with training can be significant part of the future work.

A. Bucur [8] elaborates the importance of genomic frequency patterns in biological events is greatly under survey mainly because there are few systematic tools accessible. For sequence alignment, clustering algorithms are currently used which are not suitable for spectral alignment and analysis. Thus, in the proposed paper, they have used the following spectral sorting method in order to achieve improvements such as Clearer patterns and Scaling. To lessen the human effort required for the analysis of the spectral images, the authors investigated the automatic data mining of the datasets in the Fourier space to detect relevant features. Easy detection of strong patterns in both single frequencies and multiple frequencies is proposed in paper.

K. Cheng [9] explains DNA sequences useful in areas such as biology, therapy, and genetics, DNA sequences have been gathered extensively in recent years. To reduce storage space and transmission load it is important to compress these sequences. A new challenge for DNA compression has introduced the advancement in sequencing techniques. Effective encoding can be done under repeated patterns within the DNA sequence. Proposed algorithms for multiple sequences compression are constructed on reference-based compression for a single sequence. In the future investigate complexity improvement in finding subsequence matches.

A. Haron [10] proposes the process of DNA sequence alignment that accelerates the process and presents the design and development of high-performance acceleration and optimization techniques. The author concentrates on memory and speed optimization by optimizing and mapping the DNA sequence data before alignment. This method has been designed and developed on a hardware-based acceleration device and targeted to Altera Cyclone II 2C70 FPGA and using 50MHz oscillator for clock source. Due to exponential increase in the amount of DNA sequences data, the DNA sequences alignment system from time to time experiences complexity. The proposed paper is verified by Data Optimization Technique that can be designed and implemented via FPGA plan to accelerate DNA sequences alignment process.

H. Zhou [11] explains the detection of Tandem Repeats in DNA Sequences on the base of parametric spectral estimation. The spectrogram of a DNA sequence is observed based on the autoregressive model. Repetitive DNA is divided into two parts first one is interspersed repeats, where repeat units are distributed in the genome in an apparently random fashion and the second one is tandem repeats, where repeat units are placed next to each other in an array. The purpose of DNA repeats is still mostly unsure. In gene regulation number of tandem repeats is related to diseases and plays an important role in their projection. The

output of the proposed paper shows that this method has a superior performance in comparison with other algorithms.

S. Ray [12] discusses the sequence alignment technique as the method of the layout sequences of DNA, RNA, or protein to recognize the area of similarity and to disclose the effect of functional, structural, or evolutionary relationships between the sequences. The process of alignment method could be classified into two broad parts i) pair-wise sequence alignment (PSA) and (ii) multiple sequence alignment (MSA). The main focus of PSA is to compute to what extent a particular set of genes or proteins are alike or to identify the belongingness of a particular sequence of interest (DNA, RNA or protein) to a pair of sequences. Local alignment and global alignment are two methods of PSA. The proposed technique is significantly better at differentiating than the technique based on Needleman-Wunsch even though it uses two matrices.

S. Liu [13] narrates that an immense amount of biological sequence data, is especially intricate in the progress of genome projects, demand the development of computational methods and tools to annotate, cluster, or describe their sequences, structures and functions corresponding to living processes such as transcriptions, regulatory factors, and translations. DNA binding site observing is the main issue in biology experiments as well as in computational methods. The authors are developing a Robust Mixed Effect Mixture Model (RMEMM) to find DNA binding sites that bind to specific transcription factors. Results show that the mean effect is similar to position-specific scoring matrices (PSSM), providing new insight into the sequence.

M. Shahzad [14] exploring DNA sequences is really important in the context of bioinformatics, as it maintains every function of the body being the building block. Bioinformaticians have flattened numerous characteristics of this sequence and it is probable that still there are more properties to discover within the DNA sequence. The DNA sequence is comprised of a long series of units called nucleotides and each of them holds a variable chemical constituent called a base. In the proposed paper software tool which can be helpful in traversing DNA features and analyzing it in different representations simultaneously or individually to discover new patterns.

M. Nuser [15] estimates that the speedy increase in the number of DNA sequences and the increased need to analyze these sequences and to find the similarity between these sequences has caused several techniques to be formulated and submitted by researchers. These techniques either allow insertions and deletions of elements in the DNA sequence or treat the DNA sequences as complete with no insertions or deletions. The authors presented a second technique for the similarity analysis of DNA sequences which is primitive discrimination substrings of sequence S and Q to define a new discrimination measure DM(S, Q) to observe the similarity such that the tinier the discrimination measure is, the more similar the species.

III. PROPOSED SYSTEM

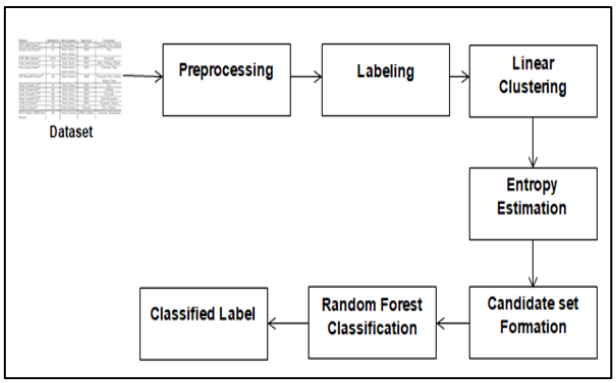

Fig. 1: Proposed System Architecture

1) Data Collection: This is the primitive step of the proposed model where a retail dataset is being collected from the publicly available repository kaggle.
2) Preprocessing: Once the dataset is collected it is converted into clean dataset for next steps.
3) Labelling: Labeling typically takes a set of unlabeled data and augments each piece of that unlabeled data with meaningful tags that are informative.
4) Linear Clustering: The preprocessed labelled list is subjected to hashing process to collect the unique patterns. Then for each of these unique patterns respective rows are being collected to create the cluster.
5) Entropy Estimation: Each of the unique pattern is counted for its presence in the number of the clusters. Then this count is used to estimate the 36distribution factor of the pattern using the Shannon information gain theory.This Shannon information gain theory yields a numerical value in between the 0 and 1. The value nearer to 1 indicates the pattern is more important and nearer to 0 indicates pattern is having less importance.
6) Candidate Set Formation: In this step candidate sets of the selected patterns are being generated based on the power set creation technique.
7) Random Forest Classification & Classified Labels: As the frequent patterns are being generated they are subject to the random forest tree classification to get the best outcome. In this process initial candidate set is consider as the right child ,if the value is bigger and as the left child if the value is smaller than the root value. Once the tree is formed, then this tree is traversed in the pre-order form to estimate the number of the nodes in each of the levels of the tree. Then the node serial numbers for the tree level that contain the maximum number of nodes or candidate set are considered as the classified labels.

IV. CONCLUSION

This paper has studied the various different techniques that have been published for the extraction of frequent item sets from a DNA sequence. The extraction of the frequent item sets is an important exercise in determining the various different forms of gene expression as a lot of diseases are genetic and their expression can hold clues towards early identification and preventive measures that can be performed to help the patient. Traditionally, for this purpose, the algorithms that are used have a very high time

complexity due to the use of horizontally computing algorithms such as Apriori, etc. which hinders a quick remedial action by the doctor. Therefore, to reduce the time taken for the extraction of frequent item sets the proposed methodology utilizes a tree-based architecture to shorten the time taken by implementing the Random forest Framework.

REFERENCES

[1] Mingjun Zhang, Chaman L. Sabharwal,Weimin Tao, Tzyh-Jong Tarn," Interactive DNA Sequence and Structure Design for DNA Nanoapplications" IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 3, NO. 4, DECEMBER 2004.

[2] Xin Ma, Jing Guo, Hong-De Liu, Jian-Ming Xie, and Xiao Sun," Sequence-Based Prediction of DNA-Binding Residues in Proteins with Conservation and Correlation Information" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 9, NO. 6, NOVEMBER/DECEMBER 2012.

[3] Kimberly A. Aeling, Nicholas R. Steffen, Matthew Johnson, G. Wesley Hatfield, Richard H. Lathrop, and Donald F. Senear," DNA Deformation Energy as an Indirect Recognition Mechanism in Protein-DNA Interactions " IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 4, NO. 1, JANUARY-MARCH 2007.

[4] Zakaria Elyazghi, Loubna El Yazouli, Khalid Sadki, and Fouzia Radouani," ABI Base Recall: Automatic Correction and Ends Trimming of DNA Sequences " IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 16, NO. 8, DECEMBER 2017.

[5] Boonsit Yimwadsana, Paramita Artiwet," On Optimizing DNA Sequence Design for DNA Logic AND Circuit" Proceedings of TENCON 2018 - 2018 IEEE Region 10 Conference (Jeju, Korea, 28-31 October 2018).

[6] Peter Eaton, Gonçalo Doria, Eulalia Pereira, Pedro Viana Baptista, and Ricardo Franco," Imaging Gold Nanoparticles for DNA Sequence Recognition in Biomedical Applications" IEEE transactions on nanobioscience, vol. 6, no. 4, december 2007.

[7] Ankit Agrawal, Xiaoqiu Huang, "Pairwise DNA Alignment with Sequence-Specific Transition-Transversion Ratio Using Multiple Parameter Sets", International Conference on Information Technology DOI 10.1109/ICIT.2008.62.

[8] Anca Bucur, Jasper van Leeuwen, Nevenka Dimitrova, and Chetan Mittal," Alignment Method for Spectrograms of DNA Sequences " IEEE Transactions On Information Technology In Biomedicine, VOL. 14, NO. 1, JANUARY 2010.

[9] Kin On Cheng, Paula Wu, Ngai Fong Law, and Wan Chi Siu," Compression of Multiple DNA Sequences Using Intra sequence and Inter sequence Similarities " IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, TCBB-2014-09-0365.

[10] Al Junid, S. A. M.; Haron, M.A.; Abd Majid, Z.; Osman, F.N.; Hashim, H.; Idris, M.F.M.; Dohad, M.R.," Optimization of DNA Sequences Data for Accelerate DNA Sequences Alignment on FPGA" Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation, 2010.

[11] Hongxia Zhou, Member, IEEE, Liping Du, and Hong Yan," Detection of Tandem Repeats in DNA Sequences Based on Parametric Spectral Estimation" IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 13, NO. 5, SEPTEMBER 2009

[12] Sanchita Saha Ray∗, Ananya Banerjee†, Anurupa Datta‡, Surajeet Ghosh§," A Memory Efficient DNA Sequence Alignment Technique Using Pointing Matrix " 978-1-5090-2597-8/16/c IEEE, 2016.

[13] Sheng Liu, Qing Song, Aize Cao, Xulei Yang, Yilei Wu," Robust Mixture Model Clustering of DNA Binding Sites " Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006

[14] M.Shahzad, Nazish Alia, Sadaf Mahmood," DNA Innovate: Visualizing DNA Sequences978-1-4244-4609-4/09/2009 IEEE

[15] Maryam Nuser, Izzat Alsmadi, "Evaluating graphical and statistical techniques for measuring similarity in DNA sequences", 978-1-4673-1550-0/12/,2012 IEEE

