Algorithm for Microarray Cancer Data Analysis using Frequent Pattern Mining and Gene Intervals

Alagukumar. S
Research Scholar in Computer Science,
Ayya Nadar Janaki Ammal College,
Sivakasi – 626124,
Tamil Nadu, India

Lawrance. R
Director, Department of Computer Applications,
Ayya Nadar Janaki Ammal College,
Sivakasi – 626124,
Tamil Nadu, India

ABSTRACT
Microarray technology allows for the simultaneously monitor of expression levels for thousands of genes or entire genomes. Diseases are often controlled by groups of genes, rather than individual ones. Association rule mining technique in data mining plays a vital role in the field of bioinformatics. In this paper, it has been proposed a novel approach for analysis of microarray gene expression profiling data. It discovers frequent patterns, expressions profiles using transcript expression intervals and extract significant relations among microarray genes. It is important to get efficient and important patterns to reveal fatal and crucial reasons for diseases. It provides improving prediction for diseases and treatment decisions for cancer patients.

Keywords
Data mining, gene expression analysis, frequent pattern mining, gene expression analysis using gene intervals.

1. INTRODUCTION
Microarray technologies provide a powerful tool by which the expression patterns of thousands of genes can be monitored simultaneously whose application range from cancer diagnosis to drug response. 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 the opportunity to compute the expression level of tens of thousands of genes in cells simultaneously.

The expression level is associated with the corresponding protein made under different conditions. Microarray experiments produced large volume of data. Microarray data presents the main challenge that is high density of data. The data collected from a microarray experiments is commonly in the form of an M x N matrix of expression level, where M represents columns(genes) and N represents rows(samples). In this paper it has been presented a novel approach, it has been focused on analysis of microarray gene expression profiling data. Frequent pattern mining is the most important task of association rules mining methods. It discovers frequent patterns, expressions profiles using transcript expression intervals and extract significant relations among microarray genes.

The structure of the paper is organized as follows. Section 2 reviewed past works in this field. Section 3 proposed the methodology of frequent pattern mining for gene expression with intervals. In section 4 it has been presented that the experimental results. Finally, Conclusion and Future work are explained in section 5.

2. RELATED WORKS
Association Rule Mining [1], has become one of the vital role in data mining tasks as well as computational biology. Association rule mining is an unsupervised data mining technique, which produces understandable rules.

Definition 1 (Association Rule) Let $I = \{i_1, i_2, i_3 \dots i_m\}$ be a set of m elements called items [2]. A rule is defined as an implication of the form $X \rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$ [2]. The left-hand side of the rule is named as antecedent and right-side of the rule is named as consequent.

The basic task of mining association rules technique extract interesting relationships among set of items. For instance, an association rules between genes in the form of $gene_1[-inf:2.03] \rightarrow gene_2[0.3,0.87]$, $gene_3[0.54,0.89]$ which means gene1 is expressed it is also very likely to observe an expression of $gene_2$ and $gene_3$.In general, every association rule must satisfy both support and confidence values. So, the target is to generate all association rules that satisfy user threshold minimum support and confidence values.

Becquet *et al*., [5] have analyzed and extracted collections of rules indicated that a very strong co-regulation of mRNA encoding ribosomal proteins occurs in the dataset.

McIntosh *et al*., [6] have proposed algorithm is a support-free algorithm and mining confidence rules, which describe interesting gene relationships from microarray data sets.

Zakaria, W., *et al.,* [7] have proposed algorithm based on the column (gene) enumeration method and mining association rules for up/down-expressed genes in microarray dataset.

3. METHODOLOGY
Mining association rules is currently a vital data mining technique for many applications [1,2]. Mining association rules technique is applied to microarray dataset to extract interesting relationships among sets of genes [5,6,7].

Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational data sets [2]. An instance of frequent itemset mining is market basket analysis. The aim of association rule mining is, to extract the frequent patterns using gene expression intervals. Before mining frequent patterns gene expression data converted continuous values into discretized values and discretized values are substituted by gene intervals. The discretized gene expression data converted into transaction data to discover frequent rule items from gene intervals. The proposed association rules using gene intervals system shown in figure 1.

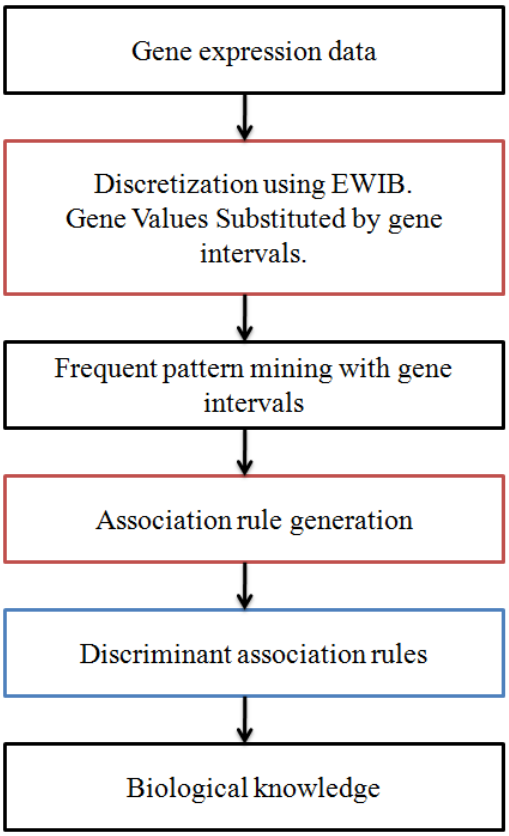

Fig 1: The association rules using gene intervals system

3.1 Data formats

The microarray gene expression dataset can be form of an M x N matrix D of expression values, where the row represents samples S={ s_1,s_2,s_3… s_n }and column represents genes G= {g_1,g_2,g_3…, g_n}, An illustration of microarray gene data shown in Table 1. The matrix usually contains large amount of data [9], therefore data mining techniques are used to extract useful knowledge.

Table 1. Microarray Data

Samples	Attributes(genes)			
	Gene1	Gene2	…	Gene m
1	G(1,1)	G(1,2)	…	G(1,m)
2	G(2,1)	G(2,2)	…	G(2,m)
3	G(3,1)	G(3,2)	…	G(3,m)
…	…	…	…	…
…	…	…	…	…
n	G(n,1)	G(n,2)	…	G(n ,m)

3.2 Discretization

The discretization process transforms quantitative data into qualitative data[3], that is, numerical attributes into discrete or nominal attributes with a finite number of intervals, obtaining a non overlapping partition of a continuous domain [3,4]. Discretization techniques are often used by the rule generation algorithms and a wide range of learning algorithms. Use of discrete values has a number of advantages, discrete features require less memory, and learning will be more accurate and faster using the discrete features.

An instance is the transformation of a continuous variable such as height of a human from a numerical measure into tall, short and medium categories.

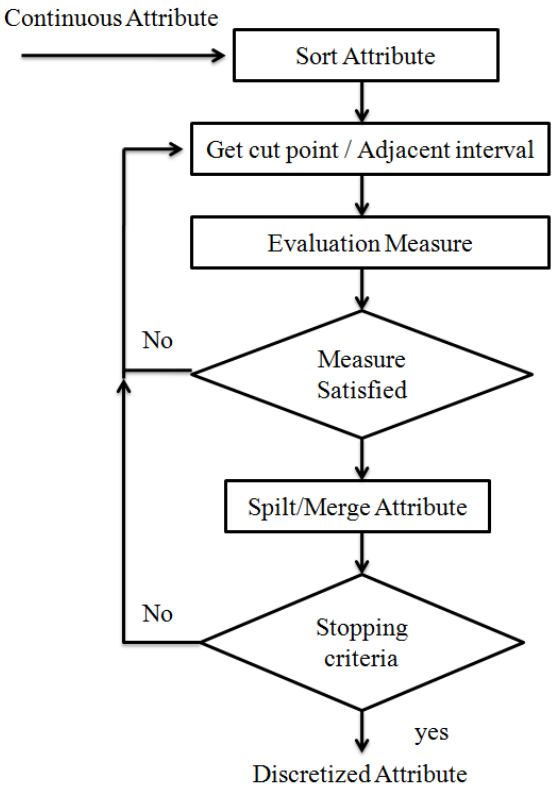

Fig 2: Discretization process

Liu H. *et al*.,[4] have formally defined a discretization process generally consists of four steps such as, sorting the continuous values of the feature to be discretized, evaluating a cut-point for splitting or adjacent intervals for merging, according to some criterion, splitting or merging intervals of continuous value, and finally stopping at some point.

Discretization methods can be supervised or unsupervised depending on data sets. Supervised methods make use of the knowledge of class label when partitioning the continuous features. While, unsupervised discretization methods without the knowledge of class label to discretize continuous attributes. Supervised discretization can be further characterized as error-based, entropy-based or statistics based. Unsupervised discretization can be characterized as equal-width and equal-frequency binning methods.

Discretization methods can also be grouped in terms of top-down or bottom-up. In a top-down approach, the intervals are split while for a bottom-up approach, the intervals are merged when discretization.

3.2.1 Equal-width interval bin method

Equal-width interval bin discretization is a simplest discretization method that divides the range of observed values for genes into k equal sized bins, where k is a parameter, which is provided by the user. The process involves sorting the observed values of gene expression and finding the minimum, V_{min} and maximum, V_{max} values. The interval can be computed by dividing the range of observed values for the variable into k equally sized bins using following formula. The gene expressions values with their specific intervals separated by the cutting points. For instance $gene_1$[-inf, cutting point] and $gene_2$[cutting point, +inf].

$$\text{Interval} = \frac{V_{max} - V_{min}}{k}$$

$$\text{Boundaries} = V_{min} + (i * \text{interval})$$

The Number of bins fixed in equal-width interval bin method, there is no need for any stopping criterion. Then the boundaries can be constructed for i= 1….. k-1 using the above equation[4].

3.3 Frequent pattern mining with gene interval

Definition 2 (Frequent item set) Given a set of items I= {i_1, i_2, i_3 … i_n} and a set of transaction T = {t_1, t_2, t_3 … t_m}, a subset of I, S⊆I is called a frequent, if support(S) ≥ minimum support, where minimum support is a user defined threshold [2].

3.4 Association Rule Generation

Definition3 (Support of rule) The rule X→Y holds in the transaction set T with support s, where s is the percentage of transactions in T that contain X∪Y [2].

$$\textbf{Support}\ (X \rightarrow Y) = \frac{\textbf{support}\ (X \cup Y)}{|T|}$$

Definition4(Confident of rule) The rule X→Y has confidence c in the transaction set D, where c is the percentage of transactions in T containing X that also contain Y[2].

$$\textbf{Confidence}\ (X \rightarrow Y) = \frac{\textbf{support}\ (X \cup Y)}{\textbf{support}(X)}$$

3.5 Discriminant Association Rules

Definition 5 (Strong/Confident of rule) Rules that satisfy both a minimum support threshold and a minimum confident threshold are called strong/confident rules [2]. If support (X→Y) ≥ minimum support and confident (X→Y) ≥ minimum confident, the rule X→Y is called strong or confident association rules where, minimum confident is user defined threshold. Mining confident association rule is performed in two steps [1,8]:

1. Generate all frequent n-itemsets.
2. Using all frequent n-itemsets, generate all strong/confident association rules X→Y, where X and Y are frequent n-itemsets.

3.6 Apriori Algorithm

Apriori algorithm [8] carries out a breadth-first search to enumerate each 1-itemset.

L_1 = {Large 1-itemsets};
for (k=2;L_{k-1}≠∅; k++) do begin
    C_k= Apriori-Gen (L_{k-1});
    forall transactions t∈ D do begin
        C_t= subset (C_k,t);
        forall candidates c∈C_t do
        c.count ++;
    end
    L_k = {c∈C_k | c.count ≥ *minsupport*}
end

Algorithm: Apriori frequent itemset discovery

Input:
D, a transaction data,
Min_sup, the number of support count threshold.

Output:
L, frequent itemsets with intervals

Rule Extraction Process

```
Begin
Step1: Read the gene expression data
Step2: Convert continuous values into discretized value
Step3: Discover frequent rule items with gene intervals
Step4: Generate Association Rules from frequent items
Step5: Discriminant rules
Step6: Visualize the discover patterns.
End
```

Again and again, join(k-1) itemsets with itself to get k-itemsets=2,3,….,L; where L represents the longest frequent itemset. Apriori Algorithm [8] is an important algorithm for mining frequent itemsets [8]. A subset of a frequent itemset must also be a frequent itemset [8]. It uses pruning infrequent itemsets, if there is any itemset is infrequent, its superset should be infrequent [8]. Iteratively find all frequent itemsets with cardinality from 1 to k (k-itemsets). K is the longest frequent itemsets.

4. EXPERIMENTAL RESULTS

In this section it has been presented a novel approach, and shown experimental result of Microarray gene expression analysis using association rules mining. Breast cancer2 relevant gene accession number and gene names are listed out in Table 2. Each gene contains gene names, each gene contain individual accession number. The sample gene expression data related to breast cancer 2 dataset, Column represents the genes and rows represents the samples shown in Table 4.The gene expression values are transformed into discretized values using equal width interval bin method. Where k equal sized bins, where k is a parameter, which is provided by the user. In this experiment bin value 4 is passed by user. After discretization, gene expression values are substituted by gene interval. The experimental results of discretization processes shown in Table 5 and Table 6. After discretization the gene expression data converted into transactional data. Where transactions are represented by TID and gene expression values with gene intervals are represented by itemsets. Then, Association rules are discovering frequent pattern using Apriori algorithm [8]. Finally discover the discriminant rules that satisfy both minimum support and confident. The result of discriminant association rules are shown in Table 8.

Table2. List of the genes accession number with Gene Name

Accession Number	Gene Name
AW613732	LYPD6
PTGER3	PTGER3
AI868854	EST
AJ272267	CHDH
AF208111	IL17BR
AW006861	SCYA4
X59770	IL1R2
AL117406	ABCC11
BC007092	HOXB13
….	…..
AB000520	APS
AI087057	DOK2
M92432	GUCY2D

Table3. Gene Name with Description

Gene Name	Description
LYPD6	Homo sapiens cDNA FLJ31137 fis
PTGER3	prostaglandin E receptor 3 (subtype EP3)
EST	Highly similar to hypothetical protein
CHDH	choline dehydrogenase
IL17BR	interleukin 17B receptor
SCYA4	small inducible cytokine A4
IL1R2	interleukin 1 receptor, type II
ABCC11	ATP-binding cassette, sub-family C (CFTR/MRP), member 11
HOXB13	homeo box B13
…..	
APS	adaptor protein with pleckstrin homology and src homology 2 domains
DOK2	docking protein 2
GUCY2D	guanylate cyclase 2D, membrane

The Gene names with description are related to breast cancer2 data set as shown in Table 3.

Table4. Sample Microarray Data

	ABCC11	**HOXB13**	**CHDH**	**EST_3**	**IL17BR**	**¨**
S1	8.09	-2.78	0.45	0.54	-1.52	…
S2	4.63	-2.34	-0.28	0.39	-2.02	…
S3	5.86	-0.57	1.59	0.81	0.84	…
S4	4.91	1.22	-0.63	-0.1	-1.82	…
…	..	…	…	…	…	…

Table5. Discrete values of gene expression

	ABCC11	**HOXB13**	**CHDH**	**EST_3**	**IL17BR**	**…**
S1	4	1	2	3	1	…
S2	1	1	1	3	1	…
S3	2	3	4	4	4	…
S4	1	4	1	1	1	…
..	…	…	…	…	…	…

Table6. Discretized values substituted by gene intervals

	ABCC11	**HOXB13**	**CHDH**	**EST_3**	**IL17BR**	**…**
S1	[7.225: Inf]	[-inf : -1.78]	[-0.075 :0.480]	[0.3550 : 0.5825]	[-Inf: -1.305]	…
S2	[-Inf : 5.495]	[-inf: -1.78]	[-Inf: -0.075]	[0.3550 : 0.5825]	[-Inf: -1.305]	…
S3	[5.495: 6.360]	[-0.78: 0.22]	[1.035 :Inf]	[0.5825 :Inf]	[0.125 :Inf]	…
S4	[-Inf : 5.495]	[0.22: Inf]	[-Inf: -0.075]	[-Inf: 0.1275]	[-Inf: -1.305]	…
…	…	…	…	…	…	…

Table7. Transaction dataset with gene intervals

T ID	**Itemsets**
S1	ABCC11 [7.225:Inf], HOXB13[-inf: -1.78], CHDH[-0.075 : 0.480], EST_3[0.3550 : 0.5825], IL17BR[-Inf: -1.305]
S2	ABCC11[-Inf : 5.495], HOXB13[-inf: -1.78], CHDH[-Inf: -0.075], EST_3[0.3550 : 0.5825], IL17BR[-Inf: -1.305
S3	ABCC11[5.495 : 6.360], HOXB13[-0.78: 0.22], CHDH[1.035 :Inf], EST_3[0.5825 :Inf], IL17BR=[0.125 :Inf]
S4	ABCC11[-Inf : 5.495], HOXB13[0.22:Inf], CHDH[-Inf: -0.075], EST_3[-Inf: 0.1275], IL17BR[-Inf: -1.305]
…	…..

Table8. Discriminant association rules

#	**Rules**	**Sup.**	**Conf.**
1	{HOXB13[-inf: -1.78]} → {EST_3[0.3550 : 0.5825]}	50%	100%
2	{EST_3[0.3550 : 0.5825]} → {HOXB13 [-inf: -1.78]}	50%	100%
3	{HOXB13[-inf: -1.78]} → {IL17BR [-Inf: -1.305]}	50%	100%
4	{EST_3[0.3550: 0.5825]} → {IL17BR [-Inf : -1.305]}	50%	100%
5	{ABCC11[-inf: 5.495]} → {CHDH [-Inf: -0.075]}	50%	100%
6	{CHDH[-Inf: -0.075]} → {ABCC11 [-inf: 5.495]}	50%	100%
7	{ABCC11 [-inf: 5.495]} → {IL17BR[-Inf: -1.305]}	50%	100%
8	{CHDH [-Inf: -0.075]} → {IL17BR[-Inf: -1.305]}	50%	100%
9	{HOXB13 [-inf -1.78], EST_3[0.3550 : 0.5825]} → {IL17BR[-Inf -1.305]}	50%	100%
10	{HOXB13 [-inf: -1.78], IL17BR[-Inf: -1.305]} → {EST_3 [0.3550 : 0.5825}	50%	100%
11	{EST_3 [0.3550 : 0.5825], IL17BR[-Inf: -1.305]} → {HOXB13 [-inf: -1.78]}	50%	100%
12	{ABCC11 [-Inf 5.495],CHDH [-Inf -0.075]} → {IL17BR [-Inf: -1.305]}	50%	100%
13	{ABCC1 [-Inf 5.495],IL17B[-Inf -1.305]} → {CHDH [-Inf: -0.075]}	50%	100%
14	{CHDH[-Inf -0.075],IL17B[-Inf -1.305]} → {ABCC11 [-Inf: 5.495]}	50%	100%
..	…	…	…

The discriminant rules are extracted from generated frequent patterns with support count 2 (50%) and confidence (100%). Discriminant rules are shown in Table 8. Finally the biological information is extracted from the discriminant rules.

Mining gene expression values with gene intervals are discovering frequent genes with gene interval and extract the relations among the genes. It provides improving prediction for diseases and treatment decisions for cancer patients.

4.1 Biological Knowledge

Biological knowledge is important to get efficient and important patterns to reveal fatal and crucial reasons for diseases from the discriminant rules. For instance in rule 8 in table8, when expressed the CHDH genes also likely to expressed in IL17BR gene. {CHDH [-Inf: -0.075]} → {IL17BR [-Inf: -1.305]}, ILI7BR and CHDH by correlated expressed genes. These genes are related to breast cancer.

The ABCC11 [-Inf 5.495], CHDH [-Inf -0.075]} → {IL17BR [-Inf: -1.305] correlated genes with estrogen receptors status and Human Epidermal growth factor Receptors 2 (HER2) status in Rule 12. ABCC11 gene transcripts were over expressed in estrogen receptor positive breast cancer [12].

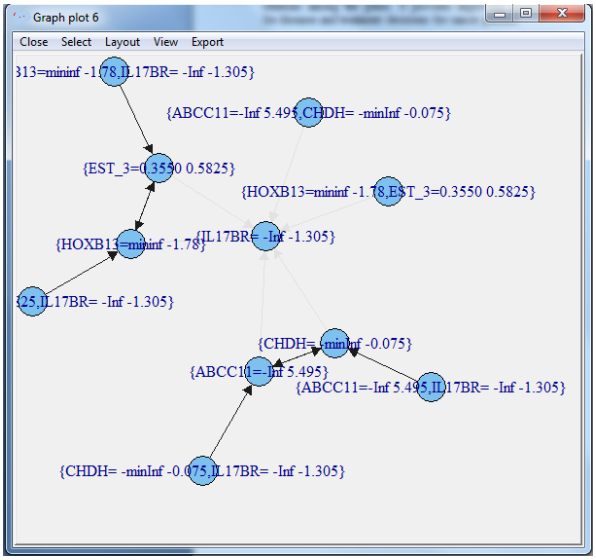

Fig 3: Relation among gene expression using Association rules

The proposed system tested in R statistical language. It has been tested using breast cancer2 dataset [11]. The dataset publically available and were downloaded from www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1379. The original data set consists of 60 samples and 22575 gene expression conditions. The experiments were carried out on a PC with Intel Core 2 Duo CPU and 2 GB of main memory.

4.2 Performance Analysis

Microarray gene expression analysis tested with breast cancer2 dataset [10]. The gene expression values are transformed into discretized values using equal width interval bin method. Time complexity of discretization to discretized one attribute of n objects is O(n). After discretization the gene expression values are generate candidate items using Apriori property [8]. Microarray gene expressions consist of large amount of data or dense data, due to large number of candidate generated so it requires large memory space. The generation of candidate itemsets takes exponential time. The exponential complexity of Apriori is $O(2^n)$.

4.3 Analysis of Algorithm

The proposed approach has been applied to extract the association rules for microarray cancer data analysis using gene expression with intervals. Proposed approach has been compared with previous frequent pattern mining approaches on microarray gene expression data, compared results and description shown in Table 9.

Table9. Comparative analysis of frequent pattern with Algorithm for Microarray Cancer Data Analysis using Frequent Pattern Mining and Gene Intervals

References	Rules	
Becquet *et al.*,[5]	Ribosomal 150→Cytochrome 255	When gene encoding the ribosomal protein is over expressed, then encoding cytochrome also over expressed.
McIntosh et al.,[6]	$\overline{ESC8} \rightarrow \overline{IMD1}, \overline{IMD2}$	When the ESC8 gene under expressed, then IMD1, IMD2 also under expressed.
Zakaria, W., *et al.*,[7]	Gene1→ gene3, $\overrightarrow{gene4}$	There are two kinds of association rules generated. One is up expressed genes and another one is UP/Down expression genes. When gene1 up expressed also expressed gene3 is up expressed and gene4 is down expressed.
Algorithm for Microarray Cancer Data Analysis using Frequent Pattern Mining and Gene Intervals	{HOXB13[-inf: -1.78]} → {EST_3[0.3550 : 0.5825]}	Our Proposed approach has been applied to extract the association rules with gene expression intervals. When gene HOXB13 [-inf:-1.78] is expressed also likely to EST_3[0.3550: 0.5825] genes also expressed. The rules are generated using gene intervals. It has been extracted discriminant rules and correlation among microarray gene expression data profiling and transcriptional regulators (positive regulator and negative regulators).

5. CONCLUSION

In this paper, it has been proposed novel approach for analysis of microarray gene expression values using intervals. It has been extracted correlation among the genes and transcriptional regulators. The market basket analysis has the

property that sparse data set, using this sparse data; the longest frequent itemsets is relatively short. But microarray gene expression datasets, that the number of items (genes) is greater than the number of transactions (samples) called dense dataset. So that, mining frequent patterns need to search efficiently for association analysis on dense data.

Apriori algorithm generates longest frequent itemsets in such dense dataset requires large memory as well as it takes high computational time. It has been generated large number of redundant rules. In future the algorithm will be modified to overcome both computational time and memory explosion problems for microarray gene expression dataset.

6. REFERENCES

[1] Agarwal.R and Srikant.R (1994)," Fast algorithm for mining association rules in large data bases", Proceedings of the 20th international conference on very Large Data Base (VLDB'94), Santiago, chile, pp 487-499.

[2] Han,J., and Kamber,M., "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Elsevier, 2002.

[3] Garcia S, Luengo J, Sez J, Lpez V, and Herrera F , "A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning". IEEE Transactions on Knowledge and Data Engineering vol.25, no.4, pp.734–750. 2013.

[4] Liu, H., Hussain, F., Tan, C. L., and Dash, M.. Discretization: An enabling technique. Data mining and knowledge discovery, vol.6, no.4, pp. 393-423, 2002.

[5] Becquet C, Blachon S, and Jeudy B, et al. "Strong-association-rule mining for large-scale gene-expression data analysis:a case study on human SAGE data". Genome Biology, pp.3:12. 2002.

[6] McIntosh T, and Chawla S.,"High confidence rule miningfor microarray analysis". IEEE/ACM Transactions, Computational Biology and Bioinformatics; vol.4, no.4, pp.611–23, 2007.

[7] Zakaria, W., Kotb, Y., andGhaleb, F., "MCR-Miner: Maximal Confident Association Rules Miner Algorithm for Up/Down-Expressed Genes". Appl. Math, vol.8, no.2, pp.799-809, 2014.

[8] Agrawal,R., Imielinski,T., and Swami,A., "Mining association rules between sets of items in large databases". In: Proceedings of the 1993 ACMSIGMOD International Conference on Management of Data. Washington, DC, USA: ACM Press, pp.207–216, 1993.

[9] Tuimala,J., and Laine,M.M., "DNA Microarray Data Analysis", Second Edition, PicasetOy, Helsinki, 2005.

[10] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1379.

[11] Ghilardi G, Biondi M, La Torre A, Battaglioli L, Scorza R (2005) Breast cancer progression and host polymorphisms in the chemokine system: role of the macrophage chemoattractant protein-1 (mcp-1)-2518 g allele. Clinical Chemistry 51: 452–5.