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*

# Emerging Pattern Analysis using Microarray Gene Sequences for Cancers

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:** Recent investigations on sub-atomic level classification of tissues have delivered striking outcomes, and demonstrated that gene expression measures could altogether help in the advancement of proficient cancer determination. Cancer classification in light of the DNA exhibit information is as yet a troublesome issue. The fundamental test is the staggering number of genes in respect to the quantity of preparing tests. It makes exact grouping of information more troublesome. This paper applies FP-Growth algorithm with some feature discretization approach in order to group similar gene data together.*
***Keywords:** DNA Microarray, Gene expression, Data mining, FP Growth, Feature discretization.*

## I. INTRODUCTION

Cancer classification through gene articulation information investigation has risen as a functioning region of research as of late. Gene articulation shows up during the time spent interpreting a gene's deoxyribonucleic acid (DNA) sequence into ribonucleic acid (RNA). A gene's appearance level shows the estimated number of duplicates of that gene's RNA created in a cell and it is related with the measure of the comparing proteins made [1]. The capacity to screen gene articulation at the transcript level has turned out to be conceivable because of the rise of DNA microarray advances which are utilized for estimating a huge number of genome wide articulation esteems in parallel. This microarray is a glass slide, onto which single-stranded DNA atoms are connected at settled spots. There might be a huge number of spots on an exhibit, each identified with a solitary gene. Investigation and giving of such information is getting to be one of the real assignments in the usage of the microarray innovation [2].

For instance, tumors are typically reflected in the difference in the articulation estimations of specific genes. Late examinations on sub-atomic level grouping of tissues have delivered amazing outcomes, and showed that gene articulation measures could altogether help in the improvement of proficient malignancy determination and characterization stages [3]. The principle challenge is the mind-boggling number of genes with respect to the quantity of preparing tests. This infers we should manage an immense number of superfluous genes. A proficient calculation is required to diminish the computational overhead. The other test is from the nearness of commotion natural in the informational collection. It makes exact grouping of information more troublesome when the example estimate is little.

## 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 exquisite selectivity of complementary stranded nucleic acids, with high sensitivity and specificity. In the customary techniques, in which radioactive labeling materials are more often than not utilized, the simultaneous hybridization of test and reference test is impossible.

## III. GENETIC ALGORITHM

The task of GA begins with a set of arrangements called population. Arrangements from one populace are taken and used to frame another populace. This is persuaded by an expectation that the new populace will be superior to the old one. Arrangements which are chosen to shape new arrangements are chosen by their wellness - the more appropriate they are, the more possibilities they need to replicate [5].

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887*
*Volume 6 Issue IX, Sep 2018- Available at www.ijraset.com*

*A. Chromosome Encoding*

The chromosome is composed as a vector with 1× n components so that:

$$chromosome\ x = [x_1, x_2, \cdots, x_n]$$

*B. Selection*

Selection begins with picking parents from the populace, a procedure achieved with a one-sided roulette wheel on which every chromosome, characterized as:

$$P_j = \frac{fitness\ (x)_j}{\sum_{i=1}^{N} fitness\ (x)_i} \qquad j = 1,2 \cdots N$$

*C. Crossover*

This exploration utilizes the two-point gene trade for hybrid activity. The hybrid rate is expected to be Pcr. The hybrid point is haphazardly assigned by:

$$\theta_i = ceil(random * n), i = 1,2.$$

*D. Mutation*

After a hybrid activity is performed, transformation happens. In this investigation, singular genes of new posterity are changed haphazardly with likelihood Pmr. The transformation point is characterized as

$$\delta = ceil(random \times n).$$

## IV. ASSOCIATION RULE

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 strong association rules from the frequent thing sets: By definition, these rules must satisfy minimum support what's more, minimum certainty.

The second step is easier of the two. The by and large execution of mining association rules is determined by the initial step. 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], 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.

## V. LITERATURE SURVEY

E. Baralis et al. [7], presents a novel way to deal with finding gene correlations from GEDs which does not require data discretization. By representing per-gene expression esteems as item weights, frequent weighted itemsets can be extracted. The discovery of weighted itemsets rather than conventional (not weighted) ones keeps experts from discretizing GEDs before analyzing them and consequently improves the adequacy of the knowledge discovery process. Experiments performed on genuine GEDs exhibit the adequacy of the proposed approach.

S. Ji et al. [8], propose a novel efficient algorithm FCPminer to mine best k frequent closed patterns (FCPs) of higher support with length no not as much as minL from gene expression information. FCPminer utilizes a prefix fp-tree information structure, with top-down best first search strategy, to such an extent that FCPs of adequate length with most elevated supports are firstly mined.

S. Mishra et al. [9], the frequent patterns acquired are considered as the arrangement of initial population. For the selection criteria, we had considered the mean squared residue score rather utilizing the threshold value. It was observed that out of the four fuzzy based frequent mining techniques, the PSO based fuzzy FP growth technique finds the best individual frequent patterns. Additionally, the run time of the algorithm and the quantity of frequent patterns generated is far superior to the rest of the techniques utilized.

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.

## VI. METHODOLOGY

In this section we present our proposed methodology in detail.

The proposed framework is shown in fig. 1.

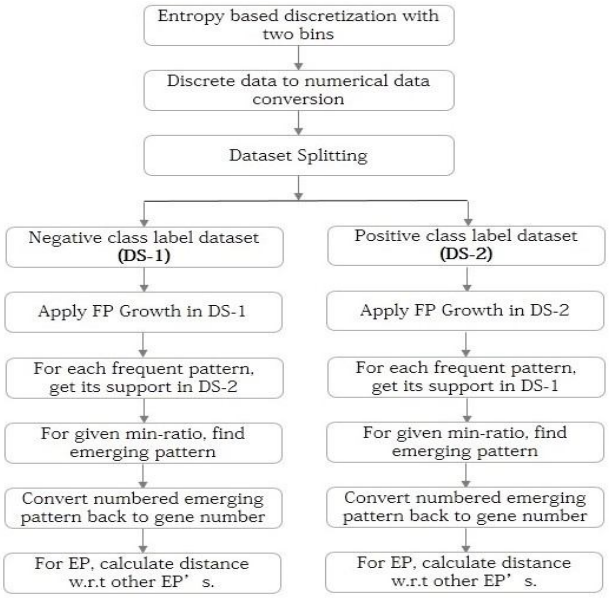



Fig. 1. Proposed system architecture for emerging pattern analysis

The proposed framework consist of various modules. Such as:

*1)* Entropy based discretization
*2)* Gene data conversion
*3)* Data spilling into positive and negative labels
*4)* FP Growth algorithm implementation
*5)* Emerging pattern recognition

### *A. File Scan And Entropy Based Discretization*

The dataset is scanned once to get the number of rows or tuples and columns or attributes. Entropy based discretization is performed for the given data with two bins. The discretization data is converted to numbered data, wherein the two states of a gene low or high are represented by numbers 2i - 1 and 2i. Where 'i' being the gene number.

### *B. Finding Frequent Pattern Using Fp-Growth Algorithm*

The numbered data is split into two datasets, one with tuples containing positive class label and other dataset with tuples containing negative class label. FP growth algorithm is applied to these two datasets to find frequent patterns, with given minimum support threshold.

### *C. FP Growth*

FP-tree growth adopts a divide-and-conquer that mines the entire arrangement of frequent item sets without applicant generation. FP growth calculation builds the contingent frequent pattern (FP) - tree and plays out the mining on this tree. FP-tree is an expanded prefix tree structure, putting away pivotal and quantitative data about frequent sets. The tree hubs are frequent things and are orchestrated such that all the more frequently happening hubs will have preferred odds of sharing hubs over the less frequently happening ones. The strategy begins from frequent 1-itemsets as an underlying addition pattern and looks at just its contingent pattern base (a subset of the database), which comprises of set of frequent things co-happening with the postfix pattern. The calculation includes two stages.

## VII. RESULT

In this section we present results performed using gene expression dataset.

When the top 35 genes based on information gain order are considered for mining frequent patterns using FP-growth with,

*1)* Minimum support of 0.6 – to find frequent patterns
*2)* Minimum ratio of 20 – to find emerging patterns
*3)* Minimum threshold of 0.004 for obtaining gene sets

The following gene sets H that are strongly correlated with the corresponding class label were obtained.

TABLE I. Strongly correlated labeled class

| Genesets strongly correlated with negative class label | Genesets strongly correlated with positive class label |
|---|---|
| {1227, 822} | {652, 249} |
| {964, 765} | {1325, 493} |
| {897, 513} | {642, 66} |
| {249, 513, 822} | {399, 249, 493} |
| {415, 513, 822} | {1153, 1325, 493} |
| {1060, 1227, 822} | {1325, 780} |

TABLE II. Emerging patterns based on item sets

| Emerging patterns – negative class label | Total distance of pattern | Emerging patterns – positive class label | Total distance of pattern |
|---|---|---|---|
| {652+, 1060+} | 1424.65 | {1325-, 897+, 652-, 377+} | 1222.03 |
| {1325+, 415-} | 1388.53 | {780-, 1325-, 897+, 377+} | 1221.76 |
| {964+, 66-} | 1378.08 | {964-, 493+, 1325-, 467-} | 1195.37 |
| {467+, 513+} | 1375.68 | {964-, 493+, 1325-, 43-} | 1191.54 |
| {964+, 493-} | 1370.35 | {780-, 1325-, 897+, 1047-} | 1214.17 |

Table I and II shows the outcome of proposed framework. The distance of positive label set and negative label set are shown. From this experiment the similar genes are extracted which can in future cause cancer.

## VIII. 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. In this paper we applied FP-Growth algorithm in combination with feature discretization in order to mine the emerging patterns for the cancer disease gene data. Table I and II shows the outcomes of extracted gene sequences which are having minimum distances from each other.

## 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.

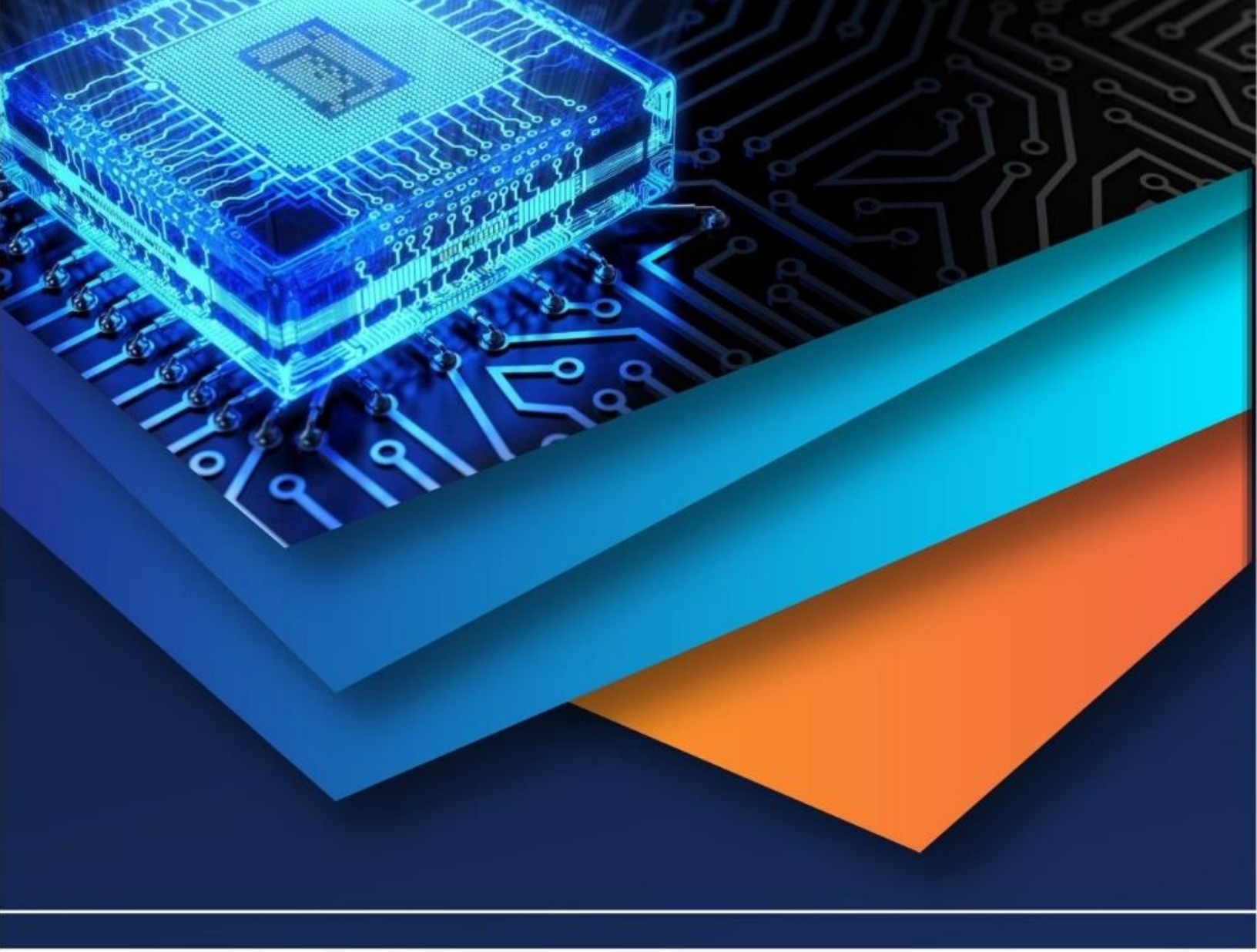
[12] M. Khashei, Z. H. Ali and M. Bijari, “A fuzzy intelligent approach to the classification problem in gene expression data analysis”, Elsevier 2012.

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

[14] R.Agrawal, T.Imielinki and A.Swami, “Mining association rules between set of item of large databases” in Proc. Of the ACM SIGMOD Intl’l Conf. on Management of data, Washington, D.C.,USA, 1993, pp 207-216.

[15] M.Anandhavalli Member, IACSIT, IAENG, M.K.Ghose, K.Gauthaman.” Association Rule Mining in Genomics” International Journal of Computer Theory and Engineering, Vol. 2, No. 2 April, 2010. 1793-8201 pages:12-15.

[16] Chad Creighton, Samir Hanash. Bioinformatics Program and 2Pediatrics and Communicable Diseases, University of Michigan, Ann Arbor, MI 48109, USA.” Mining gene expression databases for association rules”. Received on April 19, 2002.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY