Association Rule Mining in Genomics

Article *in* International Journal of Computer Theory and Engineering · January 2010

DOI: 10.7763/IJCTE.2010.V2.151

CITATIONS
31

READS
1,182

3 authors, including:

Mrinal Kanti Ghose
(Formerly, Sr. Scientist, Vikram Sarabhai Space Centre & RRSC (East), ISRO)
278 PUBLICATIONS **1,979** CITATIONS

SEE PROFILE

International Journal of Computer Theory and Engineering, Vol. 2, No. 2 April, 2010
1793-8201

Association Rule Mining in Genomics

M.Anandhavalli *Member, IACSIT, IAENG*, M.K.Ghose, K.Gauthaman

***Abstract*— Association rules, used widely in the area of market basket analysis, can be applied to the analysis of expression data as well. Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression. An association rule has the form LHS→RHS, where LHS and RHS are disjoint sets of items, the RHS set being likely to occur whenever the LHS set occurs. Items in gene expression data can include genes that are highly expressed or repressed, as well as relevant facts describing the cellular environment of the genes (e.g. the diagnosis of a tumor sample from which a profile was obtained). In this paper, association rule mining techniques that have been recently developed and used for genomic data analysis have been reviewed and discussed.**

***Index Terms*—Association Rule Mining (ARM), Gene Expression data.**

I. Introduction

There has been a great explosion of genomic data in recent years. This is due to the advances in various high-throughput biotechnologies such as gene expression microarrays. These large genomic data sets are information-rich and often contain much more information than the researchers who generated the data might have anticipated. Such an enormous data volume enables new types of analyses, but also makes it difficult to answer research questions using traditional methods. Global Gene expression data can be a valuable tool in understanding of genes, biological networks, and cellular states. Analysis of these massive genomic data has two important goals:

First goal is try to determine how the expression of any particular gene might affect the expression of other genes; the genes involved in this case could belong to the same gene network. By a gene network, we mean a set of genes being expressed together in a non-random pattern.

Second goal of expression data analysis is try to determine what genes are expressed as a result of certain cellular conditions, e.g. what genes are expressed in diseased cells that are not expressed in healthy cells.

In this paper, an attempt has been made to review the novel concepts and techniques proposed for mining association rule from the genomic data have been reviewed.

Manuscript received September 12, 2009. This work has been carried out as part of Research Promotion Scheme (RPS) Project under AICTE, Govt. of India.

M.Anandhavalli, Dr. M.K.Ghose, Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology,Majitar, East Sikkim, India - 737136. e-mail: anandhigautham@gmail.com.

Dr. K.Gauthaman Department of Pharmacognosy, Himalayan Pharmacy Institute, Majitar, East Sikkim-737136, India.

II. Notations

→ means "implies" ↑: Highly expressed
↓: Highly repressed U:Union

III. Gene Expression Data

The gene expression data in microarray are presented in M×N matrix where M is the number of microarray experiments and N being the number of genes [1]. The number of experiments M can range from dozens to thousands. On the other hand, the number of genes N can range from hundred to tens of thousands. In some context, M can be referred to as number of transactions or item sets where each gene represents an item. To add to the complexity of representation, each gene is measured in terms of absolute values. However, biologists are more interested in how gene expression changes under different environments in each respective experiment. Thus, these absolute values are discretized according to some predetermined thresholds and grouped under three different levels, namely unchanged, up regulated and down regulated.

IV. Association Rules

Association rules are used widely in the retail industry under the name 'market basket analysis'. Association rules have been used as well to mine medical record data [2],[3]. The general definition for association rules is as follows:

- An 'association rule' is a pair of disjoint itemsets. If LHS and RHS denote the two disjoint itemsets, the association rule is written as LHS→RHS i.e LHS and RHS are sets of items, the RHS set being likely to occur whenever the LHS set occurs.
- The 'support' of the association rule LHS→RHS with respect to a transaction set T is the ratio of support(LHS U RHS)/ T.
- The 'confidence' of the rule LHS→RHS with respect to a transaction set T is the ratio of support(LHS U RHS)/ support(LHS).

In market basket analysis, an association rule represents a set of items that are likely to be purchased together; for example, the rule {cereal}→{milk, juice} would state that whenever a customer purchases cereal, he or she is likely to purchase both milk and juice as well in the same transaction. In the analysis of gene expression data, the items in an association rule can represent genes that are strongly expressed or repressed, as well as relevant facts describing the cellular environment of the genes (e.g. a diagnosis for a tumor sample that was profiled, or a drug treatment given to cells in the sample before profiling). An example of an

association rule mined from expression data might be {cancer}→{gene A↑, gene B↓, gene C↑}, meaning that, for the data set that was mined, in most profile experiments where the cells used were cancerous, gene A was measured as being up (i.e. highly expressed), gene B was down (i.e. highly repressed), and gene C was up, altogether.

V. Association VS Clustering

Most of the available gene-expression data-analysis methods are based on clustering algorithms that try to establish synexpression groups [4], that is, groups of genes whose expression is correlated in different biological situations. The basis for all clustering algorithms is their ability to generate groups of genes that fulfill two related constraints: maximum intragroup similarities and minimum intergroup similarities. Although such algorithms have been quite successful, most notably in the molecular profiling of human cancers [5], their biological validity can be questioned when the identification of molecular networks is the goal. In this context, they have three main drawbacks. First, a gene which functions in numerous physiological pathways will have to be clustered in one and only one group. Second, no relationship can be inferred between the different members of a group. That is, a gene and its target genes will be co-clustered, but the type of relationship cannot be rendered explicit by the algorithm. Third, most clustering algorithms will make comparisons between the gene-expression patterns in all the conditions examined. They will therefore miss a gene grouping that only arises in a subset of cells or conditions.

To overcome these problems, the potential impact of the association-rule discovery (ARD) technique is investigated. This is an unsupervised data-mining technique that seeks descriptive rules in potentially very large datasets [4]. This method should resolve the above drawbacks of existing clustering approaches for the following reasons. First, any gene can be assigned to any number of rules as long as its expression fulfills the assignment criteria. This means that a gene involved in many synexpression groups will appear in each and every one of those groups, without limitation. Second, rules are orientated (If … then …) and thus to a certain extent describe the direction of a relationship. For example, in the overall dataset, a specific subset of cells exhibit highly characteristic patterns of gene expression, the algorithm should be able to detect it. Last but not least, by focusing on strong rules, the biologist does not have to browse and study a huge number of redundant rules.

VI. Role of Arm in Genomics

A. Distance-based Association Rules Mining (DARM)

Gene expression is the effective production of the protein that a gene encodes. Control of gene expression remains one of the fundamental unsolved problems of biology. The basic problem is deceptively simple. The primary sequences that control most gene expression (defined here as transcription of DNA into RNA) are known to be located in the non-coding DNA upstream from the coding region. If several genes are expressed in the same temporal and spatial pattern in an organism, then it seems there must be DNA sequences in common among the non-coding regions of these genes that control the timing and location of expression. Although the complete genome sequence for many organisms is now available, most sequences known to be involved in control of transcription have been identified by painstaking molecular and genetic analyses rather than through computational analysis comparing DNA sequences. There are many reasons for the difficulty in translating knowledge of DNA sequence into understanding of transcriptional control. Molecular analysis has shown that the DNA sequences or motifs that control transcription act by allowing the binding of protein transcription factors to non-coding DNA. See Figure 1. For a review, see [6].

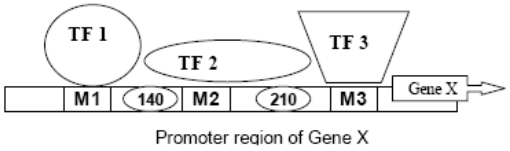

Fig. 1. Gene expression: Transcription factors TF1, TF2, and TF3 bind to motifs M1, M2, and M3, respectively, and allow transcription of Gene X to occur. Numbers in ovals represent distances between motifs in base pairs.

DARM [7] has been proposed to focus on the characterization of the expression patterns of genes based on their promoter regions. The promoter region of a gene contains short sequences called motifs to which gene regulatory proteins may bind, thereby controlling when and in which cell types the gene is expressed. DARM addresses two important aspects of gene expression analysis: (1) Binding of proteins at more than one motif is usually required, and several different types of proteins may need to bind several different types of motifs in order to confer transcriptional specificity. (2) Gives the order in which the proteins controlling transcription may need to be interact physically. Here the association rules are used to involve multiple motifs and to predict expression in multiple cell types. To address the second aspect, the association rules are enhanced with information about the distances among the motifs, or items, which are present in the rule. Rules of interest are those whose set of motifs deviates properly, i.e. set of motifs whose pair-wise distances are highly conserved in the promoter regions where these motifs occur.

B. Duplicate Association Rules

The Association rules have been used for determining the biological data duplicates [10]. The protein or DNA sequences submitted by biologists from numerous sequencing centres and laboratories around the world to the public sequence databases are subjected to various sources of redundancy:

- The same sequence may be submitted by the biologist to more than one database without cross-referencing these records.
- The sequence is submitted more than once to a same database.
- Annotations of the same sequence are submitted separately by different research groups.

- Fragments and partial entries of the same protein or DNA sequence may be stored in different database records.

Biological data duplicates are varying representations of the same protein or DNA sequences in different database records. They provide hints of the redundancy in biological datasets. For example, record 1 and 2 in Table 1 refer to the same protein found separately in a PIR [8] and a Swiss-Prot database [9] records. The example is likely resulted from the submission of the same protein sequence to both PIR and Swiss-Prot without cross referencing to each other records.

TABLE I. DUPLICATE PROTEIN RECORDS

Fields	Record 1	Record 2
Locus ID	P34180	S22388
Definition	Phospholipase A2, neutral precursor (Ammodytin I2) (Phosphatidylcholine 2-acylhydrolase).	phospholipase A2 (EC 3.1.1.4) ammodytin I2 Precursor - western sand viper.
Database	swissprot: locus	pir: locus S22388;
source	PA2N_VIPAA, accession P34180;	
Organism	Vipera ammodytes Ammodytes	Vipera ammodytes ammodytes
Sequence	MRTLWIVAVCLIGV EGNLYQFGNMIFK MTKKSALLSYSNYG CYCGWGGKGKPQD ATDRCCFVHDCCY GRVNGCDPKLSIYS YSFENGDIVCGGDD PCLRAVCEC DRVAAICFGENLNT YDKKYKNYPSSHCT ETEQC	MRTLWIVAVCLIGV EGNLYQFGNMIFKM TKKSALLSYSNYGC YCGWGGKGKPQDA TDRCCFVHDCCYGR VNGCDPKLSIYSYSF ENGDIVCGGDDPCL RAVCEC DRVAAICFGENLNT YDKKYKNYPSSHCT ETEQC

C. Heterogeneous Association Rules Mining

Bioinformatics databases are highly heterogeneous, not only do they differ in their representation but they also offer radically different query capabilities across the diverse information held in distributed resources. The need to extract and link knowledge from these very large databases is increasing. Gene mutation is one of the promising application areas. Extracting interesting patterns and rules from gene mutation datasets can be important in identifying cause of gene tumours and diseases. The discovery of interesting association relationships among huge amount of gene mutation can help in determining the cause of mutation in tumours and diseases.

The Human Gene Mutation Database (HGMD) represents an attempt to collate known (published) gene lesions responsible for human inherited disease. All HGMD entries comprise a reference to the first literature report of a mutation, the associated disease state as specified in this report, the gene name, symbol and chromosomal location [11].

The Mammalian Gene Mutation Database (MGMD) stores the mutation spectra information such as Mutagen, Species, Tissue, Cell line, Gene, Mutation Class, Mutation or first name author from the reference, or the Medline abstract number of the studies. It is a single relational table which has 39134 records [12].

From above databases, sets of items whose elements tend to be in both databases have been retrieved to discover the interesting association rules among genes, mutations, mutagens and diseases.

D. Transcription Factors Analysis using Association Rules

Gene expression data are stored in the database grouped by the tissues or organs where they are present. A gene (an item) in the database is identified by its access number, name and the sequence of the gene. One way to understand how different genes regulate each other is to measure gene expression levels [13] produced by a cell using microarray technologies [14]. Typically, biologists conduct a number of experiments measuring gene expression levels of a cell or a group of cells under various conditions affecting these expression levels. Biologists are interested in how different gene expressions change depending on the type of a tissue, age of the organism, therapeutic agents, and environmental conditions. Most genes are expressed consistently in every tissue and organ, which are defined as housekeeping genes. Some genes are induced to express by other gene expressions or factors. Transcription factors consist of a large number of proteins that were classified into different families.

The rules of transcription factors (x_1, x_2,…,x_n) and their target gene (y) are defined as follows:

1) If xs (one or more) exists→y exists in the dataset;
2) If one of the xs does not exists →y does not exist in the dataset.
3) An xs exists →different y exists, i.e., a transcription factor may be participated in different target gene expressions.

To apply the association rules for mining the transcription actors of the target gene, each type of tissues is defined as a set of transactions or a dataset (e.g. HL60, one of the blood tissues). In a dataset, each tissue sample that consists of many genes (transcription factors and target genes) is viewed as one transaction, given in Table 2.

TABLE II. DUPLICATE PROTEIN RECORDS

Tissue 1 (trans 1)	Tissue 2 (trans 2)	...	Tissue n (trans n)
Tf 11	Tf 21	...	Tf m1
Tf 12	Tf 22	...	Tf m2
.	.	...	.
.	.	...	.

Association rules have been effectively applied to obtain transcription factors associated with gene expressions.

E. Interesting Rule Group Analysis using Association Rules

Gene expression data has a large number of columns which poses a great challenge for existing rule mining algorithms, since their basic approaches are the column-wise enumerations where combinations of columns are tested incrementally to search for frequent occurrences of certain combinations. Column wise association rule mining algorithms generally have the following three problems on gene expression data:

Problem 1: Extremely long running time due to the huge column enumeration space.

Problem 2: Too many association rules found due to the combinatorial explosion of frequent itemsets.

To address these 2 problems, a novel row-wise depth-first

algorithm FARMER [15] has been proposed to mine all the interesting rule groups (IRGs) [16] satisfying user-specified minimum measure (support, confidence, chi square value) thresholds, instead of finding individual association rules.

F. Fuzzy Association Rules

Biological data are often heterogeneous, imprecise and noisy. Integration and analysis of this information are required to understand gene roles in cell behaviour. Fuzzy set theory is generally suitable to model imprecise and noisy data and association rules are very appropriate to deal with imprecise and uncertain concepts. Classical quantitative association rule mining methods partition continuous domains into crisp intervals [17]. Fuzzy logic is proved to be a superior technology to enhance the interpretability of these intervals. The fuzzification of the continuous domains is carried out by partitioning them into fuzzy sets. Fuzzy confidence and support measure the significance of the rule. Thus, fuzzy association rules are expressions of the form X→Y, but in this case, X and Y are sets of fuzzy attribute-value pairs. Fuzzy association rules [18] have been proposed to consider simultaneously gene expression data, Go annotations and gene structures. Linguistic labels (e.g. HIGH, MEDIUM, and LOW) are often used when gene structural features are considered.

G. Co-regulated Gene Mining using Association Rules

Genes with similar patterns of mRNA expression and similar functions are likely to be regulated via the same mechanisms. Co-regulated genes are those regulated by atleast one common transcription factor [20], [21]. Research shows that in order for two genes to have a greater that 50% chance of sharing a common transcription factor binder, the correlation between their expression profiles must be greater than 0.84 [20]. Liping Ji and Kian-Lee [22] analysed the dynamic range of the gene expression value to find the co-regulated genes, which precisely reflects their relations. Since gene expression values are static data, but the regulation is a dynamic process, it is more reasonable to mine the co-regulated genes by their changing tendency. Association rules combined with hash tree and genetic algorithm has been proposed to mine regulated genes from yeast genome dataset by generating lots of co-regulated genes [19]. The method for mining the co-regulated genes from gene expression matrix O includes three steps and shown in Fig. 2 :

1) Transforming the gene expression value O'.
2) Mining the frequent itemsets from matrix O".
3) Generating rules from frequent itemsets with genetic algorithm.

The entry in gene expression matrix $O_{i,j}$ is transformed to $O_{i,kj}$ according to equation (1), and generates matrix O', which reflects the relative changing tendency of genes. Then O" can be obtained by binning the values of O' in equation (2) to prevent noise introduced by experimental errors and make it clear the general increasing or decreasing tendency of gene values where t is a threshold for binning.

$$O'_{i,kj} = \begin{cases} \dfrac{O_{i,j} - O_{i,k}}{|O_{i,k}|} \text{ if } O_{i,k} \neq 0 \\ 1 \text{ if } O_{i,k} = 0 \text{ and } O_{i,j} > 0 \\ -1 \text{ if } O_{i,k} = 0 \text{ and } O_{i,j} < 0 \\ 0 \text{ if } O_{i,k} = 0 \text{ and } O_{i,j} = 0 \end{cases} \quad (1)$$

$$O''_{i,kj} = \begin{cases} 1 \text{ if } O'_{i,kj} \geq t \\ -1 \text{ if } O'_{i,kj} < 0 \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

Under the regulation of the same transcription factors, co-regulated genes have the same changing tendency, which is increasing or decreasing at the same time.

H. Ant-based Association rule mining (Ant-ARM)

An ant-based association rule mining (Ant-ARM) algorithm which makes use of natural behaviour of ants such as cooperation and adaptation to allow for a flexible robust search for a good candidate solution, have been proposed for gene expression data analysis.

Ant-ARM is applied to discover classification rules for one particular class only. That is, the training set will contain cases from one particular class. One ant is used to construct one Classification Association Rule (CAR) and it adds one itemsets at a time. The Ant-ARM approach consists of four main processes: Initialization, Tour Construction, Pheromone Update, and Rule Set Update. For the procedure and design issues of Ant-ARM illustration, see [23].

Ant-ARM has been tested on the acute lymphoblastic leukaemia (ALL)/acute myeloid leukaemia (AML) dataset and generated about 30 classification rules with high accuracy [23].

VII. Conclusion

Basically, data mining is an application-dependent issue and different applications may require different mining techniques to copy up with. To apply mining association rules in gene expression analysis, we need to understand the properties of gene expression data efficiently. We believe that there are several following reasons why these association rule mining approaches for genomic data have been successful and represent a promising direction for future work:

- It has been found that evaluation of duplicate association rule mining on a real-world dataset shows that duplicate association rules can accurately identify up to 96.8% of the duplicates in the dataset at the accuracy of 0.3% false positives and 0.0038% false negatives.
- Extracting interesting patterns and rules from gene mutation datasets can be important in identifying cause of gene tumours and diseases. Heterogeneous association rules has been used to access and link various databases to extract knowledge and patterns of interest, while not needing to be aware of the representation details in the individual resources.
- Distance-based Association Rules Mining used to involve multiple motifs and to predict expression in multiple cell types because set of motifs whose pair-wise distances are highly conserved in the promoter regions where these motifs occur.

- ARM has been used to discover transcription factors that are essential and required for the gene expression of a target gene in the database in very short time. It can be a useful tool to identify genes directly activated or repressed by expression of a transcription factor.
- IRG has been used to mine all the Interesting Rule Groups of association in gene expression data, instead of finding individual rules.
- Fuzzy association rules have been used to obtain interesting relations between functional and structural gene features.
- ARM for Co-regulated gene expression stores all the frequent itemsets into a unit hash tree for improving the time and the space consumption of algorithm. However, after using genetic algorithm, the searching time of association rules is remarkably shortened.
- Ant-ARM method based on the ant colony optimization, a nature inspired algorithm emerging from the collective behaviour of social ant colonies, used for a flexible robust search for a good candidate solution.

Furthermore, unlike data mining in business applications, the size of a transaction in gene analysis is relatively small since the number of tissues present in an organism (e.g. human tissues) is limited. However, the number of items (genes) in one single transaction is very large. When we select an algorithm to facilitate this analysis, the number of passes is not a major factor to be considered

Finally, most importantly, based on the significant efforts by the NIH (Gene Expression Omnibus, GEO) and the EBI (ArrayExpress), many precious microarray data sets of cancer—both cell lines and patients—have been archived for public access. For example, GEO currently has archived over 5,550 microarray data sets on >150K different biomedical samples and human patients with >1,500 sets for cancer alone. Furthermore, despite their technical differences, microarray data sets from different time points, different laboratories, and even different platforms contain quite consistent information for many genes' expression patterns, so that we can successfully perform our investigations across those different genomic data sets. This large and rapidly increasing compendium of data demands data mining approaches, particularly association rule mining ensures that genomic data mining will continue to be a necessary and highly productive field for the foreseeable future.

REFERENCES

[1] Tuzhilin A, Adomavicius G. Zaıane O, Goebel R, Hand D, Keim D, Ng R, "Handling very large numbers of association rules in the analysis of microarray data", Proc. of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, p. 396-404, 2002,.

[2] Doddi,S., Marathe,A., Ravi,S.S. and Torney,D.C, "Discovery of association rules in medical data", Med. Inform. Internet. Med.,vol. 26, p.25–33,2001.

[3] Stilou,S., Bamidis,P.D., Maglaveras,N. and Pappas,C., "Mining association rules from clinical databases: an intelligent diagnostic process in healthcare", Medinfo, vol.10, 1399–1403, 2001.

[4] Niehrs C, Pollet N, "Synexpression groups in eukaryotes", Nature, 402:483-487,1999.

[5] Liotta L, Petricoin E, "Molecular profiling of human cancer", NatRev Genet, 1:48-56, 2000,.

[6] White, R., Gene Transcription, Mechanisms and Control, Blackwell science, 2001.

[7] Aleksandar Icev, Carolina Ruiz, Elizabeth F. Ryder, "Distance-based Association Rules Mining," BIOKDD03: Proc. 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics, p.34– 40,2003.

[8] Barker, W.C., Garavelli, J.S., Hou, Z., Huang, H., Ledley, R.S., McGarvey, P.B., Mewes, H.W., Orcutt, B.C., Pfeiffer, F., Tsugita, A., Vinayaka, C.R., Xiao, C., Yeh, L.S., Wu,C., "Protein Information Resource: a community resource for expert annotation of protein data", Nucleic Acids Res. 29, p.29- 32, 2001.

[9] Boeckmann, B., et al. The SWISS-PROT protein knowledge base and its supplement TrEMBL. Nucleic Acids Res. 31,p.365–370, 2003.

[10] Judice L.Y.Koh1,2, Mong Li Lee2, Asif M. Khan1, Paul T.J. Tan1 and Vladimir Brusic, "Duplicate Detection in Biological Data using Association Rule Mining", Proc. of the Second European Workshop on Data Mining and Text Mining in Bioinformatics, p.34–41,2005,.

[11] Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, Cooper DN, Human Gene Mutation Database - a biomedical information and research resource, Human Mutation 15(1):45-51,2000.

[12] P.D. Lewis, J.S. Harvey, E.M. Waters, and J.M. Parry,The Mammalian Gene Mutation Database,Mutagenesis,15(5):411- 414,2000.

[13] Lewin, Benjamin. Genes VI. Oxford; New York: Oxford University Press, 1997.

[14] Pevsner P.A., Lysov Y., Khrapko K.R., Belyavsky A., Floreny'ev, Mirzabekov A, Improved Chips for Sequencing by Hybridization. Journal of Biomolecular Structure and Dynamics 9(2), pp 399-410, 1991.

[15] G.Cong, Anthony K. H. Tung, X. Xu, F. Pan, and J. Yang., "Farmer: Finding interesting rule groups in microarray datasets". Proc. Of the 23rd ACM SIGMOD International Conference on Management of Data, 2004.

[16] Xin Xu Gao Cong Beng Chin Ooi Kian-Lee Tan Anthony K. H. Tung "Semantic Mining and Analysis of Gene Expression Data", Proc. of the 30th VLDB Conference, Toronto, Canada, 2004, p-1261-1264.

[17] K. Wang, L. Tang, J. Han and J. Liu, "Top down Fp-growth for association rule mining", Proc. of the 6th Pacific Area conference on Knowledge Discovery and Data Mining, Taipei, Taiwan, 2002.

[18] E Javier Lopez, Armando Blanco, Fernando Garcia, Antonio Marin, "Extracting Biological Knowledge by association rule mining", IEEE Trans. 1-4244-1210-2,2007.

[19] Fengjun Han, Nini Rao, "Mining Co-regulated Genes using Association Rules combined with Hast-tree and Genetic algorithms", IEEE Xplore, p 858-862,2009.

[20] D. J Allocco, I. S Kohane and A. J Butte: "Quantifying the relationship between co-expression, co-regulation and gene function", BMC Bioinformatics, 5:18 2004.

[21] K. Y. Yeung, M. Medvedovic and R. E. Bumgarner, From coexpression to co-regulation: how many microarray experiments do we need?, Genome Biology, 5:R48, 2004.

[22] Liping Ji and Kian-Lee Tan, "Mining gene expression data for positive and negative co-regulated gene clusters", Bioinformatics Vol.20 no.16, page 2711-2718 dol: 10.1093,2004.

[23] He Y, Hui SC, "Exploring ant-based algorithms for gene expression data analysis", Intell Med (2009), doi:10.1016/j.artmed.2009.03.004.