iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 6 **Issue:** IX **Month of publication:** September 2018

**DOI:**

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887*
*Volume 6 Issue IX, Sep 2018- Available at www.ijraset.com*

# Review on Emerging Pattern Analysis using Gene Sequences

Nayan Shivhare[1], Mr. Vaibhav Chandrakar[2]
[1, 2] *Central College of Engineering and Management, Dept. of Computer Science and Engineering, Raipur, Chhattisgarh, India*

***Abstract: DNA microarrays permit the estimation of expression levels for a huge number of genes, perhaps all genes of an organism, inside various diverse experimental samples It is particularly imperative to extract biologically important data from this huge amount of expression data to know the present state of the cell because most cellular processes are regulated by changes in gene expression. Association rule mining techniques are useful to find association relationship between genes. Various association rule mining algorithms have been developed to analyze and associate this immense amount of gene expression data. This paper presents survey on some of the popular mining algorithms developed to analyze gene expression data.***

***Keywords: DNA Microarray, Gene expression, Data mining.***

## I. INTRODUCTION

GENE is a segment of DNA, which contains the formula for the chemical composition of one particular protein. Genes serve as the blueprints for proteins and some extra items, and mRNA is the first intermediate amid the production of any genetically encoded molecule. The concentration of a particular mRNA molecule is generally called the expression level of the respective gene, and it serves as an indicator of the measure of finished result that is right currently being produced. These days, the expression levels of thousands of genes, possibly all genes in an organism, can be measured simultaneously in a single investigation utilizing microarrays. This new innovation gives rise to present challenge: to interpret the meaning of this gigantic measure of biological data arranged in numerical matrices. To meet the difficulty, different strategies have been created utilizing both customary also, creative techniques to extract, analyze and visualize gene expression data generated from DNA microarrays. A critical advance in the investigation of quality articulation data is to find affiliation and connection between quality articulation designs.

## II. MICROARRAY TECHNOLOGY

Microarray is a technology which empowers the scientists to explore and address problems which were once thought to be non-traceable. Microarray innovation has enabled the scientific community to comprehend the essential perspectives underlining the growth and development of life and also to investigate the genetic reasons of anomalies happening in the working of the human body.

The fundamental principle underlying microarray innovation is that complementary nucleic acids will hybridize. This is too the reason for customary gene expression examinations, for example, Southern and Northern blotting. Hybridization gives dazzling selectivity of reciprocal stranded nucleic acids, with high affectability and specificity. In the standard strategies, in which radioactive marking materials are as a general rule used, the synchronous hybridization of test and reference test is outlandish.

## III. DNA MICROARRAY EXPERIMENT

A DNA chip is the instrument that measures simultaneously the concentration of thousands of mRNA molecules. It moreover alludes to as a DNA microarray. They can measure simultaneously the expression levels of up to 20,000 genes.

The DNA microarrays are produced as follows:

Divide a glass or silica plate of 1 cm over (the chip) into pixels. Here every pixel will be dedicated to one gene. Millions of 25 base pair long single strand DNA, replicated from a particular segment of gene, is combined on the dedicated pixel. These are called probs. The mRNA molecules are extracted from the cell taken from tissue of interest, (for example, cancer tissue). They are reversed transcribed from RNA to DNA and their concentration is improved. At that point the subsequent DNA is interpreted again into fluorescently marked single strand RNA. The arrangement of marked and mRNA molecules (copies of the mRNA molecules that were initially removed from the tissue) is put on the chip and named RNA diffuse over the thick forest of single strand DNA tests. Whenever such an mRNA experiences a bit of the test, of which the RNA is the ideal copy, it joins to it with high fondness which is called hybridization. After the mRNA arrangement is washed off, just those molecules that discovered their ideal match

©IJRASET: All Rights are Reserved

remain fixed to the chip. Presently the chip is enlightened with a laser, and those fixed targets fluoresce. By estimating the light radiating from every pixel, one gets a measure of focuses on that stuck. It is proportional to the centralization of those mRNA in the examined tissue.

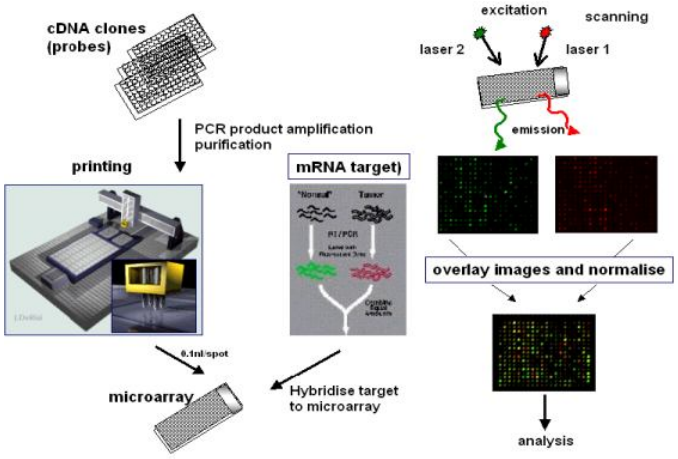



Fig. 1. DNA microarray experiment

## IV. GENE EXPRESSION DATA

A typical DNA microarray experiments gives the expression profiles of a few several examples (say Ns % 100), more than a few thousand (Ng) genes. These results are plot in a Ng×Ns articulation table; every section identifies with one specific quality and every segment to a test. Passage Egs of such an articulation table stands for the articulation level of quality g in test s. The first quality articulation framework obtained from a checking procedure contains commotion, missing characteristics, and precise varieties rising up out of the test method. Information pre-preparing is indispensable before any affiliation rules investigation can be performed.

### *A. Analysis of Gene Expression Data*

Analysis of gene expression information helps the atomic researcher in numerous angles, such as, gathering data about distinctive cell states, functioning of genes, identifying genes that reflect biological process of interest and so on. A few self-evident implications of gene expression information analysis are the accompanying:

*1)* Identify genes whose expression levels reflect biological processes of interest, (for example, development of cancers).

*2)* Group the tumors into classes that can be separated on the premise of their expression profiles, possibly as it were that can be interpreted as far as clinical classification. In the event that one can partition tumors, based on their expression levels into relevant classes, (for example, e.g. positive versus negative responders to a particular treatment), the classification got from expression analysis can be utilized as a diagnostic and therapeutic tool.

*3)* Finally the analysis can give clues and guesses to the function of genes (proteins) of yet unknown role.

## V. ASSOCIATION RULES

Association rule mining finds interesting association and correlation relationships among a substantial arrangement of daat items [1]. The rules are viewed as interesting on the off chance that they satisfy both a minimum support threshold and minimum confidence threshold [2]. The most widely recognized way to deal with finding association rules is to separate the issue into two sections [3].

*1)* Find frequent thing sets: By definition, each of these thing sets will happen in any event as frequently as a pre-determined minimum support count [1].

*2)* Generate in number affiliation rules from the successive thing sets: By definition, these principles must fulfill least help what's more, least assurance [1].

The second step is less demanding of the two. The all-around execution of mining affiliation rules is dictated by the underlying advance. As appeared in [4], the execution, for vast databases, is most impacted by the combinatorial explosion of the quantity of conceivable frequent item sets that must be considered and furthermore by the quantity of database scans that has to be performed. Numerous conventional association rule mining algorithms, (for example, A priori [5], FP-growth [4], DynFPgrowth [6],

©IJRASET: All Rights are Reserved

Partitioning, Dynamic Item set counting (DIC), Direct Hashing also, Pruning DHP and so forth.) have been adopted or on the other hand specifically connected to gene expression data. These association rules mining algorithms have been demonstrated helpful for identifying biologically relevant association among the genes.

*A. Significance of association rule mining Techniques in Gene Expression*

Utilizing association rule mining approach, we can analyze:

*1)* The expression of one gene prompts the enlistment of a serial of target gene expressions. This expression pattern is signified control of gene expression. The relationship between one gene and the other target genes can be seen as an associative relation.

*2)* A few gene expressions prompt the expression of one target gene. Transcription factors and their target gene is one of numerous cases in this classification.

*3)* Gene expression prompts the induction of new biological function.

## VI. LITERATURE SURVEY

Baralis et al. [7], presents a novel method to manage discovering quality relationships from GEDs which does not require information discretization. By speaking to per-quality articulation regards as thing weights, visit weighted itemsets can be separated. The revelation of weighted itemsets as opposed to regular (not weighted) ones shields specialists from discretizing GEDs before examining them and therefore enhances the ampleness of the learning disclosure process. Tests performed on honest to goodness GEDs show the ampleness of the proposed approach.

S. Ji et al. [8], propose a novel effective calculation FCPminer to mine best k visit shut examples (FCPs) of higher help with length no not as much as minL from quality articulation data. FCPminer uses a prefix fp-tree data structure, with top-down best first hunt methodology, to such a degree, to the point that FCPs of satisfactory length with most hoisted underpins are initially mined.

S. Mishra et al. [9], the successive examples gained are considered as the plan of starting populace.

For the determination criteria, we had considered the mean squared deposit score rather using the edge esteem. It was seen that out of the four fluffy based continuous mining methods, the PSO based fluffy FP development procedure finds the best individual incessant examples. Furthermore, the run time of the calculation and the amount of regular examples created is far better than whatever is left of the strategies used.

S. Alagukumar et al. [10], Associative Classification techniques are utilized to make better decision in basic circumstances. The proposed associative classification called as Classification of microarray gene expression information utilizing associative classification and gene expression intervals used to arrange the gene expression with gene intervals in influenced gene expression. The experimental results are completed by utilizing the gene expression of breast cancer. The associative classification on gene expression information acquired the best prediction and accuracy of the classification result.

V. Rajput et al. [11], Author plan a novel model for forecast the dengue sickness. Here, we utilize genetic algorithm to figure the actual weight of attributes afterwards applied the FP-Growth with actual weight.

Theoretical investigation and experiments have shown that the changed approach can detect the virtual significance of attributes in requirements of their weights.

This model are ponder and the parameters are set to get ideal forecast execution.

M. Khashei et al. [12], a new hybrid model of artificial neural systems is proposed as an elective classification model for situations where inadequate information are accessible, utilizing the unique soft computing of the fuzzy logic. A hierarchical version of the proposed model is produced by analyzing three distinct methodologies including "one versus one", "one versus rest", and "one versus all".

Among these methodologies, the "one versus all" approach yield more accurate outcomes and apply for constructing the hierarchical version of the proposed model.

©IJRASET: All Rights are Reserved

Table I. Shows comparisons of existing methods and its features

| Ref. No. | Author Name/year | Dataset used | Method | Findings | Conclusion |
|---|---|---|---|---|---|
| 7 | E. Baralis et al, 2013 | T-ALL, and BRC-ABL | weighted item set mining algorithm | Author propose to consider gene expression values as item weights, which indicate gene expression intensity within each sample, and apply a weighted itemset mining algorithm [20] directly to non-discretized GED. | The experimental results show the applicability and usefulness of the proposed approach on real GEDs. |
| 8 | S. Ji et al, 2014 | ALL-AML leukemia dataset | FCPminer algorithm | Author addresses the problem of mining top-k frequent closed patterns of minL pattern length from gene expression data. A novel algorithm FCPminer is proposed with prefix fptree data structure and best first search strategy. | Experimental results on real microarray data and synthetic data show that, the proposed FCPminer is much more efficient than state-of-the-art top-k frequent closed pattern mining algorithms. |
| 9 | S. Mishra, et al, 2011 | Fuzzy dataset | FP-growth algorithm | Author considered the fuzzified dataset and have implemented various frequent pattern mining techniques. Out of the various frequent pattern mining techniques it was found that Frequent Pattern (FP) growth method yields us better results on a fuzzy dataset. | FP growth method it even yields us much better results as expected to discover perfect patterns |
| 10 | S. Alagukumar, et al, 2016 | breast cancer gene expression dataset | Associative Classification techniques | The experimental results carried out by using breast cancer gene expression data which are available on the NCBI online biological database. It has been tried with two class and multi-class datasets and contrasted and the traditional classification calculations, for example, Linear Discriminant Analysis, SVM, and Decision Tree. | The experimental results are carried out by using the gene expression of breast cancer. The associative classification on gene expression data obtained the best prediction and accuracy of the classification result. |
| 11 | V. Rajput, et al, 2017 | real dataset | genetic algorithm | use genetic algorithm to calculate the actual weight of attributes afterwards applied the FP-Growth with actual weight. Theoretical study and experiments have displayed that the modified approach is able to detect the virtual significance of attributes in requisites of their weights. | The outcome displays that the model produces the better prediction. |
| 12 | M. Khashei, et al, 2012 | microarray dataset | Artificial neural network | A new hybrid model of artificial neural networks is proposed as an alternative classification model for cases where inadequate data are available, using the unique soft computing advantages of the fuzzy logic. | The obtained results indicate that the proposed model to be superior to all alternative models in both training and test data sets. |
| 13 | S. Ramos, et al, 2010 | Microarray dataset | Bayesian classification | Results on microarray data sets (Leukemia, Prostate and Breast) show that BOSc performance is competitive with, and in some cases significantly better than, quadratic and linear discriminant analyses and support vector machines classifiers. | It allows also the incorporation of prior information on the prevalence of the disease or type of disease in the model, improving the performance of the classifier. |

©IJRASET: All Rights are Reserved

## VII. CONCLUSION

Microarrays have become a standard research tool for present laboratory. Microarray analysis has been used successfully characterize transcriptional signatures to allow for patient-tailored therapy strategy in breast cancer or to classify better tumors having no histological counterparts in normal tissues. It very well may be a helpful instrument to distinguish genes straightforwardly actuated or curbed by articulation of an interpretation factor. In this manner, essential reaction genes can be identified by computational seeking of consider particular responsive components a DNA district found upstream of genes observed to be differentially communicated in microarray tests. In any case, every one of these things and conceivable outcomes rely upon proficient and appropriate investigation of quality articulation information. This paper presents study on different calculations created to break down quality articulation information.

## REFERENCES

[1] J. Han M.Kamber, “Data Mining Concepts and Techniques”. Morgan Kaufmann Publishers, San Francisco, USA, 2001, ISBN 1558604898.

[2] C.Gyorodi, R.Gyorodi. “Mining Association rules in Large Databases”. Proc. of Oradea EMES’02: 45-50, Oradea, Romania, 2002

[3] M.H. Dunham. “Data Mining – Introductory and Advanced Topics”. Prentice Hall, 2003, ISBN 0-13-088892-3.

[4] J.Han, J.Pei, Y.Yin, “Mining Frequent Patterns without candidate generation”. Proc. Of ACM-SIGMOD, 2000.

[5] R.Agrawal, R.Srikant, “Fast algorithms for mining association rules in large databases”. Proc. of 20th Int’l conf. on VLDB: 487-499, 1994.

[6] C.Gyorodi, R.Gyorodi, T.Cofeey & S.Holban – “Mining association rules using Dynamic FP-Trees” – in Proc. of The Irish signal and Systems Conference, University of Limerick, Limerick, Ireland, 30th June- 2nd July 2003, ISBN 0-9542973-1-8, page 76-82.

[7] E. Baralis, L. Cagliero, T. Cerquitelli, S. Chiusano and P. Garza, "Frequent weighted itemset mining from gene expression data," 13th IEEE International Conference on BioInformatics and BioEngineering, Chania, 2013, pp. 1-4.

[8] S. Ji, X. Wang, Y. Zong and X. Gao, "Mining Top-K Frequent Closed Patterns from Gene Expression Data," 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, 2014, pp. 732-739.

[9] S. Mishra, D. Mishra and S. K. Satapathy, "Particle swarm optimization based fuzzy frequent pattern mining from gene expression data," 2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011), Allahabad, 2011, pp. 15-20.

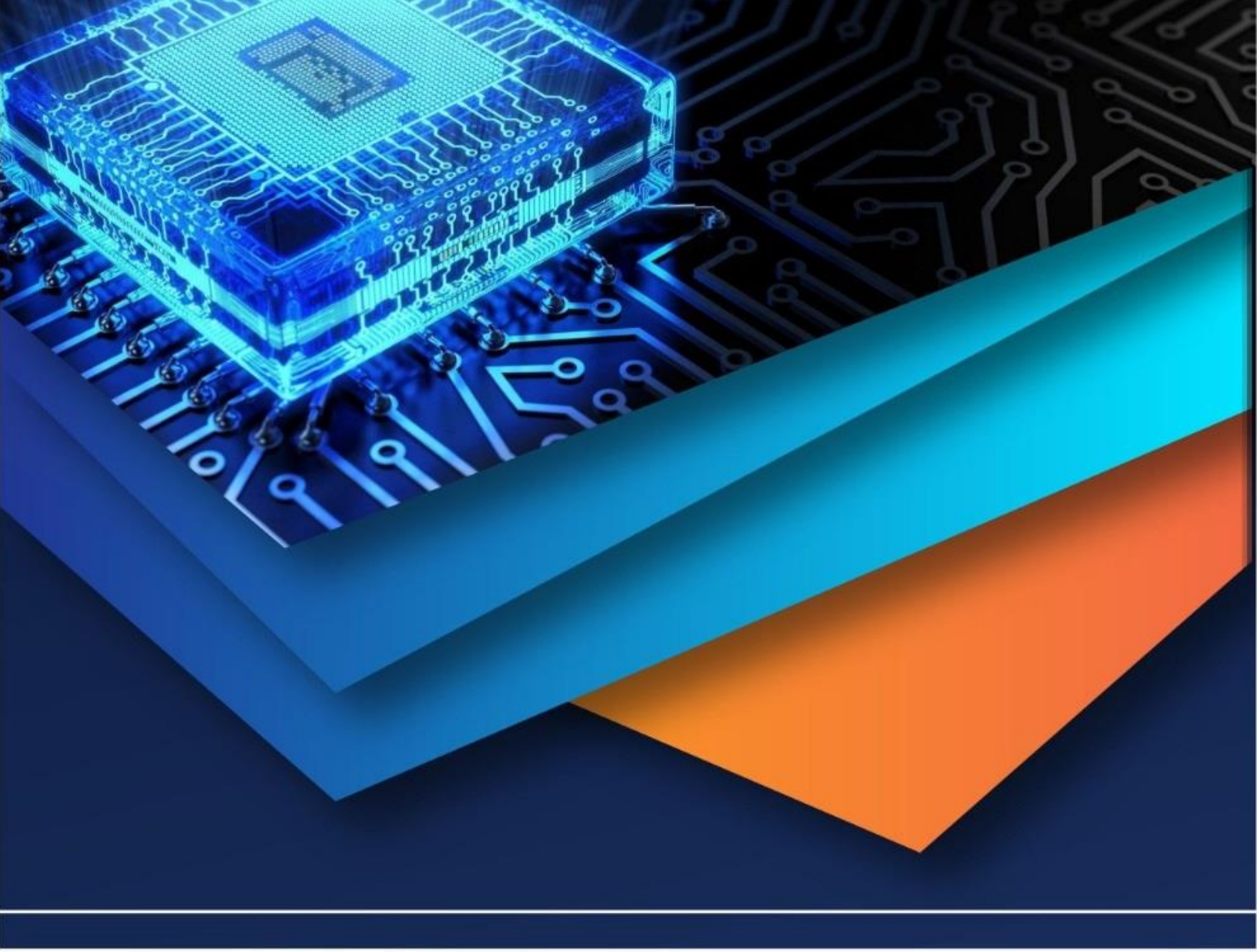
[10] S. Alagukumar and R. Lawrance, "Classification of microarray gene expression data using associative classification," 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, 2016, pp. 1-8.

[11] V. Rajput and A. Manjhvar, "A model for forecasting dengue disease using genetic based weighted FP-growth," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, 2017, pp. 944-948.

[12] M. Khashei, Z. H. Ali and M. Bijari, “A fuzzy intelligent approach to the classification problem in gene expression data analysis”, Elsevier 2012.

[13] S. Ramos a,c, A. T. Antónia, A. Marília, “Bayesian classification for bivariate normal gene expression”, 2010 Elsevier.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)