BIOINFORMATICS

Vol. 25 no. 21 2009, pages 2795–2801

Bi-correlation clustering algorithm for determining a set of co-regulated genes

Anindya Bhattacharya[1] and Rajat K. De[2,*]

[1]Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata 700152 and [2]Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, West Bengal, India

ABSTRACT

Motivation: Biclustering has been emerged as a powerful tool for identification of a group of co-expressed genes under a subset of experimental conditions (measurements) present in a gene expression dataset. Several biclustering algorithms have been proposed till date. In this article, we address some of the important shortcomings of these existing biclustering algorithms and propose a new correlation-based biclustering algorithm called bi-correlation clustering algorithm (BCCA).

Results: BCCA has been able to produce a diverse set of biclusters of co-regulated genes over a subset of samples where all the genes in a bicluster have a similar change of expression pattern over the subset of samples. Moreover, the genes in a bicluster have common transcription factor binding sites in the corresponding promoter sequences. The presence of common transcription factors binding sites, in the corresponding promoter sequences, is an evidence that a group of genes in a bicluster are co-regulated. Biclusters determined by BCCA also show highly enriched functional categories. Using different gene expression datasets, we demonstrate strength and superiority of BCCA over some existing biclustering algorithms.

Availability: The software for BCCA has been developed using C and Visual Basic languages, and can be executed on the Microsoft Windows platforms. The software may be downloaded as a zip file from http://www.isical.ac.in/~rajat. Then it needs to be installed. Two word files (included in the zip file) need to be consulted before installation and execution of the software.

1 INTRODUCTION

Cluster analysis on gene expression data is a popular tool for identification of groups of co-expressed genes under all experimental conditions (measurements) present on the input dataset. Many clustering algorithms have been proposed in this regard. However, common disadvantage of all these clustering algorithms is that they try to find group of genes that remain co-expressed through all experimental conditions (measurements). But in reality genes tends to be co-regulated and thus co-expressed under only a few experimental conditions. They may start behaving differently under different conditions. If an input dataset has many measurements and an algorithm tries to find out group of genes expressed similarly under all measurements, then chances of finding such a group with success is very less. To overcome this problem, the concept of biclustering has emerged.

Biclustering is a technique that performs simultaneous grouping on genes and conditions (measurements) of a dataset to determine subgroups of genes that exhibit similar behavior over a subset of experimental conditions. The technique introduced by Hartigan (1972) was first applied on gene expression data by Cheng and Church (2000). Several biclustering algorithms have been proposed till date. They include, among others, Block Clustering by Hartigan (1972), δ-biclusters by Cheng and Church (2000), Coupled Two-Way Clustering (CTWC) by Getz *et al.* (2000), FLOC by Yang *et al.* (2002, 2003), δ-pClusters by Wang *et al.* (2002), Spectral biclustering by Kluger *et al.* (2008), Iterative Signature Algorithm (ISA) algorithm by Ihmels *et al.* (2002, 2004), Interrelated Two-Way Clustering (ITWC) algorithm by Tang *et al.* (2001), Plaid model by Lazzeroni and Owen (2002), Order Preserving Sub Matrix (OPSM) algorithm by Ben-Dor *et al.* (2002), SAMBA by Tanay *et al.* (2002) and xMOTIF by Murali and Kasif (2003). Prelic *et al.* (2006) have compared performance of different biclustering algorithms, and proposed a fast divide-and-conquer biclustering algorithm (Bimax). Teng and Chan (2006, 2008) have developed a biclustering algorithm based on weighted correlation coefficient, which involves sorting of gene expression data matrix both by rows and columns, followed by their comparison. They have used weight values in weighted correlation coefficient to avoid finding the already identified biclusters.

Apart from different approaches of biclustering, Pascual-Montano *et al.* (2006) have applied the notion of non-negative matrix factorization (NMF; Kim and Tidor, 2003) to the analysis of gene-array experiments and designed a software tool called bioNMF. It is capable of recognizing similarity between sub-portions of the data corresponding to localized features in expression space and is able to produce biclusters as subsets of genes behaving similarly in a subset of expressions. Another competitive tool with different biclustering techniques regarding the analysis of gene expression data is mining attribute profile (MAP; Gyenesei *et al.*, 2007). The algorithm can be characterized as a depth-first search, divide-and-conquer algorithm. Application of MAP to gene expression data allows for identification of genes whose expressions follow the same pattern in response to different biological conditions.

Common shortcoming of all these biclustering algorithms including bioNMF is that they may be able to find a group of genes that show similar expression pattern over a group of samples, but none of them can determine a group of co-regulated genes having common transcription factors among all the genes in a bicluster as a support toward their co-regulation. A recent biclustering algorithm, called cMonkey (Reiss *et al.*, 2006), groups genes and conditions into biclusters on the basis of (i) coherence in expression data across subsets of experimental conditions; (ii) co-occurrence of putative *cis*-acting regulatory motifs in the regulatory regions of bicluster members; and (iii) the presence of highly connected subgraphs in metabolic and functional association networks. For complex types of input data, cMonkey only supports gene expression data biclustering for a small number of species, namely, *Halobacterium* NRC-1, *Helicobacter pylori*, *Saccharomyces cerevisiae* and *Escherichia coli*.

Here, we propose a new biclustering algorithm, called bi-correlation clustering algorithm (BCCA), based on a correlation coefficient. The algorithm is able to produce biclusters of co-regulated genes over a subset of samples where all the genes in a bicluster not only obtain a similar change of expression pattern over a subset of sample, but also can have common transcription factors.

In our study, the superior capability of clustering by BCCA over a number of other algorithms, namely, δ-biclusters (Cheng and Church, 2000), OPSM (Ben-Dor *et al.*, 2002), SAMBA (Tanay *et al.*, 2002), ISA (Ihmels *et al.*, 2002, 2004), Bimax (Prelic *et al.*, 2006), bioNMF (Pascual-Montano *et al.*, 2006) and the method of Teng and Chan (2006, 2008) is demonstrated through experiments with five gene expression datasets that are publicly available. These datasets are Yeast CC dataset (YCCD; Cheng and Church, 2000), Spellman *et al*. dataset (SPTD; Spellman *et al.*, 1998), GDS958 (Wills-Karp and Ewart, 2004), GDS2547 (Chandran *et al.*, 2007; Yu *et al.*, 2004) and GDS2938 (Wang *et al.*, 2007). Various issues related to the characteristics of the algorithm are also discussed.

2 BI-CORRELATION CLUSTERING ALGORITHM

Let us consider a set of n genes $X=\{g_1,g_2,\ldots,g_n\}$, for each of which m expression values are given. That is, for each gene g_i there is an m-dimensional vector $\mathbf{x}_i$, where x_{ij} is the j-th expression value of g_i. Let us also consider a set of m microarray experiments (measurements) $Y=\{e_1,e_2,\ldots,e_m\}$. For each experiment, we have n expression values corresponding to n genes in X. These n genes will have to be grouped into K overlapping biclusters $\{C_1,C_2,\ldots,C_K\}$. In our algorithm, we define a bicluster based on a correlation coefficient as similarity measure.

Bicluster: generally a bicluster C_k can be defined as a subset of genes I_k possesing a similar behavior over a subset of experiments (measurements) J_k. Thus, a bicluster C_k can be represented as $C_k=(I_k,J_k)$. A bicluster $C_k=(I_k,J_k)$ contains a subset I_k ($I_k\subseteq X$) of genes and a subset J_k ($J_k\subseteq Y$) of experiments where each gene in I_k is correlated with a correlation value greater than or equal to a specified threshold (θ), with all other genes in I_k over the measurements in J_k.

BCCA uses Pearson correlation coefficient for measuring similarity between expression patterns of two genes g_i and g_j, and is defined as

$$\mathrm{Corr}(\mathbf{x}_i,\mathbf{x}_j)=\frac{\sum_{l=1}^{m}(x_{il}-\bar{x}_i)(x_{jl}-\bar{x}_j)}{\sqrt{\sum_{l=1}^{m}(x_{il}-\bar{x}_i)^2\sum_{l=1}^{m}(x_{jl}-\bar{x}_j)^2}} \quad (1)$$

where x_{il} and x_{jl} are l-th expression values of the i-th and j-th genes, respectively. The terms $\bar{x}_i$ and $\bar{x}_j$ are mean values over m expression values (corresponding to microarray experiments) of the i-th and j-th genes, respectively. $\mathrm{Corr}(\mathbf{x}_i,\mathbf{x}_j)>\theta$ indicates that g_i and g_j are similarly expressed, i.e. their expression values are changing in a similar way. Staring with a pair of genes, BCCA augments a bicluster by including a new gene based on the correlation values with all the other genes in the bicluster. Thus, the algorithm minimizes the chance of misplacement of genes in a bicluster.

In Step 1 of BCCA, the set of biclusters S is initialized to *NULL* and number of biclusters *Bicount* is initialized to 0, as a bicluster is yet to be determined. BCCA generates a bicluster (C) for each pair of genes in a dataset as any pair of genes may be a pair of co-regulated genes if they have similar change in expression values under a set of conditions (measurements). For each pair of genes, g_i, g_j ($i\neq j$), BCCA creates a bicluster $C=(I,J)$, in Step 2A, where $I=\{g_i,g_j\}$ and $J=Y$. For a pair of genes in C, if $\mathrm{Corr}(\mathbf{x}_i,\mathbf{x}_j)<\theta$ then a sample is detected (in Step 2C) from C, deletion of which causes maximum increase in correlation value between g_i and g_j. If $m'=|J|>r$, $r\geq 3$ being a threshold, the sample is deleted from J. Otherwise, C is discarded. The sample is deleted because we want only a subset of conditions (measurements) for which all genes present in a bicluster are highly positively correlated with all other genes in that bicluster. Deletion of a measurement for which genes differ in expression value the most will result in the highest increase in correlation value. Note that here we have considered $r\geq 3$ as otherwise, correlation coefficient will be $+1$ or -1, or cannot be computed. In this way BCCA deletes one measurement at a time from J, which causes the highest increase in correlation value, and continues to delete measurements until a pair of genes in C become correlated with correlation value equal to or higher than θ for the remaining measurements in J, and the number of measurements is greater than or equal to r. Thus, C contains a pair of genes and a set of measurements for which that pair of genes are highly positively correlated.

In Step 2D(a), other genes from $X-I$, which satisfy the definition of a bicluster are included in C for its augmentation. The algorithm checks in Step 2D(b), whether the present bicluster C has already been found. If it is so then we do not need to include C, otherwise, C is considered as a new bicluster.

Algorithm Input: (i) A set $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ of expression vectors corresponding to genes in X, for each of which m expression values $Y=\{e_1,e_2,\ldots,e_m\}$ are given. (ii) A correlation threshold value θ. (iii) The threshold value r for minimum number of expression values.

Output: A finite set of overlapping biclusters $S=\{C_1,C_2,\ldots,C_O\}$, where each bicluster $C_i=\{I_i,J_i\}$, $I_i\subseteq X$ and $J_i\subseteq Y$.

Steps:

1. Initially, set $S=NULL$ and set number of biclusters $Bicount=0$.
2. For each pair of genes g_i, g_j, $i\neq j$, do:
   - A. Set $C=(I,J)$, $I=\{g_i,g_j\}$ and $J=Y$.
   - B. Assign number of expression values in C to a variable m', i.e. set $m'=|J|$.
   - C. While $\mathrm{Corr}(\mathbf{x_k},\mathbf{x_i})<\theta$, $g_i,g_j\in I$ and $m'\geq r$, do:
     - a. From m' expression values, find out the measurement e_l elimination of which expression from J will cause maximum increase in $\mathrm{Corr}(\mathbf{x}_i,\mathbf{x}_j)$ [Equation (1)].
     - b. Set $J=J-\{e_l\}$.
     - c. Set $m'=m'-1$.
   - D. If $\mathrm{Corr}(\mathbf{x}_i,\mathbf{x}_j)\geq\theta$, for $g_i,g_j\in I$ over m' expression values in J, where $m'\geq r$, then
     - a. Set $X'=X-I$;
     - b. For each $g_p\in X'$, do:
       - i. If $\mathrm{Corr}(\mathbf{x}_j,\mathbf{x}_p)\geq\theta$, for all $\mathbf{x}_j\in I$ over m' expression values in J, then set $I=I\bigcup\{g_p\}$.
       - ii. Set $X'=X'-\{g_p\}$;
     - c. If there exists no $C_k\in S$ such that $I_k=I$ and $J_k=J$, then
       - i. Set $Bicount=Bicount+1$.
       - ii. Set $I_{Bicount}=I$, $J_{Bicount}=J$ and $C_{Bicount}=(I_{Bicount},J_{Bicount})$.
       - iii. Set $S=S\bigcup\{C_{Bicount}\}$.

3 RESULTS

Here, we demonstrate the affectiveness of BCCA in determining a set of co-regulated genes (i.e. the genes having common transcription factors) and functionally enriched clusters (and attributes) on five datasets. The superior performance of BCCA over some other biclustering algorithms, namely, δ-biclusters (Cheng and Church, 2000), OPSM (Ben-Dor *et al.*, 2002), SAMBA (Tanay *et al.*, 2002), ISA (Ihmels *et al.*, 2002, 2004), Bimax (Prelic *et al.*, 2006), bioNMF (Pascual-Montano *et al.*, 2006) and the method of Teng and Chan (2006, 2008) have also been depicted in terms of the aforesaid criteria. These gene expression datasets deal with two yeasts (http://sgdlite.princeton.edu/download/yeast_datasets/) and three mammals (http://www.ncbi.nlm.nih.gov/geo/). Moreover, a discussion on various characteristics of the algorithm is provided.

3.1 Datasets

A short description of five gene expression datasets used in analysis is given in Table 1. Null rows/columns and rows/columns with all zeros are deleted from the datasets before applying biclustering algorithms. For example, five such rows are deleted from original YCCD.

3.2 Variation with respect to threshold

Correlation threshold can be any value between -1 and $+1$ but most likely to be positive, and depends on data. Thus, for determination of correlation threshold θ, one can vary correlation threshold between 0 and 1, and then for each biclustering result, the average number of functionally enriched attributes is determined. From a plot of average number of functionally enriched attributes (computed using P-values) versus correlation threshold value, the correlation threshold value associated with the highest average number of functionally enriched attributes can be selected. Supplementary Figure 9 shows such a plot for YCCD dataset. In Supplementary Figure 9, for correlation threshold value of 0.91, maximum average number of functionally enriched attributes is found. Assumption of 0.91 as the correlation threshold value is due to the following facts. If the threshold value is <0.91 (for YCCD data) then some functionally dissimilar genes are included in a bicluster. On the other hand, if the threshold value is >0.91 (for YCCD dataset) then some functionally similar genes cannot be included in the same bicluster. However, the exact choice of the threshold value ~0.9 does not make a big difference.

The selection of optimum correlation threshold value by varying correlation threshold and judging each biclustering result takes huge amount of time. For this reason, we have followed a guideline on this value from a previous study by Allocco *et al.* (2004), which has concluded that if two genes have a correlation between their expression profiles >0.84 then there is >50% chance of being bounded by a common transcription factor. For YCCD dataset, Supplementary Figure 9 shows that a very high (almost equal to the largest number) average number of functionally enriched attributes for biclustering results of BCCA is obtained for correlation threshold value of 0.85.

Table 1. Short description of the datasets used in analysis

Name (Organism)	Number of genes	Number of samples
YCCD (Yeast)	2879	17
SPTD (Yeast)	6178	77
GDS958 (Mouse)	22690	12
GDS2547 (*Homo sapiens*)	12646	164
GDS2938 (*H.sapiens*)	22283	12

3.3 Performance comparison

For the purpose of comparison, we have, at first, considered only those biclusters that have less than or equal to 50 genes. The reason behind this is that finding functional enrichment in larger group of genes is much more easier than that in a smaller group of genes. Genes from such biclusters (containing at most 50 genes) that have been obtained by BCCA on five datasets are listed in Supplementary Tables 3–12. Detailed analysis with these small biclusters is presented in Sections 3.3.1 and 3.3.2.

Bicluster size could influence the comparison study as it would be much easier to identify common transcription factor binding sites in small gene sets than in larger ones. Thus, 100 non-overlapping biclusters of varying sizes generated by BCCA have been selected for comparisons with all the biclusters obtained by other biclustering algorithms. Results of the analysis with these biclusters are presented in Section 3.3.3.

Expression profile plots of biclusters obtained by different algorithms are also examined. Such a plot corresponding to BCCA (Supplementary Fig. 10) for $Bicluster_1$ of GDS958 dataset shows that the expression values of all the genes change in a similar way over different conditions (measurements). Such plots for biclusters generated by δ-bicluster (Cheng and Church, 2000) (Supplementary Fig. 11), SAMBA (Supplementary Fig. 12), ISA (Supplmentary Fig. 13), Bimax (Supplementary Fig. 14) and OPSM (Supplementary Fig. 15) do not show such similar variation in the expression profiles.

For the purpose of comparison with other biclustering algorithms, we have started with the parameter values (of these algorithms) recommended/used in the original papers. Then we have adjusted these values to maximize the average number of functionally enriched attributes per bicluster. Supplementary Table 2 shows this fact.

3.3.1 By locating common transcription factors We have used a software TOUCAN 2 (described in the Supplementary Material; Aerts *et al.*, 2005) for performance comparison by extracting information on the number of transcription factors present in proximal promoters of all the genes in a single bicluster. A transcription factor that is found in promoter region of all the genes in a bicluster, is considered as a common transcription factor for that bicluster. Presence of common transcription factors in the promoter regions of a set of genes is a good evidence toward co-regulation. The results obtained by different biclustering algorithms including BCCA have been compared in terms of the number of common transcription factors that can bind to the promoter regions of all the genes in a bicluster. Higher the number of common transcription factors, better is the chance of finding a set of co-regulated genes in a bicluster.

As mentioned above, we have considered the biclusters containing less than or equal to 50 genes. We have got many such biclusters, for

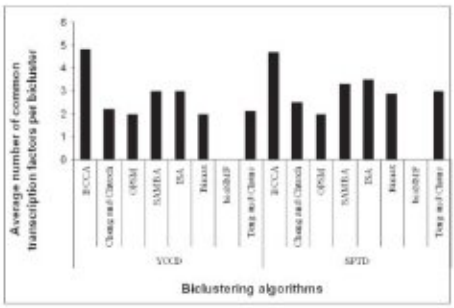

Fig. 1. Comparison of transcription factors on small biclusters for Yeast datasets.

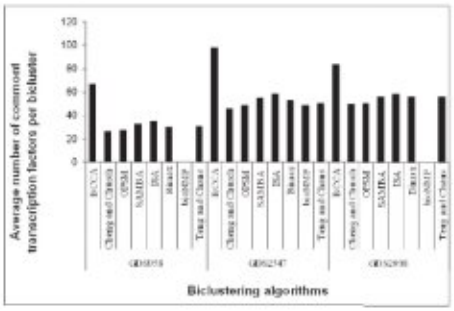

Fig. 2. Comparison of transcription factors on small biclusters for Mammalian datasets.

several datasets, obtained by BCCA and few by the other algorithms considered here. For BCCA, we have picked up 10 most distinct biclusters for each of these datasets. However, we have considered all the biclusters for the other algorithms. Average number of common transcription factors per bicluster obtained by BCCA and other aforesaid algorithms on the five datasets are shown in Figures 1 and 2.

Figures 1 and 2 show that average number of common transcription factors per bicluster obtained by BCCA for all the datasets is significantly greater than that of the other biclustering algorithms. Supplementary Table 13 shows the list of common transcription factors of all the selected biclusters obtained by BCCA for two yeast datasets (YCCD and SPTD). Similarly, Supplementary Tables 14–45 show the lists of common transcription factors of all biclusters obtained by BCCA for GDS958, GDS2547 and GDS2938 datasets. Presence of quite a large number of common transcription factors in all biclusters obtained by BCCA as shown in Figures 1 and 2 and Supplementary Tables 13–45 suggest that they contain co-regulated genes.

Supplementary Figure 16 shows that sequences of all the five genes have been found present in a bicluster generated by BCCA from SPTD dataset. Name of the transcription factors found by MotifLocator (available in TOUCAN 2) in any of the five gene sequences are listed in the legend. Positions of each transcription factor binding site for each transcription factor in five gene sequences are also marked. Supplemantary Figure 16 shows that any transcription factor may be found present in more than one location in upstream region of a gene. For example, transcription factor GCN4 has been found 16 times in upstream region of the gene YAL003W (Supplementary Fig. 16). The same transcription factor has been found present for all the genes in a bicluster. For example, GCN4 has been found in upstream region of the genes, namely, YAL003W, YDR418W, YBR118W, YBR189W and YBR191W (Supplementary Fig. 16).

3.3.2 Functional enrichment: analysis and comparisons using P-values The enriched functional categories for each bicluster obtained by BCCA on five datasets are listed in Supplementary Tables 46–53. The functional enrichment of each GO category in each of the biclusters has been calculated by its *P*-value. To compute the *P*-value, we have employed the software Funcassociate (Berriz *et al.*, 2003). *P*-value represents the probability of observing the number of genes from a specific GO functional category within each cluster. A low *P*-value indicates that the genes belonging to the enriched functional categories are biologically significant in the corresponding clusters. In the present article, only functional categories with $P < 5.0 \times 10^{-7}$ are reported in order to restrict the size of the article.

Analysis: of the 10 biclusters obtained for the YCCD (Supplementary Tables 46 and 47), the highly enriched category in bicluster $Bicluster_1$ is the 'ribosome' with *P*-value of 4.2×10^{-17}. The biclusters $Bicluster_1$ to $Bicluster_{10}$ contain several enriched categories including 'ribosome', 'cytosolic ribosome', 'DNA replication', 'replication fork', 'structural molecule activity' and 'bud neck' as shown in Supplementary Tables 46 and 47.

For the SPTD (Supplementary Table 48), the highly enriched category in bicluster $Bicluster_2$ is the 'cytosolic ribosome (sensu Eukaryota)/80S ribosome' with *P*-value of 2.2×10^{-30}. Different GO categories 'nucleosome', 'ribosome', 'ribosome biogenesis' and 'rRNA processing' are enriched in $Bicluster_1$ to $Bicluster_{10}$.

As in the above datasets, for GDS958 dataset (Supplementary Tables 49 and 50), the biclusters $Bicluster_1$ to $Bicluster_{10}$ contain several enriched categories on 'ribosome'. The highly enriched category in bicluster $Bicluster_9$ is the 'ribosome' with *P*-value of 6.6×10^{-22}. The biclusters $Bicluster_1$ to $Bicluster_{10}$ contain several enriched categories including 'hydrogen ion transporter activity/proton transporter', 'striated muscle thin filament', 'lytic vacuole', 'ATP synthesis coupled proton transport' as shown in Supplementary Tables 49 and 50.

For GDS2547 dataset (Supplementary Table 51), the highly enriched category in bicluster $Bicluster_1$ is the 'thiamin diphosphokinase activity' with *P*-value of 9.0×10^{-10}. Similarly, for GDS2938 dataset (Supplementary Tables 52 and 53), the highly enriched categories in bicluster $Bicluster_2$ is the 'immune response' with *P*-value of 6.24×10^{-126}. Several other GO categories are enriched in $Bicluster_1$ to $Bicluster_{10}$ of both GDS2547 and GDS2938 datasets.

From the results of Supplementary Tables 46–53, we see that all the 10 distinct biclusters obtained by BCCA for each of the above datasets are functionally enriched. Ten biclusters obtained for the YCCD dataset are enriched with 83 functionally enriched categories, whereas 10 distinct biclusters obtained for the SPTD dataset are found to be enriched with 68 functionally enriched categories. Similarly, 10 distinct biclusters obtained for GDS958, GDS2547 and GDS2938 datasets are enriched with 102, 50 and 113 functionally enriched categories, respectively.

Comparisons: here we describe the ability of detecting functionally enriched biclusters (categories) by the aforesaid

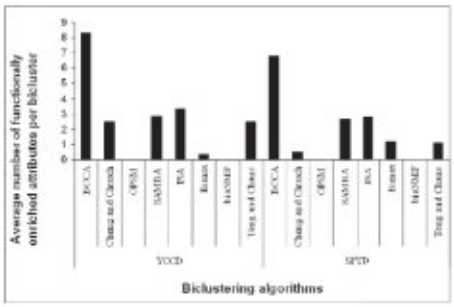

Fig. 3. Comparison of functionally enriched attributes on small biclusters for Yeast datasets.

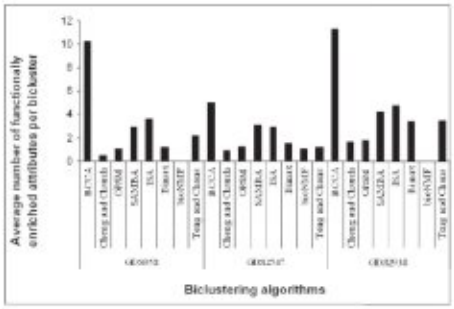

Fig. 4. Comparison of functionally enriched attributes on small biclusters for Mammalian datasets.

biclustering algorithms. Figures 3 and 4 show that BCCA provides much higher average number of functionally enriched categories per bicluster than the other aforesaid algorithms for all the datasets.

3.3.3 Comparing biclusters of varying sizes The software TOUCAN 2 used for transcription factor analysis has been found to be very time consuming and is not scaled well to large datasets. Hence, we have considered PRIMA available in EXPANDER (Tanay *et al.*, 2002) for analysis of transcription factor binding sites corresponding to the biclusters of varying sizes. Number of enriched transcription factors for each bicluster of yeast datasets is found based on *P*-values. Figure 5 shows that, for YCCD and SPTD datasets, BCCA has resulted in the highest average number of significant transcription factors per bicluster compared with that obtained by the other algorithms. Higher the average number of significant transcription factors per bicluster better is the algorithm. In the present article, only transcription factors with $P < 1.0 \times 10^{-4}$ are reported as significance. Regarding functional enrichment, BCCA has again resulted in the highest average number of enriched attributes per bicluster compared with that obtained by the other algorithms (Fig. 6).

3.4 Finding relationship among genes mediating allergic asthma

The dataset GDS958, generated by Wills-Karp and Ewart (2004) from lung tissue of mouse, has been used to show the relationship among different genes mediating allergic asthma. From some earlier investigations on asthma mediation (Cormier *et al.*, 2002; Grunig *et al.*, 1998; Wills-Karp, 2004), we know that IL-13 (InterLeukin 13)

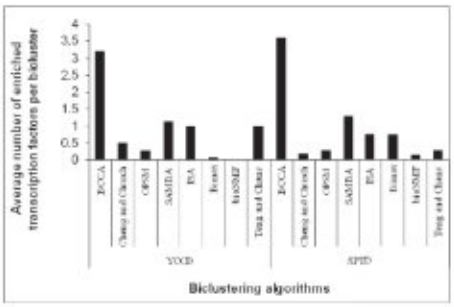

Fig. 5. Comparison of transcription factors on diverse set of biclusters for Yeast datasets.

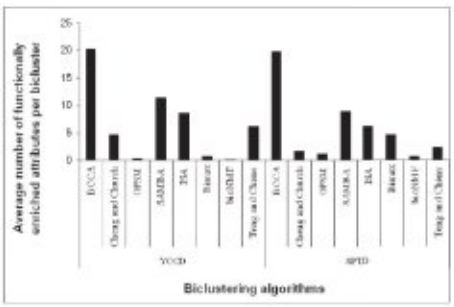

Fig. 6. Comparison of functionally enriched attributes on diverse set of biclusters for Yeast datasets.

is the central mediator of asthma. List of genes known to be associated with asthma mediation is given in The GEArray S Series Mouse Autoimmune and Inflammatory Response Gene Array (http://www.eurogentec.com/). It includes different adaptor proteins, cell surface receptors, cytokines and receptors, chemokines and receptors, signal transduction proteins and some other related genes.

For GDS958 dataset, the biclusters ($Bicluster(IL\text{-}13)_1$, $Bicluster(IL\text{-}13)_2$, $Bicluster(IL\text{-}13)_3$ and $Bicluster(IL\text{-}13)_4$) obtained by BCCA in Supplementary Table 54 contain central mediator of asthma gene IL-13. Moreover, these biclusters also contain different adaptor proteins (Traf3), cell surface receptors (Fcer1a, Cd68 and Cd24a), cytokines and receptors (Il4i1, Il13ra2, Tnfrsf11a and Il7r), chemokines and receptors (Ccl6, Cxcl16, Cxcl12 and Ccr5), signal transduction proteins (Serpine2, Map2k4) and some other related genes (Col1a2) that are also listed in the GEArray as responsible for asthma. These observations suggest that IL-13 is co-regulated with different adaptor proteins, cell surface receptors, cytokines and receptors, chemokines and receptors, signal transduction proteins and some other related genes listed in the GEArray.

Moreover, Chi3l3, Serpine2, Serpina3n, Arg1 and IgK-V1 present in the aforesaid biclusters along with IL-13 may have role to play in asthma mediation. Role of these genes in asthma mediation is also suggested by earlier investigations on Chi3l3 (Welch *et al.*, 2002), Serpine2 (http://geneticassociationdb.nih.gov), Serpina3n (http://geneticasso-ciationdb.nih.gov), Arg1 (Vercelli, 2003; Zimmermann *et al.*, 2003) and IgK-V1

(http://geneticassociationdb.nih.gov). We also mention here that IL-13 may be co-regulated with these asthma mediating genes, namely, Chi3l3, Serpine2, Serpina3n, Arg1 and IgK-V1.

3.5 Discussions

Here, we provide some important characteristics of BCCA based on the results obtained for YCCD dataset. In order to restrict the size of the article, we have not included the results using other datasets.

3.5.1 Effect on variation of bicluster size Size of the biclusters obtained by BCCA depends on correlation threshold values. Average number of genes and measurements (samples) per bicluster decreases with increase in the correlation threshold value, which is depicted in Supplementary Figures 17 and 18 for YCCD dataset. However, this decrease in size of biclusters does not always lead to decrease in functional enrichment of attributes in biclusters. Supplementary Figure 9 also shows this fact, where average number of functional enrichment per bicluster increases as correlation threshold value increases.

3.5.2 Diverse set of biclusters BCCA is able to generate a diverse set of overlapping biclusters. Supplementary Figures 7 and 8 show for YCCD dataset that each gene or sample is a member of more than one bicluster. There are genes in YCCD dataset that are included in more than 500 biclusters and some of the samples in the same dataset are included in more than 1000 biclusters. The total number of biclusters for YCCD dataset is about 2455. All these data implies that the chances of getting overlapping and diverse set of biclusters is high as the algorithm does not result in the same biclusters multiple times. Similar findings have also been obtained for the other datasets.

3.5.3 Comparisons with Spearman's rank correlation coefficient and Euclidean distance BCCA is a general algorithm and any pairwise correlation measure can be used as a similarity measure instead of Pearson correlation coefficient. We have compared the results using Spearman's rank correlation coefficient and Euclidean distance. For example, the numbers of functionally enriched attributes per bicluster, for YCCD dataset, with Spearman's rank correlation and Euclidean distance have been found to be 20.4 and 18.5, respectively, while that using Pearson's correlation is 24.9. Moreover, the numbers of significant transcription factor per bicluster, for YCCD dataset, with Spearman's rank correlation and Euclidean distance have been found to be 2.6 and 2.2, respectively, while that using Pearson's correlation is 3.2. Thus, Pearson's correlation coefficient results in the best for YCCD dataset.

3.5.4 Main difference of BCCA over the other algorithms BCCA generates biclusters following two major principles that were overlooked by some other biclustering algorithms. First, BCCA tries to obtain one bicluster for each pair of genes in a dataset as any pair of genes can be important for further analysis. In order to obtain one bicluster for a pair of genes, BCCA fixes a set of conditions (measurements) for which the pair of genes is correlated. Then the algorithm tries to augment the biclusters by including other genes on the given set of measurements as any change in this set may cause removal of the starting pair of genes. Second, BCCA augments a bicluster by including a new gene based on correlation values with all the other genes in the bicluster. It helps in minimizing the chance of misplacement in a BCCA bicluster. However, the main difference of BCCA with the other biclustering algorithms lies in the use of correlation coefficient for forming biclusters.

3.5.5 Time complexity Upper bound on the execution time of BCCA for a single iteration is $O(n^3)$, for a dataset of n genes and m samples. Since the number of iterations is $O(n^2)$, time complexity of BCCA is $O(n^5)$. (One may refer to the Supplementary Material for its analysis.) The high time complexity of BCCA is due to its three major steps, i.e. Step 2C—selection of samples for a pair of genes; Step 2D(a)—augmentation of a bicluster; and Step 2D(b)—comparison of a bicluster with all the others for similarity. BCCA has taken about 1 h 40 min to generate all the (2455) biclusters for YCCD dataset in a server with 2 GHz Quad core processor and 2 GB RAM. For the algorithm of Cheng and Church, upper bound on the execution time for a single iteration is $O(nm)$ (Cheng and Church, 2000); while that for OPSM is $O(nm^3l)$, where l is the number of biclusters (Ben-Dor *et al.*, 2002). SAMBA has the time complexity of $O(n2^d)$, where d is the upper bound on the degree of each vertex (Tanay *et al.*, 2002). The algorithm of Teng and Chan (2006, 2008) need time $O(m^2)$ for executing a single iterations. The running time complexity of Bimax is $O(nm\beta \times \min\{n,m\})$, where β is a parameter (Prelic *et al.*, 2006). Although nominally ISA runtime scales linearly with number of genes and samples (Bergmann *et al.*, 2003), it scales linearly with the number of seeds, which, to get a good modularization, is required to be larger for a bigger set of data.

4 CONCLUSIONS

Here, we have developed a biclustering algorithm called BCCA based on Pearson correlation coefficient as a similarity measure. The algorithm is able to find a group of genes that show similar pattern of variation in their expression profiles over a subset of measurements (microarray experiments). Interestingly, the genes in a bicluster obtained by BCCA have many common transcription factors. Thus, we may say that the genes in such a bicluster are co-regulated, i.e. they have some common transcription factors. It has also been found that BCCA has been able to find higher number of common transcription factors of a set of genes in a bicluster than that of some other biclustering algorithms. Moreover, BCCA has found a diverse set of biclusters that are more functionally enriched than that of some other biclustering algorithms. Regarding the limitation of BCCA, the dataset must contain expression profile of various genes for at least three experiments (measurements). Some of the steps [Steps 2C, 2D(a) and 2D(b)] of BCCA are time consuming.

Conflict of Interest: none declared.

REFERENCES

Aerts,S. *et al.* (2005) Toucan 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acid Res.*, **33**, 393–396.

Allocco,D.J. *et al.* (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, **5**, 18.

Ben-Dor,A. *et al.* (2002) Discovering local structure in gene expression data: the order-preserving submatrix problem. In *Proceedings of the Sixth International Conference on Computational Biology (RECOMB 02)*, Washington DC. ACM, New York, pp. 49–57.

Bergmann,S. *et al.* (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E*, **67**, 031902.

Berriz,F.G. *et al.* (2003) Characterizing gene sets with funcassociate. *Bioinformatics*, **19**, 2502–2504.

Chandran,U.R. *et al.* (2007) Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer*, **7**, 64.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93–103.

Cormier,S.A. *et al.* (2002) Th2-mediated pulmonary inflammation leads to the differential expression of ribonuclease genes by alveolar macrophages. *Am. J. Respir. Cell Mol. Biol.*, **27**, 679–687.

Getz,G. *et al.* (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12084.

Grunig,G. *et al.* (1998) Requirement for il-13 independently of il-4 in experimental asthma. *Science*, **282**, 2261–2263.

Gyenesei,A. *et al.* (2007) Mining co-regulated gene profiles for the detection of functional associations in gene expression data. *Bioinformatics*, **23**, 1927–1935.

Hartigan,J.A. (1972) Direct clustering of a data matrix. *J. Am. Stat. Assoc.*, **67**, 123–129.

Ihmels,J. *et al.* (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.

Ihmels,J. *et al.* (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics*, **20**, 1993–2003.

Kim,P.M. and Tidor,B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.*, **13**, 1706–1718.

Kluger,Y. *et al.* (2008) Spectral biclustering of microarray cancer data: Co-clustering genes and conditions. *Genome Res.*, **13**, 703–716.

Lazzeroni,L. and Owen,A. (2002) Plaid models for gene expression data. *Statistica Sinica*, **12**, 61–86.

Murali,T.M. and Kasif,S. (2003) Extracting conserved gene expression motifs from gene expression data. *Pacific Symposium on Biocomputing*, pp. 77–88.

Pascual-Montano,A. *et al.* (2006) bioNMF: a versatile tool for non-negative matrix factorization in biology. *BMC Bioinformatics*, **7**, 366.

Prelic,A. *et al.* (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Reiss,D.J. *et al.* (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, **2**, 280–302.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell.*, **9**, 3273–3297.

Tanay,A. *et al.* (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18**, S136–S144.

Tang,C. *et al.* (2001) Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. In *Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering*, Bethesda. IEEE Computer Society, pp. 41–48.

Teng,L. and Chan,L.W. (2006) Biclustering gene expression profiles by alternately sorting with weighted correlated coefficient. In *Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, Arlington. IEEE, pp. 289–294.

Teng,L. and Chan,L.W. (2008) Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data. *J. Signal Process. Syst.*, **50**, 267–280.

Vercelli,D. (2003) Arginase: marker, effector, or candidate gene for asthma? *J. Clin. Invest.*, **3**, 1815–1817.

Wang,H. *et al.* (2002) Clustering by pattern similarity in large data sets. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, Madison. ACM, New York, pp. 394–405.

Wang,S. *et al.* (2007) Microarray analysis of cytokine activation of apoptosis pathways in the thyroid. *Endocrinology*, **10**, 4844–4852.

Welch,J.S. *et al.* (2002) Th2 cytokines and allergic challenge induce ym1 expression in macrophages by a stat6-dependent mechanism. *J. Biol. Chem.*, **277**, 42821–42829.

Wills-Karp,M. (2004) Interleukin-13 in asthma pathogenesis. *Curr. Allergy Asthma Rep.*, **4**, 123–131.

Wills-Karp,M. and Ewart,S.L. (2004) Time to draw breath: asthma-susceptibility genes are identified. *Nat. Rev. Genet.*, **5**, 376–387.

Yang,J. *et al.* (2002) δ-clusters: capturing subspace correlation in a large data set. In *Proceedings of the 18th IEEE International Conference on Data Engineering*, San Jose. IEEE Computer Society, pp. 517–528.

Yang,J. *et al.* (2003) Enhanced biclustering on expression data. In *Proceedings of the 3rd IEEE Conference on Bioinformatics and Bioengineering*, Bethesda. IEEE Computer Society, pp. 321–327.

Yu,Y. *et al.* (2004) Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.*, **22**, 2790–2799.

Zimmermann,N. *et al.* (2003) Dissection of experimental asthma with dna microarray analysis identifies arginase in asthma pathogenesis. *J. Clin. Invest.*, **3**, 1863–1874.