Bi-k-bi clustering: mining large scale gene expression data using two-level biclustering

Levent Çarkacıoğlu

Department of Computer Engineering,
Middle East Technical University,
Ankara, Turkey
E-mail: leventc@ceng.metu.edu.tr

Rengül Çetin Atalay and Özlen Konu

Department of Molecular Biology and Genetics,
Bilkent University,
Ankara, Turkey
E-mail: rengul@bilkent.edu.tr
E-mail: konu@fen.bilkent.edu.tr

Volkan Atalay and Tolga Can*

Department of Computer Engineering,
Middle East Technical University,
Ankara, Turkey
E-mail: volkan@ceng.metu.edu.tr
E-mail: tcan@ceng.metu.edu.tr
*Corresponding author

Abstract: Due to the increase in gene expression data sets in recent years, various data mining techniques have been proposed for mining gene expression profiles. However, most of these methods target single gene expression data sets and cannot handle all the available gene expression data in public databases in reasonable amount of time and space. In this paper, we propose a novel framework, *bi-k-bi clustering*, for finding association rules of gene pairs that can easily operate on large scale and multiple heterogeneous data sets. We applied our proposed framework on the available NCBI GEO *Homo sapiens* data sets. Our results show consistency and relatedness with the available literature and also provides novel associations.

Keywords: biclustering; APD; association pattern discovery; Spearman rank correlation; gene expression analysis.

Reference to this paper should be made as follows: Çarkacıoğlu, L., Atalay, R.Ç., Konu, Ö., Atalay, V. and Can, T. (2010) 'Bi-k-bi clustering: mining large scale gene expression data using two-level biclustering', *Int. J. Data Mining and Bioinformatics*, Vol. 4, No. 6, pp.701–721.

Biographical notes: Levent Çarkacıoğlu received his MSc in Computer Engineering at the Middle East Technical University, Turkey, in 2002.

He is currently a senior PhD student of the Middle East Technical University, Department of Computer Engineering. His primary research interests include issues related to data mining, database management systems, large scale databases, bioinformatics and J2EE (Java) frameworks.

Rengül Çetin Atalay is an Associate Professor of Molecular Biology. She obtained her PhD Degree from Universite de Paris-Sud, Orsay, France in 1997 and MD Degree from Hacettepe Univ. Medical School in 1992. She worked as a research assistant at Ecole Polytechnique, France during her PhD, and as an Assistant Professor at Virginia Bioinformatics Institute, Virginia, during her sabbatical leave in 2004. Her research focuses on application of computational techniques for the analysis of post genomic data in relation to liver cancer *in silico* and *in vivo*.

Özlen Konu received her PhD in Biology at the Texas Tech University, USA, in 1999. She is currently an Assistant Professor of the Bilkent University, Department of Molecular Biology and Genetics. Her primary research interests include highthroughput data analysis of cellular signalling pathways and comparative genomics.

Volkan Atalay is a Professor at Department of Computer Engineering, Middle East Technical University, Ankara, Turkey. His main research interests include pattern recognition in bioinformatics and computer vision. He has obtained PhD in Computer Science from Universite Paris Descartes, Paris, France in 1993.

Tolga Can received his PhD in Computer Science at the University of California at Santa Barbara in 2004. He is currently an Assistant Professor of the Department of Computer Engineering, Middle East Technical University, Ankara, Turkey. His main research interests are in bioinformatics, especially protein structure analysis and analysis of protein-protein interaction networks, and statistical methods such as graphical models and kernel methods.

1 Introduction

Microarray technology allows researchers simultaneously monitor expression levels of thousands of genes in a single experiment. These experiments are valuable tools in the understanding of genes, biological networks, and cellular states. Studies of microarray experiments have targeted many important goals, such as: finding differentially expressed genes, defining pathways, drug targeting, co-expressed gene clustering. Therefore, the most important goals of these studies is to survey patterns of gene expressions (Eisen et al., 1998; Jiang et al., 2004).

A microarray sample is a biological experiment performed on a single microarray chip to monitor expression levels of thousands of genes. Microarray samples experimented for specific studies are grouped into data sets. NCBI Gene Expression Omnibus (GEO) project is a public repository for microarray sample submissions from all over the world (Barrett et al., 2007). NCBI GEO hosts tens of thousands of microarray samples, grouped into hundreds of data sets, done for different purposes and submitted by different experimenters. Since many experiments were done on the same organism with a common gene set, it has been a challenge for scientists to mine gene expression data compiled at such a large and heterogeneous scale.

In recent years, biclustering and Association Pattern Discovery (APD) (a.k.a Association Rule Mining (ARM)) methods have been proposed to discover genes with similar expression profiles in a subset of conditions (samples) (Berrar et al., 2001; Cheng and Church, 2000; Kotala et al., 2001).

Cheng and Church (2000) introduced the concept of biclustering for gene expression data and proposed a greedy approach based on a uniformity criterion. Cheng and Church (2000) used randomly generated values to replace missing values in the data set. This approach is likely to reduce the quality of the discovered biclusters. To address this problem, Yang et al. (2003) proposed a probabilistic algorithm (FLOC) that handles missing values and discovers higher quality, overlapping biclusters.

Ben-Dor et al. (2002) introduced the Order Preserving Submatrix (OPSM) model for biclustering that focuses on the coherence of relative order of conditions rather than the coherence of expression values. Liu et al. (2004) improved the OPSM model by allowing some flexibility among conditions in an order equivalent group. Their model, OP-Cluster, is able to tolerate the effect of noise that reduces the efficiency of the more strict OPSM model.

Pattern based clustering problem proposed by Wang et al. (2002) is related to the biclustering problem. The p-Clustering algorithm developed by Wang et al. (2002) is able to discover genes that exhibit similar expression patterns in a subset of conditions. Pei et al. (2003) improved the p-Clustering algorithm by proposing a more efficient and scalable method, MaPle, to find maximal pattern-based clusters. All of the biclustering methods mentioned above discover clusters of genes that behave similarly in a subset of experimental conditions. However, the computational resources required by existing biclustering methods do not allow for analysis of gene expression data on a large scale. The proposed methods generally work on a single gene expression data set and cannot handle large number of gene expression data sets in public databases in reasonable amount of time and space. As a solution we propose a two level biclustering approach that works at the data set and experiment (i.e., condition) levels and discovers similar behaving gene pairs in multiple data sets. Our approach does not produce a subset of genes in a subset of conditions; rather, we report *pairs of genes* that behave similarly in a subset of conditions. This property is a setback compared to existing methods; however, it allows for mining gene expression patterns on a larger scale on a desktop computer with limited computational resources. In our experiments, the available biclustering methods we tested were not able to produce an output for the Breast Cancer Data Set with about 20,190 genes and 188 conditions.

Similar to biclustering methods, APD methods can be used to discover genes with similar expression profiles in a subset of conditions. APD methods were first used to discover associations among subsets of items from large transaction databases (Agrawal et al., 1993). APD methods detect sets of elements that frequently co-occur in a database and establish relationships between them of the form of $X \rightarrow Y$, meaning that, when X occurs it is likely that Y also occurs (Agrawal et al., 1993).

Kotala et al. (2001), Creighton and Hanash (2003), Becquet et al. (2002), Tuzhilin and Adomavicius (2002), and Georgii et al. (2005) have worked on associations and relationships among subsets of genes (e.g., $X = \{\text{Condition}_1$ (sample_1), Condition_2 $(\text{sample}_2)\}$, $Y = \{\text{gene } A\uparrow, \text{ gene } B\downarrow, \text{ gene } C\uparrow\}$, which means, in Condition_1 and Condition_2 when gene A is up-regulated, gene B is down-regulated and gene C is also up-regulated).

Most of the studies with APD methods work on homogeneously curated one or two data sets experimented for a specific study. APD methods that work on binary data use thresholding on the expression values of genes (Georgii et al., 2005). It is clear that a crude discretisation such as using thresholding lead to certain loss of information. As a solution, Georgii et al. (2005) have proposed a quantitative ARM approach. However, microarray experiments are generally not sufficiently robust and noise-resistant. It is not easy to decide whether or not an association of a small quantity is noise. Therefore, quantitative associations among genes do not generally give valuable information to a biologist. Another disadvantage of existing APD methods is that these methods try to find out rules over the whole set of genes. However, focusing on association rules among genes having similar expression profiles reduces the working data set by eliminating uncorrelated data patterns and is therefore expected to give more accurate results in a more efficient manner.

In this paper, we combine ideas from biclustering and association pattern discovery approaches and propose a novel framework, *bi-k-bi clustering*, for finding association rules of gene pairs that can easily operate on large scale and multiple heterogeneous data sets. Our framework mainly consists of a preprocessing phase followed by three phases. In the preprocessing phase, we scale gene expression values of all data sets into a comparable platform and compute an all-to-all similarity measure over the whole set for finding gene pairs with similar expression profiles. In the first phase of the bi-k-bi clustering, we apply biclustering on gene pairs in order to eliminate unrelated gene pairs and data sets. In the second phase, we assign labels to each gene in each sample of the working data set to indicate expression levels (high-expressed or low-expressed). Finally, we generate rules of gene pairs associated with samples by applying biclustering on the working set. Our method outputs clusters of labelled gene pairs with their associated microarray samples. We aim to help biologists to discover significant relationships among gene pairs and work easily on these relationships by concentrating on their associated microarray samples. Our contributions in this study can be summarised as follows:

- By scaling all expression values into a comparable platform, our framework enables working on multiple data sets. However, previous studies in this area generally work on a single gene expression data set.
- By defining a similarity measure and applying biclustering using this measure at the data set level, we automatically generate a reduced working set consisting of gene pairs with similar expression profiles with their associated data sets. This reduction enables focusing on correlated gene pairs therefore, it becomes easy to work on large scale data.
- We use dynamic thresholding by applying *k*-means clustering, for assigning expression level labels (i.e., High-expressed/Low-expressed) to genes.

- In this study, we extend a maximum frequent itemset algorithm to a biclustering algorithm and use it for biclustering in an efficient manner whereas available biclustering algorithms have high space complexity. Furthermore, our two-level biclustering strategy filters out unrelated gene pairs in the first level for increased efficiency.

We applied our bi-k-bi clustering framework on all available NCBI GEO *Homo sapiens* data sets. Gene pairs in the resulting clusters were classified as housekeeping genes with steady yet high expression values (i.e., ranks). Since, housekeeping genes are expected to have similar and robust behaviour across different experimental treatments this result is consistent with our expectations. As another study, we applied our framework on five different groups of selected NCBI GEO data sets (i.e., Breast Cancer, Normal Human Tissue, Obesity, Liver, and Colon). Results in these groups also showed consistency and relatedness with the available literature. Novel transcript associations were revealed and contextually analysed using the available literature. Moreover, existence of gene pairs were searched in protein-protein interaction databases to recover those pairs that act in concert at the transcriptional as well as post-transcriptional levels.

Our manuscript is organised as follows. First, we give detailed information about methods, metrics, and implementation of the bi-k-bi clustering framework. Then, we provide an illustrative example to explain each step of the framework. In Section 3, we discuss the results of our bi-k-bi clustering framework applied on NCBI GEO data sets. Conclusions are provided in the last section.

2 Methods

We follow a three step methodology in our study. In the first step we apply normalisation in order to scale all microarray samples in our database into a comparable platform. In the second step, we define a function to find gene pairs with similar expression profiles from the first step. Finally, we apply the bi-k-bi clustering algorithm to the gene pairs that have similar expression profiles and construct rules consisting of gene pairs and associated samples. The overview of the proposed methodology is given in Figure 1 and explained in detail here.

2.1 Data preperation

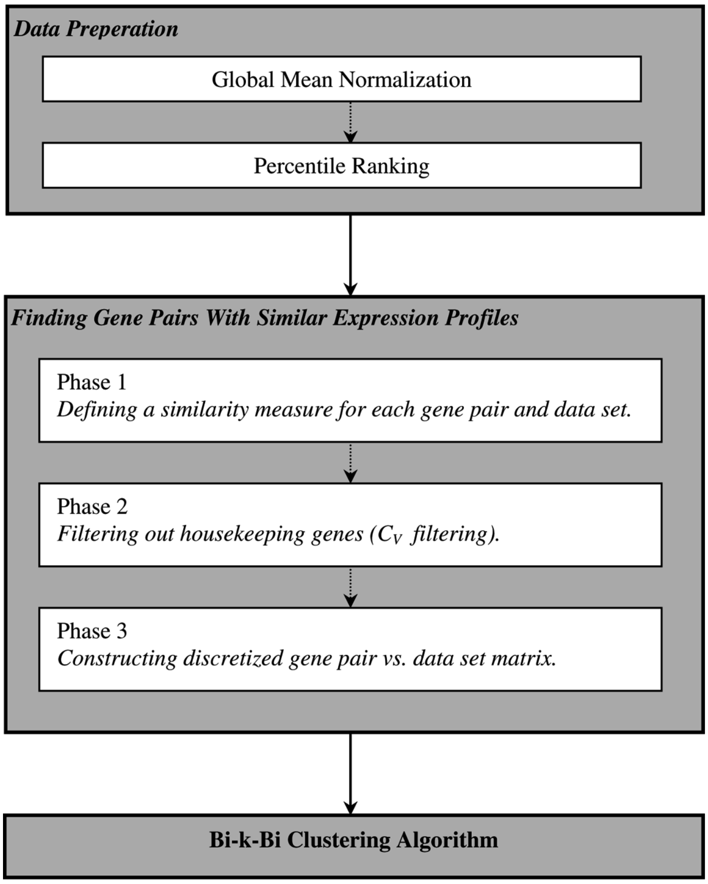
We have downloaded the *Homo sapiens* microarray data sets from NCBI GEO. In order to analyse gene expression values across several microarray experiments, we define and apply a two step methodology on each sample in our database:

- global mean normalisation over all spots in a microarray sample
- percentile ranking over the normalised spots within the same sample.

There exist many normalisation methods to scale gene data into a comparable platform (Yang et al., 2002). We preferred global mean normalisation since it is one of the least intrusive methods available and does not alter the distribution of microarray data. Accordingly, a log-based sample mean value is subtracted from each spot's log based expression value, for each microarray sample in the database.

In order to scale the normalised expression values, percentile ranking have been applied on these globally normalised expression values of each microarray sample. It is important to note that, each spot is ranked among the other spots only within the sample it belongs to. We calculate, for each gene, an average change of its rank in a data set. The idea behind this approach is that some genes may have multiple probes within the same sample. Let $GS = \{G_1, G_2, \ldots, G_k\}$ be the set of genes and $S = \{e_1, e_2, \ldots, e_n\}$ be a set of n microarray samples. For each gene G_i in a sample e_i, a single rank value, $r(G_i, e_i)$ is computed. For a gene that occurs in multiple probes in a sample, if the variation across rank values of the spots of that gene probe does not exceed a threshold then the average gene probe rank value is used. Otherwise all the probes for that gene is ignored for this sample. Given a set S, that has n samples, we have at most n rank values for a gene G_i.

Figure 1 The overview of the three step methodology for mining gene expression data from multiple data sets


Data Preperation
Global Mean Normalization
Percentile Ranking
Finding Gene Pairs With Similar Expression Profiles
Phase 1
Defining a similarity measure for each gene pair and data set.
Phase 2
Filtering out housekeeping genes (C$_V$ filtering).
Phase 3
Constructing discretized gene pair vs. data set matrix.
Bi-k-Bi Clustering Algorithm

2.2 Finding gene pairs with similar expression profiles

In order to find gene pairs with similar expression profiles we follow a three phase procedure. In the first phase, we define a similarity measure for finding similar rank behaving gene pairs and calculate this measure for each gene pair and each data set. In the next phase, we apply a filter to eliminate gene pairs that consists of housekeeping genes. Finally, in the last phase, we construct a matrix from the values calculated in the first phase and apply thresholding to discretise this matrix.

Phase 1: Three commonly used similarity measures for finding similar behaving gene pairs are the Euclidean Distance, Pearson's Correlation Coefficient, and Spearman Rank Correlation Coefficient (ρ) (Balasubramaniyan et al., 2005). Euclidean Distance and Pearson's Correlation Coefficient methods are sensitive to magnitude and shape (Kyungpil et al., 2007). Whereas, Pearson's Correlation Coefficient assumes approximate Gaussian distribution of the points and therefore it is not robust for non-Gaussian distributions (Jiang et al., 2004; Kyungpil et al., 2007). On the other hand, Spearman Rank Correlation Coefficient does not require Gaussian distribution and it is more robust against outliers. Therefore, we decided to use Spearman Rank Correlation Coefficient in our analysis.

Spearman Rank Correlation Coefficient has the drawback of data loss due to consequence of ranking (Jiang et al., 2004). Since we look for gene pairs with similar patterns over scaled expression values, the drawback of ranking has a negligible effect over our analysis.

Spearman Rank Correlation Coefficient is also sensitive to missing values. One of the approaches for missing values is to ignore the missing value pairs during correlation computation. Since not every gene is observed in all microarray samples, missing expression value pairs are ignored during computation. In order to cope with the negative effect of the '*ignore*' approach, Weighted Spearman Rank Correlation Coefficient, ρ_w is used in our analysis.

Let $S = \{S_1, S_2, \ldots, S_m\}$ be the data sets in the database where $S_i = \{e_{i1}, e_{i2}, \ldots, e_{in}\}$ is a set having n microarray samples, where $n \geq 3$ and $GS = \{G_1, G_2, \ldots, G_k\}$ be the set of distinct genes in the database. Also let P_S be the percentage of non-ignored value pairs during Spearman Rank Correlation Coefficient calculation. We define a log based non linear weight function given in equation (1). Since $P_S \in [0, 100]$ then by using equation (1), $f_w(P_S) \in [1, 2]$.

$$f_w(P_S) = \begin{cases} \log(P_S) & \text{for} \quad P_S > 10 \\ \log(10) & \text{for} \quad P_S \leq 10. \end{cases} \tag{1}$$

Let $r(G_i, e_k)$ be the average percentile ranked expression value of gene G_i and $r(G_j, e_k)$ be the average percentile ranked expression value of gene G_j in the microarray sample e_k of the data set S_a. We then define the Weighted Spearman Rank Correlation Coefficient as in equation (3) by using the equation (2).

$$R(G_{i,j}, S_a) = \{\forall e_k \in S_a \,|\, (r(e_k, G_i),\ r(e_k, G_j))\} \tag{2}$$

$$\rho_w(G_{i,j}, S_a) = \rho(R(G_{i,j}, S_a)) \times f_w(P_{Sa}). \tag{3}$$

Phase 2: Gene pairs that are not modulated in microarray samples are expected to (and does) have high ρ_w values. In this study we mainly focus on genes whose expression profiles are similarly affected among samples. Therefore, genes whose expressions are not modulated in any of the experimental conditions, (i.e., housekeeping genes) does not have much impact on the association rules of genes which we aim to find.

Previous studies by Hsiao et al. (2001), and Eisenberg and Levanon (2007) have provided gene lists involved in cellular maintenance functions; thus these genes are called housekeeping genes that are generally assumed to have expression levels unaffected by experimental conditions. However, recent studies indicated that several widely used housekeeping genes might have altered expression under different experimental conditions (Vandesompele et al., 2002; Warrington et al., 2000). Therefore, in order to eliminate genes that are not effected in microarray samples, we use the Coefficient of Variation, C_v, as a filter.

Let $C_v(G_i, S_a)$ be the Coefficient of Variation of gene G_i and $C_v(G_j, S_a)$ be the Coefficient of Variation of gene G_j calculated using the percentile ranked expression values of microarray samples of the data set S_a, as defined in equation (4). Let also, $\rho_w(G_{i,j}, S_a)$ be the Weighted Spearman Rank Correlation Coefficient for $G_{i,j}$ among the expression values of G_i and G_j in the data set S_a, calculated by using equation (3). We then apply a filter on $\rho_w(G_{i,j}, S_a)$ using the C_v values which is defined in equation (5).

$$C_v(G_i, S_a) = C_v(\{r(G_i, e_k)\}) \;\; \forall e_k \in S_a \tag{4}$$

$$\rho_w(G_{i,j}, S_a) = \begin{cases} 0 \;\text{ if }\; C_v(G_i, S_a) \leq t_{Cv} \;\text{ and }\; C_v(G_j, S_a) \leq t_{Cv} \\ \rho_w(G_{i,j}, S_a) \;\text{ otherwise.} \end{cases} \tag{5}$$

Phase 3: For each gene pair, $G_{i,j} = (G_i, G_j)$ where $G_i, G_j \in GS$, and each data set S, we calculate Weighted Spearman Rank Correlation Coefficient by equation (5).

Given a query gene G_i, this calculation forms out an $k \times m$ matrix, **A**, with k rows and m columns. This matrix is defined by its set of rows, $X = \{G_{i1}, G_{i2}, G_{i3}, \ldots, G_{ik}\}$ for the given gene and its set of columns, $Y = \{S_1, S_2, \ldots, S_m\}$.

A cell in this matrix, a_{MN}, is a real value representing the Weighted Spearman Correlation Coefficient of Mth gene pair for the Nth working set as defined in equation (6).

$$a_{MN} = \rho_w(G_{i,j}, S_N) \;\text{ where } M = G_{i,j} \;\text{ and } N = S_N. \tag{6}$$

Since $\rho \in [-1, +1]$ and $f_w(P) \in [1, 2]$ then $\rho_w \in [-2, +2]$. Setting the threshold, $t_{\rho w} = 1.5$, for ρ_w makes $\rho \leq 0.75$; which is a reasonable threshold for two random variables that show similar behaviour.

Weighted Spearman Rank Correlation gives a scale on the strength of the similarity of two random variables. Therefore, in order to decide whether two genes behave similarly, we apply thresholding and discretise the matrix to mark gene pairs vs. sets showing similar behaviour. A cell, a_{MN}, in the matrix is then defined in equation (7).

$$a_{MN} = \begin{cases} 1 \;\; \rho_w(G_{i,j}, S_N) \geq t_{\rho w} \;\text{ where } M = G_{i,j} \;\text{ and }\; N = S_N \\ 0 \;\; \rho_w(G_{i,j}, S_N) < t_{\rho w} \;\text{ where } M = G_{i,j} \;\text{ and }\; N = S_N. \end{cases} \tag{7}$$

There exist approximately 20,200 distinct genes among the 371 NCBI GEO data sets complied in our database. Therefore, the matrix we used in our analysis has more than 150,000,000,000 elements. By the use of the symmetry property of Spearman Rank Correlation Coefficient, thresholding, filtering weight values during ρ_w calculation and using the fact that microarray samples generally do not include all genes, we had performed 12,000,000,000 ρ_w calculations to construct this matrix.

2.3 Bi-k-bi clustering algorithm

The bi-k-bi clustering algorithm was applied in three steps (Figure 2 and Algorithm 1) which operates on the discretised matrix, **A**, defined in Section 2.2.

In the first step we perform a coarse analysis in order to reduce the working set by focusing on gene pairs with their associated data sets in which they have similar expression profiles. The second step is the assignment of labels for genes in microarray samples by k-means clustering over all available gene expression data in our database. Finally, a detailed analysis of labelled gene pairs associated with microarray samples, so called rules, is performed. As it can be deduced, the first and last step should be done in order, however, the second step, where we assign a label for each gene in each microarray sample, can be performed in any order.

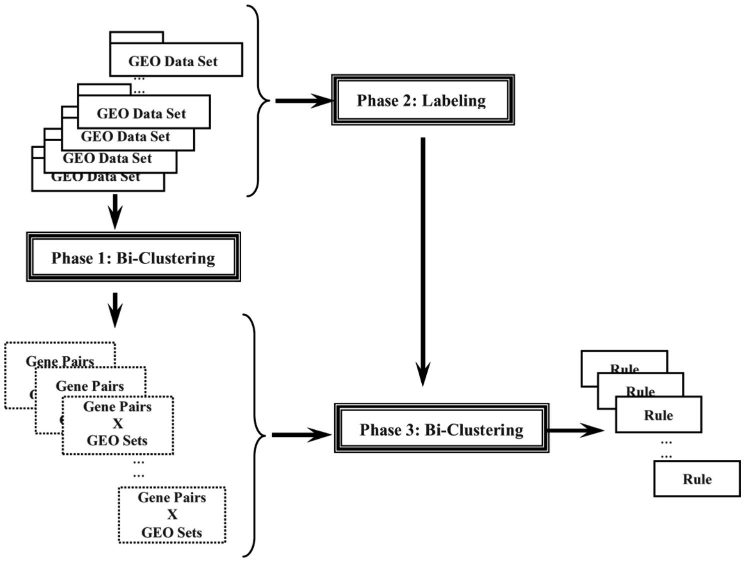
Algorithm 1 Bi-k-Bi Clustering Algorithm

```
Find clusters of similar behaving gene pairs versus sets using the biclustering
  algorithm, as described in Section 2.3.
Associate a label to each gene in each experiment using k-means clustering
  algorithm, as described in Section 2.3.
loop
  For each cluster of similar behaving gene pairs
  Construct (sample,labeled gene pairs) using the ClusterLabel function defined
    in Equation (11).
end loop
Find clusters of labeled gene pairs versus experiments, called as rules, using
  biclustering algorithm, as described in Section 2.3.
```

Finding clusters of gene pairs that have similar expression profiles: After the discretisation process of the matrix from the finding similar gene expression pair step (Figure 1), we first focus on identification of both groups of gene and set pairs in this matrix. It is important to keep in mind that we should find the groups of gene pairs and sets having 1's in the matrix which represent similarly behaved gene pairs after bi-k-bi clustering algorithm. Biclustering algorithms are used for this purpose and there exists several of them (Cheng and Church, 2000; Jiang et al., 2004; Wang et al., 2002; Zhao and Zaki, 2005) and tools available in the literature. Initially in this study we used a freely available Biclustering Analysis Toolbox (BiCAT) (Barkow et al., 2006) however due to the large size of our matrixes we faced memory problems. For this reason, we designed and implemented a time and memory efficient biclustering algorithm.

Association Pattern Discovery (APD) methods have been applied on gene expression data in order to find out groups of co-regulated gene patterns (Creighton and Hanash, 2003; Carmona-Saez et al., 2006; Georgii et al., 2005; Gyenesei et al., 2007). APD originates from market basket analysis and aim to find interesting relationships hidden in large data sets. Such relationships can be represented as frequent itemsets and association rules. APD methods are inherited from the area of frequent itemsets and association rule mining. However, studies done on this area have commonly been focused on finding gene patterns as association rules (Creighton and Hanash, 2003; Carmona-Saez et al., 2006; Georgii et al., 2005; Gyenesei et al., 2007).

Figure 2 Workflow of the bi-k-bi clustering algorithm



Maximum Frequent Itemset (MFI) in transactional databases is the problem of mining maximum itemsets from the transactional database. Thus, in a given set of items and transaction set, MFI algorithms find out maximum sets of items occur for a given support. For example: for support v, items that occur at least in v transactions are reported. It is important to note that MFI reports the maximum itemsets and does not report subsets. There have been many studies for mining frequent itemsets in the literature (Agarwal et al., 2000, 2001; Gouda and Zaki, 2001).

MAFIA is one of the MFI algorithms which performs best when mining long itemsets and it outperforms other algorithms on dense data (Burdick et al., 2001). The algorithm applies space pruning techniques and adaptive compression that makes optimal use of memory and running time. A free implementation with source code of MAFIA is publicly available (http://himalaya-tools.sourceforge.net/Mafia/doxygen-Mafia/index.html).

MAFIA, like most MFI algorithms, outputs the list of frequent itemsets for a given support. Applying a post processing on this output by adding the associated transactions of the itemsets, we can have biclusters of itemsets and transactions. (i.e., biclusters having number of transactions greater than the support, v). The post processing has at most $O(n^2)$ time complexity, which does not have much effect on the running time of the original MAFIA.

According to this approach, we modified the MAFIA algorithm in order to output associated transactions with the itemsets. We use this modified algorithm as our biclustering algorithm. We then represent transactions as data sets and itemsets as gene pairs with similar expression profile using the discretised matrix defined in Section 2.2.

Let $G_{i,j} \in GS$ and $S_k \in S$ and v_1 be the minimum number of sets in the clustered results; applying our biclustering algorithm gives clusters as in equation (8).

$$\text{Cluster}_c = [\{G_{i,j}\}, \{S_k\}] \quad \text{where size of } (S) \geq v_1. \tag{8}$$

Labelling gene pairs in experiments. In order to use association rules among the biclustered gene pairs, gene pairs with similar expression profiles (i.e., biclusters constructed in the previous subsection) should be labelled. In this step, we assign a label for each gene pair with the corresponding experiment in which it occurs.

In order to label gene pairs in the prepared data, described in Section 2.1, a preprocessing step should be applied. In this preprocessing step, rank values of the genes in the experiments are discretised. High rank values are labelled with **High** (‘*High-expressed*’), low rank values with **Low** (‘*Low-expressed*’).

It is clear that a crude discretisation such as using thresholding on rank data lead to certain loss of information (Georgii et al., 2005). In order to alleviate this loss of information we decided to use clustering on the rank values for labelling. Since ***k*-means clustering** is a fast and efficient clustering algorithm, we apply k-means clustering on all rank values and assign a label for each gene in each experiment using these clusters. When the number of clusters (i.e., $k = 2$) are known k-means clustering is also advantageous in both time and space complexity.

Let $S = \{e_1, e_2, \ldots, e_n\}$ be a set of n microarray experiments, $G_{i,j}$ be a gene pair in S and $e_k \in S$. Also, let $l_{i,k}$ be the label assigned to gene G_i in e_k and $l_{j,k}$ be the label for gene G_j within the same experiment e_k (i.e., $l_{i,k}, l_{j,k} \in \{High, Low\}$). We define the gene pair experiment labelling function (*ExpLabel*) for the gene pair $G_{i,j}$ within the experiment e_k as in equation (9).

$$ExpLabel(G_{i,j}, e_k) = l_{i,j,k} \quad \text{where } l_{i,j,k} = [G_i(l_{i,k}), G_j(l_{j,k})]. \tag{9}$$

For a given set, we use the *ExpLabel* function defined in equation (9) and define the set labelling function (*SetLabel*) as in equation (10).

$$SetLabel(G_{i,j}, S) = \{ExpLabel(G_{i,j}, e_k)\} \quad \forall e_k \in S. \tag{10}$$

Let C be a bicluster, we then define the cluster labelling function (*ClusterLabel*) as in equation (11).

$$ClusterLabel(C) = \{SetLabel(G_{i,j}, S_m)\} \quad \forall S_m, \quad \forall G_{i,j} \in C. \tag{11}$$

Finally, we apply the *ClusterLabel* function in equation (11) to all biclusters found in equation (8) and construct the data to be used in the third step of our bi-k-bi clustering algorithm.

Extracting rules. In the third step of our bi-k-bi clustering algorithm we aim to find association rules as clusters of labelled gene pairs vs. experiments. Thus, we apply biclustering algorithm on the clusters of gene pairs with similar expression profiles obtained as a result of k-means clustering. We use the same biclustering algorithm described in Section 2.3 and find out sets of clusters, which we call them as rules, for a given support v_2 (i.e., number of experiments in the rule is greater than v_2).

Let e_k be an experiment and $G_{i,j}$ be a gene pair in or database. Further let v_1 be the support for the minimum number of data sets, gene pair $G_{i,j}$ have similar expression profiles and v_2 be the support for the minimum number of experiments in a rule. This final step of the algorithm outputs sets of rules as of the form given in equation (12).

$$Rule = [\{ExpLabel(G_{i,j}, e_k)\}, \{e_k\}] \quad \text{for support } v_1 \text{ and } v_2. \tag{12}$$

2.4 Illustrative example

As an illustrative example, consider the randomly selected sample genes Gene$_1$, Gene$_2$, Gene$_3$ and NCBI GEO data sets GDS1, GDS2 and GDS3 from our database. The average percentile rank values of these sample genes are given in Table 1.

Table 1 Average percentile rank values of the genes for the illustrative example

		Average percentile rank		
NCBI-GEO GDS	*NCBI-GEO GSM*	*Gene$_1$*	*Gene$_2$*	*Gene$_3$*
GDS1	GSM11	36	51	45
	GSM12	51	60	65
	GSM13	38	57	–
	GSM14	54	–	–
	GSM15	–	–	55
	GSM16	55	–	44
GDS2	GSM21	62	90	35
	GSM22	59	86	21
	GSM23	44	84	42
GDS3	GSM31	87	87	60
	GSM32	94	93	46

First, we find gene pairs with similar expression profiles as described in Section 2.2. For this purpose, we define thresholds $t_{Cv} = 0.2$ and $t_{\rho w} = 1.5$. We then compute similarity measures (P_S, C_v, ρ, ρ_w) for each gene pair. The computed values for this illustrative example are given in Table 2. By using these computed values and thresholds, we prepare the input data for the bi-k-bi clustering framework as given in Table 3.

We apply the bi-k-bi clustering framework on the discretised matrix generated from the gene pairs with similar expression profiles and data sets.

In the first phase of the framework, we apply biclustering and generate clusters as described in Section 2.3. The resulting cluster for this illustrative example with support $v_1 = 2$ is given in Table 4.

In the second phase of the framework, we apply k-means to assign **High/Low** labels to the genes and generate (sample, labelled gene pairs) for the clusters as described in Section 2.3. List of labels for the genes in the microarray samples and the cluster for this illustrative example are given in Tables 5 and 6 respectively.

Table 2 Similarity measure values (P_S, C_v, ρ, ρ_w) of the gene pairs in the illustrative example where $t_{Cv} = 0.2$

Gene pair	*GDS 1*	*GDS 2*	*GDS 3*
Gene$_1$–Gene$_2$	$P_S = 50$ $C_v(\text{Gene}_1) = 0.19$ $C_v(\text{Gene}_2) = 0.08$ $\rho = 1$ $\rho_w = 1.69$	$P_S = 100$ $C_v(\text{Gene}_1) = 0.17$ $C_v(\text{Gene}_2) = 0.03$ $\rho = 1$ $\rho_w = 2$	$P_S = 100$ $C_v(\text{Gene}_1) = 0.05$ $C_v(\text{Gene}_2) = 0.04$ $\rho = 1$ $\rho_w = 2$
Gene$_1$–Gene$_3$	$P_S = 50$ $C_v(\text{Gene}_1) = 0.19$ $C_v(\text{Gene}_3) = 0.18$ $\rho = -1$ $\rho_w = -0.84$	$P_S = 100$ $C_v(\text{Gene}_1) = 0.17$ $C_v(\text{Gene}_3) = 0.32$ $\rho = 0.5$ $\rho_w = 0$	$P_S = 100$ $C_v(\text{Gene}_1) = 0.05$ $C_v(\text{Gene}_3) = 0.18$ $\rho = -1$ $\rho_w = -2$
Gene$_2$–Gene$_3$	$P_S = 33$ $C_v(\text{Gene}_2) = 0.08$ $C_v(\text{Gene}_3) = 0.18$ $\rho = 1$ $\rho_w = 1.51$	$P_S = 100$ $C_v(\text{Gene}_2) = 0.03$ $C_v(\text{Gene}_3) = 0.32$ $\rho = -0.5$ $\rho_w = 0$	$P_S = 100$ $C_v(\text{Gene}_2) = 0.04$ $C_v(\text{Gene}_3) = 0.18$ $\rho = -1$ $\rho_w = -2$

Table 3 NCBI-GEO data sets vs. gene pairs in the illustrative example where $t_{\rho w} \geq 1.5$

NCBI-GEO GDS	*Gene pairs*
GDS1	[Gene$_1$, Gene$_2$], [Gene$_2$, Gene$_3$]
GDS2	[Gene$_1$, Gene$_2$]
GDS3	[Gene$_1$, Gene$_2$], [Gene$_1$, Gene$_3$], [Gene$_2$, Gene$_3$]

Table 4 Clusters for the illustrative example with support $v_1 = 2$

Cluster name	*Cluster members*
$c_1(2 \times 2)$	{[Gene$_1$, Gene$_2$], [Gene$_2$, Gene$_3$]} × {GDS1, GDS3}

In the last phase of our framework, we apply biclustering and generate association rules as described in Section 2.3. The resulting rule for this illustrative example with support $v_2 = 3$ is given in Table 7.

3 Results

The proposed bi-k-bi clustering framework was applied on all available NCBI GEO *Homo sapiens* data sets (i.e., 9090 microarray samples grouped into 372 data sets). Our findings indicated that majority of the gene-pairs belonged to categories with known housekeeping gene functions, such as ribosomal protein genes and metabolic pathway genes. Housekeeping genes generally are assumed to have expression levels unaffected by experimental conditions thus are expected to exhibit relatively stable expression at high levels. Coefficient of Variation parameter of the bi-k-bi clustering approach allowed us to determine the most stably expressed genes in the database (Carkacioglu et al., 2006). By applying an experimentally defined C_v filter threshold on the C_v, i.e., 0.1, we filtered out the likely housekeeping genes thereby leaving gene-pairs potentially involved in cellular signalling processes as well as those function among tissues and pathological states in highly divergent manners.

Table 5 K-means clustering for the assignment of gene expression levels in the illustrative example: High and Low

		Assigned labels		
NCBI-GEO GDS	*NCBI-GEO GSM*	*Gene*$_1$	*Gene*$_2$	*Gene*$_3$
GDS1	GSM11	LOW	HIGH	LOW
	GSM12	HIGH	HIGH	HIGH
	GSM13	LOW	HIGH	–
	GSM14	HIGH	–	–
	GSM15	–	–	HIGH
	GSM16	HIGH	–	LOW
GDS2	GSM21	HIGH	HIGH	LOW
	GSM22	HIGH	HIGH	LOW
	GSM23	LOW	HIGH	LOW
GDS3	GSM31	HIGH	HIGH	HIGH
	GSM32	HIGH	HIGH	LOW

Table 6 Expression level assigned gene pairs vs. NCBI-GEO microarray samples for the illustrative example

Labelled gene pairs	*NCBI-GEO GSM*
Gene$_1$(HIGH), Gene$_2$(HIGH)	GDS1_GSM12, GDS3_GSM31, GDS3_GSM32
Gene$_1$(HIGH), Gene$_2$(LOW)	–
Gene$_1$(LOW), Gene$_2$(HIGH)	GDS1_GSM11, GDS1_GSM13,
Gene$_1$(LOW), Gene$_2$(LOW)	–
Gene$_2$(HIGH), Gene$_3$(HIGH)	GDS1_GSM12, GDS3_GSM31
Gene$_2$(HIGH), Gene$_3$(LOW)	GDS1_GSM16, GDS3_GSM32,
Gene$_2$(LOW), Gene$_3$(HIGH)	–
Gene$_2$(LOW), Gene$_3$(LOW)	–

Table 7 Rules for the illustrative example with support $v_1 = 2$ and $v_2 = 3$

Rule name	*Rule members*

Rule$_1$ (2 × 3)	{[Gene$_1$(HIGH), Gene$_2$(HIGH)], [Gene$_2$(HIGH), Gene$_3$(HIGH)]}
	×
	{GDS1_GSM12, GDS3_GSM31, GDS3_GSM32}

However, the results after C_v filtering may still contain so called housekeeping genes that pass the C_v threshold since recent studies have indicated that several widely used housekeeping genes have altered expression under experimental conditions (Vandesompele et al., 2002; Warrington et al., 2000). One of the important aspects of the present study is its ability to associate a set of gene-pair rules with subsets of the available microarray data. Therefore, we applied our framework on five different groups of data sets:

- Breast Cancer
- Normal Human Tissue
- Obesity
- Liver
- Colon and extracted rules.

For each group we constructed a working set consisting of NCBI GEO microarray data sets (Barrett et al., 2007). We manually curated working sets by using the free text titles and descriptions supplied by experimenters in NCBI GEO data sets. Microarray sample and data set distribution of each working set is as follows:

- Breast Cancer: 188 microarray samples grouped into 10 data sets
- Normal Human: 120 microarray samples grouped into 5 data sets
- Obesity: 275 microarray samples grouped into 8 data sets
- Liver: 35 microarray samples grouped into 2 data sets
- Colon: 132 microarray samples grouped into 9 data sets.

The corresponding NCBI GEO data sets with their titles for each working set were provided in Table 8.

We applied the bi-k-bi clustering framework on each working set independently. C_v filter threshold, t_{Cv}, was set as 0.1 to specifically focus on gene-pairs with variable expression. Results for each subset can be accessed online (http://www.i-cancer.org/~levent/rules). A search engine that enables users to query genes within the result sets can also be accessed through the website (http://www.i-cancer.org/~levent/query).

The most observed gene-pair rule in the breast cancer subsets, FOXM1-TPX2, existed in 80% of the microarray samples analysed. Interestingly, a few of the experiment sets were included or excluded as a whole for the rule determination. In particular, FOXM1-TPX1 rule was restricted to all samples of GDS817, GDS820, GDS901 while this rule was not observed in any of the samples of the experiment sets GDS901 and GDS1250. On the other hand, different samples of GDS2250 were used for determining the FOXM1-TPX2 rule; for example, normal breast tissue expression together with six non-basal-like and one basal-like tumour samples did not contribute to the rule (Richardson et al., 2006). These results might indicate that normal cells do not exhibit FOXM1-TPX2 rule unlike subsets of breast tumours, i.e., basal-like. Previous studies support our findings such that Sotiriou et al. (2006) showed that FOXM1 and TPX2 were up-regulated in breast tumours.

Table 8 NCBI GEO datasets used in working sets

Working set	*NCBI GEO dataset*
Breast Cancer	**GDS360**: BREAST CANCER AND DOCETAXEL TREATMENT **GDS817**: BREAST CANCER CELL EXPRESSION PROFILES (HG-U95A) **GDS820**: BREAST CANCER CELL EXPRESSION PROFILES (HG-U133A) **GDS823**: BREAST CANCER CELL EXPRESSION PROFILES (HG-U133B) **GDS881**: BREAST CANCER AND SELECTIVE ESTROGEN RECEPTOR MODULATORS **GDS901**: ESTROGEN RECEPTOR ALPHA L540Q MUTATION EFFECT ON GENE INDUCTION BY ESTRADIOL: TIME COURSE **GDS1250**: ATYPICAL DUCTAL HYPERPLASIA AND BREAST CANCER **GDS1326**: BREAST CANCER CELLS REEXPRESSING ESTROGEN RECEPTOR ALPHA RESPONSE TO 17BETA-ESTRADIOL **GDS1329**: MOLECULAR APOCRINE BREAST TUMOURS **GDS2250**: BASAL-LIKE BREAST CANCER TUMOURS
Normal Human Tissue	**GDS422**: NORMAL HUMAN TISSUE EXPRESSION PROFILING (HG-U95A) **GDS423**: NORMAL HUMAN TISSUE EXPRESSION PROFILING (HG-U95B) **GDS424**: NORMAL HUMAN TISSUE EXPRESSION PROFILING (HG-U95C) **GDS425**: NORMAL HUMAN TISSUE EXPRESSION PROFILING (HG-U95D) **GDS426**: NORMAL HUMAN TISSUE EXPRESSION PROFILING (HG-U95E)
Obesity	**GDS268**: OBESITY AND FATTY ACID OXIDATIONC **GDS1480**: OBESITY: PREADIPOCYTE EXPRESSION PROFILE (HG-U133A) **GDS1481**: OBESITY: PREADIPOCYTE EXPRESSION PROFILE (HG-U133B) **GDS1493**: OBESITY: ADIPOCYTE EXPRESSION PROFILE (HG-U95A) **GDS1495**: OBESITY: ADIPOCYTE EXPRESSION PROFILE (HG-U95B) **GDS1496**: OBESITY: ADIPOCYTE EXPRESSION PROFILE (HG-U95C) **GDS1497**: OBESITY: ADIPOCYTE EXPRESSION PROFILE (HG-U95D) **GDS1498**: OBESITY: ADIPOCYTE EXPRESSION PROFILE (HG-U95E)
Liver	**GDS1373**: PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR SUBTYPE ACTIVATION EFFECT ON LIVER CELL **GDS1442**: PPARI AGONIST CIPROFIBRATE EFFECT ON LIVER
Colon	**GDS559**: INFLAMMATORY BOWEL DISEASE (HG-U133A) **GDS560**: INFLAMMATORY BOWEL DISEASE (HG-U133B) **GDS709**: ENTEROCYTE DIFFERENTIATION TIME COURSE **GDS756**: COLON CANCER PROGRESSION **GDS1263**: DUKES B COLON CANCER RECURRENCE **GDS1330**: CROHN DISEASE AND ULCERATIVE COLITIS COMPARISON **GDS1386**: COLORECTAL CARCINOMA SUBTYPE WITH MICROSATELLITE INSTABILITY (HG-U133A) **GDS1387**: COLORECTAL CARCINOMA SUBTYPE WITH MICROSATELLITE INSTABILITY (HG-U133B) **GDS1942**: TRANSGENIC KRAPPEL-LIKE FACTOR 4 INDUCTION: TIME COURSE

For the tissue expression pair-rules, the two most commonly observed composite rules were [CHD1 (HIGH) – PSD4 (HIGH)] [CHD1 (HIGH) – ZNF384 (HIGH)] and [MBD6 (HIGH) – PSD4 (HIGH)] [POLDIP2 (HIGH) – SFXN4 (HIGH)]. The first rule was determined based on the tissue samples from GDS422 and GDS423. Similarly, [MBD6 (HIGH) – PSD4 (HIGH)] [POLDIP2 (HIGH) – SFXN4 (HIGH)] rule was supported only by the GDS423 and GDS424 experiment sets. These findings suggested that representative sequences of the same gene might be non-equivalent among different platforms influencing the degree of the association due to either presence/absence of alternative splicing events, multiple hits in the transcriptome, and/or dinucleotide content of the probesets. Since [CHD1 (HIGH) – PSD4 (HIGH)] [CHD1 (HIGH) – ZNF384 (HIGH)] rule was commonly observed in all of the tissues studied (i.e., bone marrow, liver, heart, spleen, lung, kidney, skeletal muscle, thymus, brain, spinal cord, prostate, and pancreas), this triplet might represent an alternative reference gene set for non-diseased tissue normalisation studies. Moreover, CDH1, PSD4, and ZNF384 or MBD6 and PSD4, or POLDIP2 and SFXN4 genes were not shown to be co-expressed previously in the literature thus represent novel links in cellular signalling pathways.

The most commonly observed pair-rule in comparing the adipocyte expression profiling was [CRIP2 (HIGH) – RGS5 (HIGH)]; however this rule was not able to separate lean vs. obese type adipocytes. Furthermore, [CRIP2 (HIGH) – RGS5 (HIGH)] rule did not hold in none of the preadipocyte/stromal vascular cells (GDS1480) or lean/obese skeletal muscle tissue (GDS268) or some of the lean/obese adipocytes (i.e., some of GDS1493, all of GDS1496, all of GDS1497, some of GDS1498). The same issue observed for the tissue datasets explained in the previous paragraph was present for the lean/obese samples (20 non-obese. BMI $25+/-3\,\mathrm{kg/m^2}$, and 19 obese. BMI $55+/-8\,\mathrm{kg/m^2}$, non-diabetic Pima Indians; Lee et al. (2005)). GDS1493, GDS1495, GDS1496, GDS1497, and GDS1498 belonged to five different platforms (HG-U95A-E) and although were used for the same set of adipocytes. Nevertheless, each platform might contain different probesets for the same gene thus might not always support the [CRIP2 (HIGH) – RGS5 (HIGH)] or other rules.

In liver datasets, [CD38 (LOW) – PENK (LOW)] [C2ORF27 (LOW) – HELZ (HIGH)] [GYS1 (HIGH) – SNX9 (LOW)] rule was observed most commonly across almost all liver experiment sets. The rule displayed low-high, high-low, and low-low expression pairs suggesting positive and negative regulation on these gene-pairs and warrants further experimental confirmation in the context of liver cancer. Analysing the pair-rules in and experiment specific-manner rather than globally provided rules with functional importance. For example, composite rule [AKR1B10 (HIGH) – NQO1 (HIGH)] [TBC1D9B (HIGH) – UBE2G1 (HIGH)] [HADHB (HIGH) – VNN1 (HIGH)] [CYP4F2 (HIGH) – DIO1 (HIGH)] [CD38 (LOW) – PENK (LOW)] [C2ORF27 (LOW) – HELZ (HIGH)] [GYS1 (HIGH) – SNX9 (LOW)] was able to separate the long-term high dose treatment (i.e., 3 samples out of 4 150 mg/kg/day for 15 days and 4 samples out of 5 400 mg/kg/day for 15 days) effects of peroxisome proliferator-activated receptor-a agonist ciprofibrate, which caused hepatocellular carcinoma, from the vehicle control (5 out of 5) and low dose treatment (3 mg/kg/day and 30 mg/kg/day for 15 days) on primate liver samples (GDS1442; Cariello et al. (2005)). Therefore, this rule might represent a novel co-expressed gene set that is dose-dependently regulated by peroxisome proliferator-activated receptor-a during liver carcinogenesis.

Colon disease datasets investigated in the present study were more divergent in nature including colon cancer cell lines (metastatic/non-metastatic, treated/untreated) as well colon biopsy samples (normal, chrons disease, colitis) and primary tumours (with/without recurrence). The most commonly observed gene-pair rule, [MRPL12 (HIGH) – RP2 (LOW)] was not able to separate colon cancer samples from other colon diseases nor from cell line experiments. On the other hand, [GPR64 (LOW) – RRAD (LOW)] rule seem to cluster almost all colon cancer cell lines together although it fails to distinguish between colon cancer recurrence or colon diseases.

NCBI GEO sets consists of biological experiments realised for a specific purpose. For example, an NCBI GEO data set may contain microarray samples done on normal and cancer tissues together where normal tissue samples are often used as control samples by the experimenters. Microarray samples in rules generated by our framework consisted of only control or target samples of data sets included. Only small number of rules included mixtures of control and target samples of the same data set while misplaced samples in these rules represented generally duplicates of original samples. Since most rules were determined by the contribution of the majority of the members of a particular microarray experiment set this suggested that experimental conditions together with biological differences also played a role in the observed co-expression patterns of the genes.

A literature survey performed on the genes in the rules generated by our framework resulted in confirmation of expressional regulation of cancer-related genes. FOXM1, TPX2, HIST1H1C, HIST1H2BK, IFIT3, RSAD2, ISG15, HIST1H2BE, DBF4, STIL, KNTC2, MELK, IFIT2, DLG7, BUB1B, HCP5, OASL genes were the most commonly occurring up-regulated genes in rules found using the Breast Cancer Working Set. In previous breast cancer studies, some of these genes have been reported as being up-regulated including FOXM1 (Madureira et al., 2006; Sotiriou et al., 2006), TPX2 (Sotiriou et al., 2006), HIST1H1C and HIST1H2BK (Barry et al., 2005), KNTC2 (Lacroix, 2006), MELK (Lin et al., 2007), BUB1B (Fridlyand et al., 2006).

In the light of all these information, we can say that rules found using the proposed bi-k-bi clustering framework show consistency and relatedness with the literature.

4 Conclusion

As gene expression profile data sets became increasingly available, application of data mining techniques on these data gives valuable hints about gene pattern associations. Several methods have been proposed for mining gene expression profiles with potential limitations.

In this paper, we proposed a novel framework, bi-k-bi clustering, for finding association rules of gene pairs that can easily operate on large scale data. Our framework outputs specific rules that consist of labelled gene pairs along with their associated microarray sample IDs.

One of the most important aspects of this study is its ability to deal with large scale and multiple heterogeneous data sets. By use of dynamic thresholding on expression profiles, we also alleviate the disadvantages of crude thresholding on expression data. Available biclustering algorithms are known to require space complexity on large amount of data. For this purpose, we also modified an existing MFI algorithm for biclustering.

In order to test our framework, we applied it on all available NCBI GEO *Homo sapiens* data sets and more specifically five different functionally concise groups of NCBI GEO data sets independently (i.e., Breast Cancer, Normal Human

Tissue, Obesity, Liver and Colon). Gene-pair rules and their association with a given sample set exhibited concordance with the literature. Furthermore, our results provided novel insights into the co-regulated gene pairs among a compendium of tissues as well as diverse conditions of human cancers.

References

Agrawal, R., Imieliński, T. and Swami, A. (1993) 'Mining association rules between sets of items in large databases', *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington DC, USA, pp.207–216.

Agarwal, R.C., Aggarwal, C.C. and Prasad, V.V.V. (2000) 'Depth first generation of long patterns', *KDD '00: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA, pp.108–118.

Agarwal, R.C., Aggarwal, C.C. and Prasad, V.V.V. (2001) 'A tree projection algorithm for generation of frequent item sets', *J. Parallel Distrib. Comput.*, Vol. 61, No. 3, pp.350–371.

Balasubramaniyan, R., Hüllermeier, E., Weskamp, N. and Kämper, J. (2005) 'Clustering of gene expression data using a local shape-based similarity measure', *Bioinformatics*, Vol. 21, No. 10, pp.1069–1077.

Barkow, S., Bleuler, S., Prelić, A., Zimmermann, P. and Zitzler, E. (2006) 'BicAT: a biclustering analysis toolbox', *Bioinformatics*, Vol. 22, No. 10, pp.1282–1283.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007) 'NCBI GEO: mining tens of millions of expression profiles–database and tools update', *Nucleic Acids Research*, Vol. 35, pp.760–765.

Barry, W.T., Nobel, A.B. and Wright, F.A. (2005) 'Significance analysis of functional categories in gene expression studies: a structured permutation approach', *Bioinformatics*, Vol. 21, No. 9, pp.1943–1949.

Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.F. and Gandrillon, O. (2002) 'Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data', *Genome Biology*, Vol. 3, No. 12, pp.research0067.1–0067.16.

Ben-Dor, A., Chor, B., Karp, R. and Yakhini, Z. (2002) 'Discovering local structure in gene expression data: the order-preserving submatrix problem', *RECOMB '02: Proceedings of the Sixth Annual International Conference on Computational Biology*, Washington DC, USA, pp.49–57.

Berrar, D., Dubitzky, W., Granzow, M. and Eils, R. (2001) 'Analysis of gene expression and drug activity data by knowledge-based association mining', *Proceedings of Critical Assessment of Microarray Data Analysis Techniques (CAMDA'01)*, Durham, NC, USA, pp.25–28.

Burdick, D., Calimlim, M. and Gehrke, J. (2001) 'MAFIA: a maximal frequent itemset algorithm for transactional databases', *ICDE'01: Proceedings of the 17th International Conference on Data Engineering*, Heidelberg, Germany, p.443.

Cariello, N.F., Romach, E.H., Colton, H.M., Ni, H., Yoon, L., Falls, J.G., Casey, W., Creech, D., Anderson, S.P., Benavides, G.R., Hoivik, D.J., Brown, R. and Miller, R.T. (2005) 'Gene expression profiling of the PPAR-alpha agonist ciprofibrate in the cynomolgus monkey liver', *Toxicol Science*, Vol. 88, No. 1, pp.250–264.

Carkacioglu, L., Can, T., Konu, O., Atalay, V. and Cetin-Atalay, R. (2006) 'Expression pattern analysis of housekeeping genes across large number of microarray experiments', *5th European Conference on Computational Biology (ECCB)*, Eilat, Israel, pp.P9 (poster 9).

Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J.M. and Pascual-Montano, A. (2006) 'Integrated analysis of gene expression by association rules discovery', *BMC Bioinformatics*, Vol. 7, No. 1, p.54.

Cheng, Y. and Church, G.M. (2000) 'Biclustering of expression data', *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, La Jolla, CA, USA, Vol. 8, pp.93–103.

Creighton, C. and Hanash, S. (2003) 'Mining gene expression databases for association rules', *Bioinformatics*, Vol. 19, No. 1, pp.79–86.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) 'Cluster analysis and display of genome-wide expression patterns', *Proc. Natl. Acad. Sci. USA*, Vol. 95, No. 25, pp.14863–14868.

Eisenberg, E. and Levanon, E.Y. (2007) 'Human housekeeping genes are compact', *IEEE Computer Society Press*, Vol. 4, No. 2, pp.167–189.

Fridlyand, J., Snijders, A., Ylstra, B., Li, H., Olshen, A., Segraves, R., Dairkee, S., Tokuyasu, T., Ljung, B., Jain, A., McLennan, J., Ziegler, J., Chin, K., Devries, S., Feiler, H., Gray, J., Waldman, F., Pinkel, D. and Albertson, D. (2006) 'Breast tumor copy number aberration phenotypes and genomic instability', *BMC Cancer*, Vol. 6, No. 1, p.96.

Georgii, E., Richter, L., Ruckert, U. and Kramer, S. (2005) 'Analyzing microarray data using quantitative association rules', *Bioinformatics*, Vol. 21, Suppl. 2, pp.ii123–ii129.

Gouda, K. and Zaki, M.J. (2001) 'Efficiently mining maximal frequent itemsets', *ICDM*, pp.163–170.

Gyenesei, A., Wagner, U., Barkow-Oesterreicher, S., Stolte, E. and Schlapbach, R. (2007) 'Mining co-regulated gene profiles for the detection of functional associations in gene expression data', *Bioinformatics*, Vol. 23, No. 15, pp.1927–1935.

Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R.V., Misra, J., Dillon, W., Lee, K.F., Clark, K.E., Haverty, P., Weng, Z., Mutter, G.L., Frosch, M.P., Macdonald, M.E., Milford, E.L., Crum, C.P., Bueno, R., Pratt, R.E., Mahadevappa, M., Warrington, J.A., Stephanopoulos, G. and Gullans, S.R. (2001) 'A compendium of gene expression in normal human tissues', *Physiol Genomics*, Vol. 7, No. 2, pp.97–104.

Jiang, D., Tang, C. and Zhang, A. (2004) 'Cluster analysis for gene expression data: a survey', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 11, pp.1370–1386.

Kotala, P., Zhou, P., Mudivarthy, S., Perrizo, W. and Deckard, E. (2001) 'Gene expression profiling of DNA microarray data using peano count trees', *Proceedings of the First Virtual Conference on Genomics and Bioinformatics*, Fargo, NC, USA, pp.15–16.

Kyungpil, K., Shibo, Z., Keni, J., Li, C., In-Beum, L., Lewis, F.J. and Haiyan, H. (2007) 'Measuring similarities between gene expression profiles through new data transformations', *BMC Bioinformatics*, Vol. 8, p.29.

Lacroix, M. (2006) 'Significance, detection and markers of disseminated breast cancer cells', *Endocr. Relat. Cancer*, Vol. 13, No. 4, pp.1033–1067.

Lee, Y.H., Nair, S., Rousseau, E., Tataranni, P.A., Bogardus, C. and Permana, P.A. (2005) 'Microarray profiling of isolated abdominal subcutaneous adipocytes from obese vs. non-obese Pima Indians: increased expression of inflammation-related genes', *Diabetologia*, Vol. 48, No. 9, pp.1776–1783.

Lin, M., Park, J., Nishidate, T., Nakamura, Y. and Katagiri, T. (2007) 'Involvement of maternal embryonic leucine zipper kinase (MELK) in mammary carcinogenesis through interaction with Bcl-G, a pro-apoptotic member of the Bcl-2 family', *Breast Cancer Res.*, Vol. 9, No. 1, p.R17.

Liu, J., Yang, J. and Wang, W. (2004) 'Biclustering in gene expression data by tendency', *CSB '04: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, Stanford, CA, USA, pp.182–193.

Madureira, P.A., Varshochi, R., Constantinidou, D., Francis, R.E., Coombes, R.C., Yao, K. and Lam, W. (2006) 'The forkhead box M1 protein regulates the transcription of the estrogen receptor alpha in breast cancer cells', *J. Bio. Chem.*, Vol. 281, No. 35, pp.25167–25176.

Pei, J., Zhang, X., Cho, M., Wang, H. and Yu, P.S. (2003) 'MaPle: a fast algorithm for maximal pattern-based clustering', *ICDM 2003. Third IEEE International Conference on Data Mining*, Melbourne, FL, USA, pp.259–266.

Richardson, A.L., Wang, Z.C., De Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J.D., Livingston, D.M. and Ganesan, S. (2006) 'X chromosomal abnormalities in basal-like human breast cancer', *Cancer Cell*, Vol. 9, No. 2, pp.121–132.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M.J., Bergh, J., Piccart, M. and Delorenzi, M. (2006) 'Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis', *J. Natl. Cancer Inst.*, Vol. 98, No. 4, pp.262–272.

Tuzhilin, A. and Adomavicius, G. (2002) 'Handling very large numbers of association rules in the analysis of microarray data', *Proceedings of the Eighth ACM SIGKDD International Conference on Data Mining and Knowledge Discovery*, Edmonton, Alberta, Canada, pp.396–404.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A. and Speleman, F. (2002) 'Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes', *Genome Biology*, Vol. 3, No. 7, pp.research0034.1–0034.11.

Wang, H., Wang, W., Yang, J. and Yu, P.S. (2002) 'Clustering by pattern similarity in large data sets', *SIGMOD Conference*, Madison, WI, USA, pp.394–405.

Warrington, J.A., Nair, A., Mahadevappa, M. and Tsyganskaya, M. (2000) 'Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes', *Physiol Genomics*, Vol. 2, No. 3, pp.143–147.

Yang, J., Wang, H., Wang, W. and Yu, P. (2003) 'Enhanced biclustering on expression data', *BIBE '03: Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering*, Bethesda, MD, USA, pp.321–327.

Yang, Y., Dudoit, S., Lin, S., Peng, V., Nga, J. and Speed, T. (2002) 'Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation', *Nucleic Acids Research*, Vol. 30, No. 4, p.15.

Zhao, L. and Zaki, M.J. (2005) 'MicroCluster: efficient deterministic biclustering of microarray data', *IEEE Intelligent Systems*, Vol. 20, No. 6, pp.40–49.