Mining for Mutually Exclusive Gene Expressions

George Tzanis and Ioannis Vlahavas

Department of Informatics, Aristotle University of Thessaloniki,
Thessaloniki 54124, Greece
{gtzanis, vlahavas}@csd.auth.gr - http://mlkd.csd.auth.gr

Abstract. Association rules mining is a popular task that involves the discovery of co-occurences of items in transaction databases. Several extensions of the traditional association rules mining model have been proposed so far, however, the problem of mining for mutually exclusive items has not been investigated. Such information could be useful in various cases in many application domains like bioinformatics (e.g. when the expression of a gene excludes the expression of another) In this paper, we address the problem of mining pairs and triples of genes, such that the presence of one excludes the presence of the other. First, we provide a concise review of the literature, then we define this problem, we propose a probability-based evaluation metric, and finally a mining algorithm that we apply on gene expression data gaining new biological insights.

1 Introduction

This paper deals with the issue of gene expression analysis. Proteins are the main structural and functional units of an organism's cell, whereas DNA and RNA have the role to carry the genetic information of the organisms. In particular, the genetic information that is coded in the genes of DNA is transcribed into messenger (mRNA) and then is translated into a protein. The functions of an organism depend on the abundance of proteins which is partly determined by the levels of mRNA which in turn are determined by the expression of the corresponding gene. Changes in gene expression underlie many biological phenomena.

The study of gene expression levels may guide to very important findings. SAGE (Serial Analysis of Gene Expression) is a method that provides the quantitative and simultaneous analysis of the whole gene function of a cell [26]. The method works by counting short tags of all the mRNA transcripts of a cell. The set of all tag counts in a single sample is called a SAGE library, and describes the gene expression profile of the sample. An important advantage of the SAGE method, against other methods like microarrays, is that the experimenter does not have to select the mRNA sequences that will be counted in a sample. This is quite important, since the appropriate sequences for studying various diseases such as cancer are not usually known in advance. This advantage of SAGE makes it a fairly promising method, especially for cancer studies as in ours.

In this paper we present a method that utilizes the concept of association rules mining for extracting mutually exclusive expressions of genes. This is a new problem that

has not been studied yet. We define the problem of mining for genes with mutually exclusive expressions. We propose two metrics and a mining algorithm that we study on SAGE data.

The paper is organized as follows. The next section presents the required background knowledge. Section 3 contains a short review of the relative literature. Section 4 contains the description of the proposed approach, definitions of terms and notions that are used, the proposed algorithm and the metrics for measuring the mutual exclusion. In section 5 we present our experiments and discuss important issues and in section 6 we conclude.

2 Preliminaries

This section provides the necessary background knowledge including mining for frequent itemsets and contiguous frequent itemsets.

2.1 Frequent Itemsets

The term "frequent itemset" has been proposed in the framework of association rules mining. The association rules mining paradigm involves searching for co-occurrences of items in transaction databases. Such a co-occurrence may imply a relationship among the items it associates. The task of mining association rules consists of two main steps. The first one includes the discovery of all the frequent itemsets contained in a transaction database. In the second step, the association rules are generated from the discovered frequent itemsets. A formal statement of the concept of frequent itemsets is presented in the following paragraph.

Let $I = \{i_1, i_2, \ldots, i_N\}$ be a finite set of binary attributes which are called *items* and D be a finite multiset of *transactions*, which is called *dataset*. Each transaction $T \in D$ is a set of items such that $T \subseteq I$. A set of items is usually called an *itemset*. The *length* or *size* of an itemset is the number of items it contains. It is said that a transaction $T \in D$ *contains* an itemset $X \subseteq I$, if $X \subseteq T$. The *support* of itemset X is defined as the fraction of the transactions that contain itemset X over the total number of transactions in D:

$$supp_D(X) = \frac{\left|\{T \in D \mid T \supseteq X\}\right|}{|D|} \quad (1)$$

Given a minimum support threshold $\sigma \in (0,1]$, an itemset X is said to be *σ-frequent*, or simply *frequent* in D, if $supp_D(X) \geq \sigma$.

2.2 Contiguous Frequent Itemsets

In the following lines we provide some definitions and formulate the problem of mining contiguous frequent itemsets as defined in our previous work [5].

Every frequent itemset $F \subseteq I$ divides the search space in two disjoint subspaces: the first consists of the transactions that contain F and from now on will be called the *F-subspace* and the second all the other transactions.

Definition 1. Let $F \subseteq I$ be a frequent itemset in D, and $E \subseteq I$ be another itemset. The itemset $F \cup E$ is considered to be a *contiguous frequent itemset*, if $F \cap E = \varnothing$ and E is frequent in the F-subspace.

Itemset E is called the locally frequent extension of F. The term locally is used, because E may not be frequent in the whole set of transactions. In order to avoid any confusion, from now on we will use the terms local and locally, when we refer to a subset of D and the terms global and globally when we refer to D. For example, we use the terms *global support* ($gsup = supp_D$) and *local support* ($lsup = supp_{FSub \subset D}$). There may be set two separate thresholds for global and local support. An itemset F that satisfies the minimum global support threshold (*min_gsup*) is considered to be globally frequent and an itemset E that is frequent in the F-subspace, according to the minimum local support threshold (*min_lsup*), is considered to be locally frequent. The local support of an itemset E in the F-subspace can be calculated as in (2).

$$supp_{F \cap T | T \in D}(E) = \frac{supp_D(E \cup F)}{supp_D(F)} \quad (2)$$

Given a finite multiset of transactions D, the problem of mining contiguous frequent itemsets is to generate all itemsets $F \cup E$ that consist of an itemset F that has global support at least equal to the user-specified minimum global support threshold and an extension E that has local support at least equal to the user-specified minimum local support threshold.

3 Related Work

Some recent efforts have utilized data mining methods for analyzing SAGE data. Decision trees (C4.5) and support vector machines were used in [7] to classify the data according to cell state (normal or cancerous) and tissue type (colon, brain, ovary, etc.). Hierarchical clustering of SAGE libraries was also studied [15]. In [24] hierarchical and partitional (K-Means) clustering algorithms as well as various cluster validation criteria were studied. Other approaches have also been applied on SAGE data, including mining of frequent patterns [25], strong emerging patterns [16], association rules [4], and frequent closed itemsets [9]. The effect of dimensionality reduction methods was studied in [3]. Data cleaning was considered in [14] as well as the process of the attribution of a tag to a gene. Finally, various feature ranking, classification, and error estimation methods were presented in [13].

Association rules were first introduced by Agrawal et al. [1] as a market basket analysis tool. Later, Agrawal and Srikant [2] proposed Apriori, a level-wise algorithm, which works by generating candidate itemsets and testing if they are frequent by scanning the database. Several algorithms have been proposed since then, others improving the efficiency, such as FPGrowth [11] and others addressing different

problems from various application domains, such as spatial [12], temporal [6] and intertransactional rules [21].

One of the major problems in association rules mining is the large number of often uninteresting rules extracted. Srikant and Agrawal [18] presented the problem of mining for generalized association rules. Thomas and Sarawagi [20] propose a technique for mining generalized association rules based on SQL queries. Han and Fu [10] also describe the problem of mining "multiple-level" association rules, based on taxonomies and propose a set of top-down progressive deepening algorithms. Teng [19] proposes a type of augmented association rules, using negative information called dissociations. A dissociation is relationship of the form "*X does not imply Y*", but it could be that "*when X appears together with Z, this implies Y*".

Another kind of association rules are negative association rules. Savasere et al. [17] introduced the problem of mining for negative associations. They propose a naive and an improved algorithm for mining negative association rules along with a new measure of interestingness. In a more recent work Wu et al. [27] present an efficient method for mining positive and negative associations and propose a pruning strategy and an interestingness measure. Their method extends the traditional positive association rules ($A \Rightarrow B$) to include negative association rules of the form $A \Rightarrow \neg B$, $\neg A \Rightarrow B$, and $\neg A \Rightarrow \neg B$. The last three rules indicate negative associations between itemsets A and B. A mutual exclusion can not be expressed by one such rule. If items a and b are mutually exclusive, then $\{a\} \Rightarrow \neg\{b\}$ and $\{b\} \Rightarrow \neg\{a\}$ concurrently, that is different from $\neg\{a\} \Rightarrow \neg\{b\}$. The problem of mining pairs of mutually exclusive items has been recently introduced [22, 23].

4 Our Approach

In this section we provide a detailed description of the proposed approach. Before presenting the basic steps of this approach we will describe the structure of the input data.

The data are structured in a gene expression matrix A. The columns of the matrix represent the tags of the genes and the rows represent the different samples (SAGE libraries). The intersection of the i^{th} row with the j^{th} column, namely the element a_{ij}, is the gene expression level for the gene j in the sample i.

4.1 Discretization

The data that will be used for mining the mutually exclusive gene expressions should contain binary values. Each value denotes if a gene in a particular SAGE library is expressed or not. We have utilized three methods for discretization that have been proposed in [4]. These methods are presented below:

- *Max minus x%.* This consists of identifying the highest expression value (HV) in any library for each tag, and defining a value of 1 for the expression of that tag in a library when the expression value is above (HV - x)/100. Otherwise, the expression of the tag is assigned a value of 0.

- *Mid-range based cutoff*. The highest and lowest expression values are identified for each tag and the mid-range value is defined as the arithmetic mean of these two numbers. Then, all expression values below or equal to the mid-range are set to 0, and all values above the mid-range are set to 1.
- *X% of highest value*. For each tag, we identified libraries in which its level of expression was in the x% of highest values (e.g. 30%). These are assigned the value 1, and the rest are set to 0.

The use of the above methods filters out a large number of infrequent tags that may correspond either to genes that have very low expression levels or to mistakenly counted tags (e.g. a tag with count 1 could be caused in an error in sequencing of the tag). Any of the above methods can be selected depending on the particular study and dataset. For example if a very short number of expressed genes is desirable, then method *10% of highest value* would be a good choice.

At this step, also, the dataset is converted to a transactional format, so that each sample contains the IDs of the genes (tags) that are expressed in the particular sample (SAGE library).

4.2 Definition of Mutual Exclusion

Definition 2. Let D be a finite multiset of transactions (samples or SAGE libraries) and I be a finite set of items (genes or tags). Each transaction $T \in D$ is a set of items such that $T \subseteq I$. If two items $i_1 \in I$ and $i_2 \in I$ are *mutually exclusive*, then there is not any transaction $T \in D$, such that $\{i_1, i_2\} \subseteq T$. If a gene i is contained in transaction T, then gene i is expressed in SAGE library T.

The above definition of mutual exclusion is strict. However, the inverse of definition 1 does not generally stand, so it cannot be used to identify mutually exclusive items. Typically in a SAGE dataset genes are in tens of thousands, whereas SAGE libraries (transactions) are in hundreds. It is possible that there is a large number of pairs of genes that are never expressed together. According to the inverse of definition 2 all of these pairs of genes have mutually exclusive expressions. But, in fact only a very small number of these pairs of genes have truly mutually exclusive expressions.

Mining for mutually exclusive items in a database possibly containing several thousands of different items, involves searching in a space that consists of all the possible pairs of items, because virtually any of them could contain two items that exclude each other. However this approach is naive and simplistic and can lead to many mutually exclusive items that in fact are not. We propose a more intuitive approach, which is based on the assumption/observation that every frequent itemset expresses a certain behavior and therefore it could be used to guide our search. Items that appear with high frequency in the subspace of a frequent itemset are more likely to be systematically mutually exclusive, because they follow a pattern and not because of pure chance or unusual cases.

Our approach consists of three steps. In the first step, all the frequent itemsets are mined. Then, the frequent itemsets are used for mining the contiguous frequent itemsets, producing the extensions that will be used in the next step as candidate mutually exclusive items. Any frequent itemset mining algorithm can be used in the first step. Step 2 works in a level-wise manner and requires a number of scans over the data-

base, which is proportional to the size of the extensions discovered. The extensions of the contiguous frequent itemsets mined at the second step are candidates for participating in a pair of mutually exclusive items.

4.3 Mutual Exclusion Metrics and Mining Algorithm

In order to distinguish when two items are mutually exclusive, we need a measure to evaluate the degree of the mutual exclusion between them. Initially, we should be able to evaluate this within the subspace of a frequent itemset (locally) and then it should be evaluated globally, with all the frequent itemsets that support this candidate pair to contribute accordingly. For this purpose, we propose the use of a metric we call MEM (Mutual Exclusion Metric) that can be calculated in two phases, the first one is local and is required for the second one, which is the global one.

Local Metric. We propose the following local metric (3), which will be called Local MEM for the evaluation of a candidate pair of mutually exclusive items that is supported by a frequent itemset I and its range is [0, 1].

$$LM_I(A,B) = \left[P(A-B)+P(B-A)\right]\min\left[P(A-B\mid A), P(B-A\mid B)\right] = \tag{3}$$
$$\left[(S_A - S_{AB})+(S_B - S_{AB})\right]\min\left[\frac{(S_A - S_{AB})}{S_A}, \frac{(S_B - S_{AB})}{S_B}\right] =$$
$$(S_A + S_B - 2S_{AB})\left[1 - \frac{S_{AB}}{\min(S_A, S_B)}\right]$$

For the above formula $P(I) = 1$. S_X is the fraction of transactions that contain X over the number of transactions that contain I.

Global Metric. We propose the following global metric (4) for the evaluation of a candidate pair of mutually exclusive items that is supported by a set IS of frequent itemsets.

$$GM_I(A,B) = IIF\left(\sum_{I\in \mathrm{IS}} S_I LM_I(A,B)\right) \tag{4}$$

IIF stands for *Itemset Independence Factor* and is calculated as the ratio of the number of the distinct items contained in all itemsets that support a candidate pair over the total number of items contained in these itemsets. For example, the *IIF* of the itemsets {*A, B, C*} and {*A, D*} is 0.8, since there are 4 distinct items (*A, B, C* and *D*) over a total of 5 ones (*A, B, C, A* and *D*). The *IIF* is used in order to take into account the possible overlapping of two candidate mutually exclusive items. We do this, because the overlapping between the transactions that contain two different itemsets implies overlapping between the transactions that contain the pair.

Using the above metrics we have implemented an algorithm for mining pairs of mutually exclusive items [22]. In our study we have extended our algorithm for mining not only pairs, but also triples of mutually exclusive items and we have adapted our approach for application on gene expression domain. We have implemented a tool that is available at our group's website: http://mlkd.csd.auth.gr/mutex/index.html.

5 Experiments

In this section we describe the dataset and the results of our experiments and we discuss some important issues.

5.1 Dataset

We have used a SAGE dataset that consists of 90 SAGE libraries and 27679 tags. This dataset has been provided by Dr Olivier Gandrillon's team (Centre de Génétique Moléculaire et Cellulaire de Lyon, France) and has been studied and presented at the ECML/PKDD Discovery Challenge Workshops in 2004 and 2005. The SAGE libraries contained in this dataset have been prepared as of December 2002 [8]. They are collected from various human tissue types (colon, brain, ovary, etc.) and are labeled according to their cell state that is either normal or cancerous.

5.2 Results

We have conducted a number of experiments in order to evaluate the behavior of our approach. Table 1 presents the mean transaction size of the datasets that are generated using many variants of the three discretization methods described in 4.1. As it is shown the counts of the tags are not equally distributed. For example, method "max minus 5%" provides a mean transaction size of 344 genes while method "5% of highest value" provides a mean transaction size of 1047 genes. This means that the tags that have at most 5% smaller count than the maximum count are not the 5% of the highest counted tags but a very smaller percentage.

Table 1. Mean transaction size for various discretization methods

Discretization Method	Mean Transaction Size
Max minus 5%	344
Max minus 10%	392
Max minus 15%	440
Max minus 20%	500
Max minus 25%	580
Max minus 30%	658
5% of highest value	1047
10% of highest value	1867
15% of highest value	2636
20% of highest value	3113
25% of highest value	3591
30% of highest value	3897
Mid-range based cutoff	1273

Fig. 1 presents the number of mutually exclusive pairs and triples of genes that are mined for various values of minimum local support threshold, when minimum global

support threshold is 0.25 and mid-range cutoff discretization is used. As shown in the graph the number of mutually exclusive genes grows exponentially with minimum local support threshold. Moreover, the number of triples is in most cases greater than the number of pairs. The same happens when minimum global support threshold varies (these results are not presented here, due to space limitations).

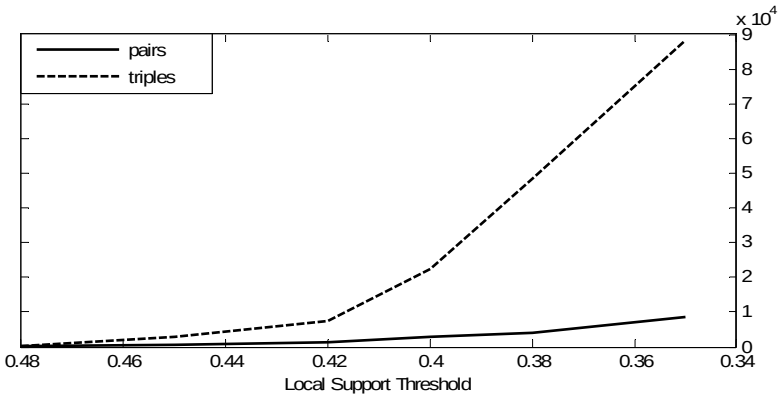

Fig. 1. Number of mutually exclusive genes against local support threshold (min_gsup = 0.25, mid-range cutoff discretization)

5.3 Discussion

In our approach a level-wise technique is applied in order to extract the contiguous frequent sets of genes and then pairs and triples of genes that their expressions are mutually exclusive. Searching for mutually exclusive pairs of items in a "blind" manner would produce a huge amount of candidate pairs. Moreover, most of the discovered mutually exclusive items would be uninteresting. The intuition behind searching for mutually exclusive items between the extensions of frequent itemsets is manifold. First, the search space is reduced sensibly. Second, genes that are expressed in particular cases and thus are expressed in a small number of samples could not be mined as globally frequent. This does not mean that these genes are not important. However, they cannot be mined guiding to possible loss of knowledge. In our approach, these genes may be found as frequent extensions of other frequent sets of genes, recovering the aforementioned loss of knowledge. Third, if a large number of frequent sets of genes share the same extensions and these common extensions are frequent in the subspace of these sets, they are likely to be mutually exclusive and possibly of the same category and the same level of a taxonomy (e.g. same tissue type).

As shown by the experiments all pairs or triples of mutually exclusive genes contain genes with different functions. In some cases there are found genes or ESTs with unknown functions. This denotes an important use of the proposed approach. This approach could be used as part of a procedure for discovering the function of a gene or EST. For example, if a gene with an unknown function is found to be mutually exclusive with other genes with known functions, then it is very possible that the

function of the gene is none of the functions of its mutually exclusive genes. Moreover, genes that suppress each other under some conditions can also be found by the proposed approach.

6 Conclusions and Future Work

In this paper we have presented the novel problem of mining for mutually exclusive gene expressions. When two or more genes have mutually exclusive expressions, this can be used as a valuable hint when looking for previously unknown functional relationships among them. In such case, this can be an interesting type of knowledge to the domain expert. For this purpose, we propose an intuitive approach, formulated the problem providing definitions of terms and evaluation metrics. We have also developed a mining algorithm. In the future, we would like to extend our algorithm for mining not only pairs and triples of mutually exclusive genes but sets of unlimited size. Moreover, we will deal with the improvement of the efficiency of our algorithm.

References

1. Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. In Proceedings of the ACM SIGMOD Conference on Management of Data, 207-216.
2. Agrawal, R., and Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the International Conference on Very Large Databases. 487-499.
3. Alves, A. Zagoruiko, N. Okun, O. Kutnenko, O., and Borisova, I.. "Predictive Analysis of Gene Expression Data from Human SAGE Libraries". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 60-71.
4. Becquet, C. Blachon, S. Jeudy, B. Boulicaut, J.F. and Gandrillon, O.. "Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data". *Genome Biol*ogy, 3(12), 2002.
5. Berberidis, C., Tzanis, G., and Vlahavas, I. (2005). Mining for Contiguous Frequent Itemsets in Transaction Databases, In Proceedings of the IEEE 3rd International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications.
6. Chen, X., and Petrounias, I. (2000). Discovering Temporal Association Rules: Algorithms, Language and System. In Proceedings of the 16th International Conference on Data Engineering.
7. Gamberoni G. and Storari S.. "Supervised and unsupervised learning techniques for profiling SAGE results". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 2004, pp. 121–126.
8. Gandrillon O.. "Guide to the gene expression data". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 2004, pp. 116–120
9. Gasmi, G. Hamrouni, T. Abdelhak, S. Ben Yahia, S. and Mephu Nguifo, E.. "Extracting Generic Basis of Association Rules from SAGE Data". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, 84–89.
10. Han, J., and Fu, Y. (1995). Discovery of Multiple-Level Association Rules from Large Databases. In Proceedings of the 21st International Conference on Very Large Databases. 420-431.

11. Han, J., Pei, J., and Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. 1-12.
12. Koperski, K. and Han, J. (1995). Discovery of Spatial Association Rules in Geographic Information Databases. In Proceedings of the 4th International Symposium on Large Spatial Databases. 47-66.
13. Lin H.-T. and Li. L. "Analysis of SAGE Results with Combined Learning Techniques". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 102–113.
14. Martinez, R. Christen, R. Pasquier, C. and Pasquier, N. "Exploratory Analysis of Cancer SAGE Data". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 72–77.
15. Ng, R.T. Sander, J. and Sleumer, M.C.. "Hierarchical cluster analysis of SAGE data for cancer profiling". In *Proceedings of Workshop on Data Mining in Bioinformatics*, 2001, pp. 65-72.
16. Rioult F.. "Mining strong emerging patterns in wide SAGE data". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 2004, pp. 484–487.
17. Savasere, A., Omiecinski, E, and Navathe, S.B. (1998). Mining for Strong Negative Associations in a Large Database of Customer Transactions. In Proceedings of the 14th International Conference on Data Engineering. 494-502.
18. Srikant, R., and Agrawal, R. (1995). Mining Generalized Association Rules. In Proceedings of the 21st VLDB Conference. 407-419
19. Teng, C.M. (2002). Learning form Dissociations. In Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery. 11-20.
20. Thomas, S., and Sarawagi, S. (1998). Mining Generalized Association Rules and Sequential Patterns Using SQL Queries. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. 344-348.
21. Tung, A. K. H., Lu, H., Han, J., and Feng, L. (2003). Efficient Mining of Intertransaction Association Rules. IEEE Transactions On Knowledge And Data Engineering. 15(1), 43-56.
22. Tzanis, G. and Berberidis, C. Mining for Mutually Exclusive Items in Transaction Databases. International Journal of Data Warehousing and Mining, 3(3), IGI Global, 2007.
23. Tzanis, G. Berberidis, C., and Vlahavas, I. On the Discovery of Mutually Exclusive Items in a Market Basket Database, In Proceedings of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery, Thessaloniki, Greece, September 6, 2006.
24. Tzanis, G. and Vlahavas, I.. "Mining High Quality Clusters of SAGE Data". In *Proceedings of the 2nd VLDB Workshop on Data Mining in Bioinformatics*, Vienna, Austria, 2007.
25. Tzanis, G. and Vlahavas, I.. Accurate Classification of SAGE Data Based on Frequent Patterns of Gene Expression. In Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), IEEE, Patras, Greece, October 29-31, 2007.
26. Velculescu, V.E. Zhang, L. Vogelstein, B., and Kinzler, K.W.. "Serial analysis of gene expression", *Science*, 270 (5235), 1995, pp. 484-487.
27. Wu, X., Zhang, C., and Zhang, S. (2004). Efficient Mining of both Positive and Negative Association Rules. ACM Transactions on Information Systems. 22(3), 381-405.