Pattern Recognition 43 (2010) 2692–2697

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

MIB: Using mutual information for biclustering gene expression data ☆

Neelima Gupta, Seema Aggarwal *

Department of Computer Science, University of Delhi, Delhi 110 007, India

ARTICLE INFO

Article history:
Received 12 February 2009
Received in revised form
27 December 2009
Accepted 7 March 2010

Keywords:
Biclustering
Gene expression data
Mutual information
GO term and transcription factor binding site

ABSTRACT

Result of any biclustering or clustering algorithm depends on the choice of the similarity measure. Most of the biclustering algorithms are based on Euclidean distance or correlation coefficient. These measures capture only linear relationships between the genes but nonlinear dependencies may exist amongst them. In this paper we propose an approach using mutual information for biclustering gene expression data. Mutual information is a more general measure to investigate relationships (positive, negative correlation and nonlinear relationships as well). To the best of our knowledge, none of the existing algorithms for biclustering have used mutual information as a similarity measure between two genes. We obtained biclusters from the gene expression data of *Arabidopsis thaliana* and compared our biclusters with those obtained by two other algorithms namely ISA and BIMAX. Biological significance of the biclusters was checked using GO database. It was found that the genes belonging to our biclusters were significantly enriched with GO terms with better p values as compared to the genes of the biclusters obtained by the other two algorithms. To further investigate the biclusters, we studied the promoter regions of the genes belonging to a bicluster for common patterns/transcription factor binding sites (TFBS) or motifs. Promoter regions of the genes of most of our biclusters were found to have a common motif patterns which existed in the motif database of *Arabidopsis thaliana*. Also, the motifs extracted from our biclusters had better E values than those of others. Thus reconfirming that use of mutual information as a similarity measure will produce better biclusters.

1. Introduction

With the help of microarray experiments biologists are able to study the expression of thousands of genes under a large number of conditions simultaneously. The large scale of the data makes it challenging to analyse it to extract any biologically significant information from it. Standard clustering algorithms like k-means clustering work well for small data sets but fair poorly when the number of experimental conditions is large as they cluster the genes based on their expression under all the conditions, whereas the cellular processes are generally affected by a small subset of conditions. Most of the other conditions which do not contribute to the cellular process add to the background noise. Moreover, these algorithms compute non-overlapping clusters, i.e. a gene belongs to at most one cluster, whereas in fact a gene may be responsible for several cellular activities and hence may be included in more than one cluster.

In [1] Cheng and Church introduced the notion of biclustering in which the biclusters are defined to be a set of genes and a set of conditions under which these genes are most tightly regulated. In other words, a bicluster is a subset of genes that are related to each other under a subset of conditions. Biclusters could overlap both on genes as well as conditions.

Result of any biclustering algorithm depends on the choice of the similarity measure used. Different similarity measures on the same expression data produce different results. The existing algorithms [1–8] for biclustering use some kind of similarity measure like Euclidean distance or correlation coefficient. Though these measures have been successfully and satisfactorily used for several years they capture only the linear relationships between the objects. In particular, a vanishing correlation coefficient implies absence of only linear dependencies. However, in many applications, like gene expression data and word document data nonlinear relationships may exist between the objects. Moreover with advances in experimental technology, increasing methodologies are available for unveiling more complex relationships. Hence we need similarity measures which exploit nonlinear dependencies. In the context of gene expression data, biologists are interested in studying how an increase or decrease in the expression of a gene affects the expression pattern of other genes. They are interested in identifying a group of genes showing similar patterns in their expression. If the increase or decrease in the expression value of a gene is linear viz a viz the increase or decrease in the expression value of another gene then similarity

☆This project is supported by University of Delhi.
* Corresponding author.
E-mail addresses: ngupta@cs.du.ac.in (N. Gupta), saggarwal@mh.du.ac.in (S. Aggarwal).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.03.002

measures like Euclidean distance and correlation coefficient will be able to extract this relationship. So far studies have focused only on this aspect. However, if changes in the expression pattern of genes are not related linearly rather are related by say a quadratic or exponential function, then these measures would fail and one would need measures which exploit nonlinear dependencies.

In this paper we propose an approach using mutual information for biclustering [10] gene expression data. To the best of our knowledge, none of the existing algorithms for biclustering gene expression data have used mutual information as a similarity measure between two genes. The only other algorithm that uses mutual information for co-clustering by Dhillon et al. [11] is discussed in Section 2.

We applied our algorithm on gene expression data of *Arabidopsis thaliana* and successfully extracted the biclusters. We validated our biclusters using external biological information by determining the functionality of the genes of the biclusters from the gene ontology database [13] using the AMIGO tool [14]. Genes belonging to our biclusters were found to be significantly enriched with GO terms with very small p values. We also used the web tool FuncAssociate [15] to compute the adjusted p values. All our biclusters were found to be statistically significant with adjusted p values <0.001. To further biologically validate our biclusters we searched for common patterns (motifs) from the promoter regions of the genes belonging to a bicluster. Promoter regions of the genes of most of the biclusters were found to have statistically significant common motif patterns which existed in the motif database of *Arabidopsis thaliana*.

We also compared our algorithm with two other biclustering algorithms namely ISA [8] and BIMAX [9]. The genes of our biclusters were associated with GO annotations with smaller p values than the genes of the biclusters of ISA and BIMAX. Also the motifs extracted from our biclusters were found to be statistically more significant (smaller E values) as compared to the motifs extracted from the biclusters of ISA and BIMAX. The overlap of our biclusters with those of ISA and BIMAX was found to be $<30\%$. Thus showing that our biclusters were better and different from those of ISA and BIMAX.

The rest of the paper is arranged as follows. In the next section we briefly describe the related work done in this area followed by the theoretical concepts used in our algorithm in Section 3. In Section 4 we describe our mutual information based biclustering algorithm. Finally, in Section 5 we discuss the results obtained by running our algorithm on expression data of *Arabidopsis thaliana*.

2. Related work

In [16] Steur et al. have shown that mutual information can be used as a measure of similarity to cluster data. They show that mutual information provides a better and more general criterion to investigate relationships (positive, negative correlation and nonlinear dependencies) between variables by showing that higher correlation coefficient implies higher mutual information but two variables having very low values of correlation coefficient (implying no linear relationship) may still be related to each other (nonlinear dependencies).

Many researchers [17–21] have used mutual information for one way clustering (clustering of genes on the entire set of conditions). These algorithms also support that information theoretic measures are responsive to any type of dependencies, including strongly nonlinear structures as compared to traditional measures which search for linear relationships only. In [21] Priness et al. have compared the use of mutual information with respect to both Euclidean distance and correlation coefficient as a similarity measure for one way clustering. With some procedural modifications they incorporated mutual information measure in some clustering algorithms like k-means, self organized maps, click and sIB [22–24]. They show that the mutual information is a more generalized measure of statistical dependence and is resistant to outliers and missing data. They also show that mutual information based methods give clusters having better homogeneity and separation scores.

In [17] Butte and Kohane compute pairwise mutual information for all genes against each other. Their work is also based on the hypothesis that an association between two genes indicated by a high amount of mutual information between them would also signify biological relationship.

In [11] Dhillon et al. have used mutual information for co-clustering word document data. An element e_{ij} of the input matrix represents the frequency of occurrence of i th word in the j th document. Dhillon et al. treat the word document matrix as a co-occurrence matrix and use it to represent the joint probability distribution of the words (represented by random variable X) and the documents (represented by random variable Y). They approximate the original matrix with a new matrix consisting of a reduced set of rows $\hat{X}$ and a reduced set of columns $\hat{Y}$, so that the new matrix contains as much information about the earlier one as possible. Thus their approach typically leads to dimensionality reduction. However, it is different from traditional dimensionality reduction in the sense that they do it simultaneously on rows as well as on columns. Though the paper beautifully exploits the information contained in the columns viz a viz rows and the vice versa, it has its limitations especially with reference to gene expression data. Firstly the entries in the gene expression data cannot be treated as a measure of co-occurrence. Secondly, to treat the input matrix as a joint probability distribution the entries must be all positive which may not be the case in gene expression data as down-regulation may be represented by negative values. Banerjee et al. in [12] propose a generalized co-clustering algorithm which works for negative entries in the input matrix as well. They assume that the probability distribution of the input data is either predefined or follows uniform distribution. Both Banerjee and Dhillon identify non-overlapping biclusters, whereas a gene may be responsible for more than one cellular function and thus may belong to more than one bicluster. Similarly biclusters may overlap on conditions as well.

3. The mutual information

The mutual information between two random variables X and Y is a measure of information contained in X about Y or the information contained in Y about X. If given a value of X, it is easy to predict the value of Y then X contains good amount of information about Y. Clearly with this definition, if X and Y are independent the mutual information between them is zero and it is high if they are highly dependent or closely related to each other. Thus Kullback has defined mutual information between two random variables as a measure of divergence from the hypothesis that X and Y are independent.

3.1. The kullback divergence

Consider a system A with N_A possible states. An experiment performed on A puts the system in one of the states $a_1, a_2 \ldots a_{N_A}$, each with its corresponding probability $p(a_i)$. The information gained by the system through a series of experiments is the amount of surprise one feels on reading the outcomes of the experiments. Thus if one hypothesize that probability distribution observed by the outcomes is $\{p^0\}$ and the actual densities are $\{p\}$,

the *Kullback* divergence $K(p/p^0)$ between the two probability distributions is given by

$$K(p/p^0) = \sum_i p_i \log \frac{p_i}{p_i^0}$$

Kullback divergence can be interpreted as the information gained when the assumed probability distribution $\{p^0\}$ is replaced by the final distribution $\{p\}$. $K(p/p^0)$ is always greater than or equal to zero [25]. It equals zero if and only if $\{p^0\}$ and $\{p\}$ are same. In our case the assumed probability distribution $\{p^0\}$ is given by the hypothesis that two variables X and Y are statistically independent. Thus $p^0_{XY}(x_i, y_j)$ is given by

$$p^0_{XY}(x_i,y_j) = p_X(x_i)p_Y(y_j)$$

The final distribution $\{p\}$ is given by the observed joint probability densities $p_{XY}(x_i,y_j)$. Thus using Kullback divergence mutual information is defined as

$$I(X,Y) = \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p_{XY}(x_i,y_j)\log\frac{p_{XY}(x_i,y_j)}{p_X(x_i)p_Y(y_j)}$$

where X takes values $x_1,x_2 \ldots x_{n_x}$, Y takes values $y_1,y_2 \ldots y_{n_y}$, $p_{XY}(x_i,y_j)$ represents the joint probability distribution of X, Y and $p_X(x_i)$, $p_Y(y_j)$ are the marginal distributions of X and Y, respectively. The mutual information is zero if and only if X and Y are statistically independent, i.e. vanishing mutual information does imply that the two variables are independent. This shows that mutual information provides a more general measure of dependencies in contrast to the commonly used measures like Euclidean distance and correlation coefficient which quantify only the linear relationships.

3.2. Estimating probability densities

We can compute the mutual information between two variables if we have explicit knowledge of the joint probability distribution and the marginal probability distributions. In general these probabilities are not known. Various methods are used to estimate the probability densities from the observed data. Consider a series (x_i,y_i) of n simultaneous observations of two continuous random variables X and Y. Since entropy (*Kullback* divergence here) is computed using discrete probabilities, we estimate probability densities using the widely used [17,18] histogram method. In histogram method, bins of width h are defined for each variable whose probability densities are to be estimated. Given an origin o (which could be different for different variables resulting in different bins), the bins are defined by the intervals $[o+rh, o+(r+1)h]$, $r=1 \ldots N_x$. Let $f_X(i)$ denote the number of observations of X falling in the bin a_i. The probabilities $\{p(a_i)\}$ are then estimated as

$$p(a_i) = \frac{f_X(i)}{n}$$

Let the bins of the random variable Y are denoted by b_j, $j=1 \ldots N_y$. Let $f_Y(j)$ denote the number of observations of Y falling in the bin b_j. The probabilities $\{p(b_j)\}$ are then estimated as

$$p(b_j) = \frac{f_Y(j)}{n}$$

Let $f_{XY}(i,j)$ denote the number of observations such that X falls in bin a_i and Y falls in bin b_j. Then the mutual information between X and Y is estimated as

$$I(X,Y) = \log n + \frac{1}{n}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} f_{XY}(i,j)\log\frac{f_{XY}(i,j)}{f_X(i)f_Y(j)}$$

4. MIB: mutual information based biclustering algorithm

In this section, we present our algorithm to find biclusters using mutual information. Let G be the set of genes and C be the set of conditions in the expression data. Let the number of genes in G be denoted by N_g and the number of conditions in C be denoted by N_c. We find biclusters which are pairs of (G', C'), where G' is the subset of genes which are most closely related to the gene seed (g^*) under the subset C' of conditions. More biclusters are obtained by running MIB for different gene seeds.

One way of choosing a gene seed is to start with a random seed. Then the set of genes most related to this seed is found. The next choice of the seed can be a gene which has not been clustered so far and so on. In this way one can develop a method to determine a set of well separated gene seeds. We are currently working on developing this method. Experimental results in this paper were obtained by choosing the gene seeds randomly.

Our algorithm takes a gene as a seed and proceeds in three steps. In the first step we find the set of genes which are most related to the input gene seed. For this we compute the pairwise mutual information of the gene seed with all other genes over all the conditions. Genes having mutual information greater than a gene threshold (t_g) are selected.

In the second step, the algorithm identifies the experimental conditions under which the set of genes found in the first step show maximum relatedness. For this the algorithm computes the pair wise mutual information of a condition c^* with all other condition over the reduced set of genes. Again only those conditions are selected whose pairwise mutual information is greater than a condition threshold (t_c).

In the third and the final step which is a bicluster refinement step the algorithm selects from the whole expression data those genes which are most related to the gene seed under the subset of conditions identified in step two. Genes not related to the gene seed under all the conditions but related under a subset of conditions will be identified in this step.

Since we do not know which c^* is best, second and third steps are repeated for every condition.

If a set of related genes is known beforehand we can skip the first step of finding the gene subset and go to step 2 directly. If no set of related genes is known, we enter into a chicken-egg problem wherein the question arises as to whether one should start grouping the genes first or conditions first. Feature selection algorithms start with grouping the conditions first, whereas most of the clustering algorithms start grouping the genes first as that is the most intuitive thing to do. The two approaches have different objectives and each involves its own tradeoffs.

The procedure MIB summarizes our algorithm. Procedure $mi(x_1,x_2,n)$ computes the pair wise mutual information between the two variables x_1 and x_2 of length n according to the formula given in Section 3.2.

Procedure: MIB
Main algorithm
Input: g^*, G, C, t_g, t_c, N_g, N_c
Method Begin:
1. G_0=compute-G0 (g^*, G, C, t_g, N_g, N_c)
2. for all $c^* \in C$
   (a) C' = compute − conditions (c^*, G_0, C, t_c, N_{g0}, N_c).
   (b) G=compute-bicluster-genes (g^*, G, C', t_g, N_g, $N_{c'}$)
   (c) output (G',C').

Method End.

Procedure: Compute-G0
Computes the initial gene set (G_0).
Input: g_*, G, C, t_g, N_g, N_c

Method Begin:

1. let g^* be the seed gene.
2. For $i=1$ to N_g
   Compute mi (g_i, g^*, N_c).
3. $\mu = \Sigma_i mi(g_i, g^*, N_c)/N_g$
4. $\sigma^2 = \Sigma_i (mi(g_i, g^*, N_c) - \mu)^2 / N_g$
5. $G_0 = \{g_i : \frac{(mi(g_i, g^*, N_c) - \mu)}{\sigma} > t_g\}$.
6. return G_0.

Method End.

Procedure: Compute-conditions
Computes the relevant set of conditions (C′).
Input: $(c^*, G_0, C, t_c, N_{g\,0}, N_c)$.
Method Begin:

1. For $j=1$ to N_c
   Compute mi $(c_j, c^*, N_{g\,0})$.
2. $\mu = \Sigma_j mi(c_j, c^*, N_{g0})/N_c$
3. $\sigma^2 = \Sigma_j (mi(c_j, c^*, N_{g0}) - \mu)^2 / (N_c)$
4. $C' = \{c_j : \frac{(mi(c_j, c^*, N_{g0}) - \mu)}{\sigma} > t_c\}$.
5. return C'.

Method End.

Procedure: Compute-bicluster-genes
Computes the final genes (G′) of the bicluster.
Input: $(g^*, G, C', t_g, N_g, N_{c'})$
Method Begin:

1. For $i=1$ to N_g
   Compute mi $(g_i, g^*, N_{c'})$.
2. $\mu = \Sigma_i mi(g_i, g^*, N_{c'})/N_g$
3. $\sigma^2 = \Sigma_i (mi(g_i, g^*, N_{c'}) - \mu)^2 / (N_g)$
4. $G' = \{g_i : \frac{(mi(g_i, g^*, N_{c'}) - \mu)}{\sigma} > t_g\}$.
5. Return G'.

Method End.

5. Results

5.1. Obtaining biclusters

We downloaded gene expression data for *Arabidopsis thaliana* from [26]. The dataset contained expression profiles of 734 genes under 69 conditions. We implemented our algorithm MIB in C++. For the Thaliana dataset we randomly chose 50 genes as gene seeds. For each gene seed we ran MIB by choosing all the 69 conditions as the reference condition. Thus we discovered 69 biclusters for each gene seed. The score of a bicluster was defined as the average mutual information of all the genes (with the gene seed) belonging to the bicluster over all the conditions of the bicluster. Out of the 69 biclusters obtained for each gene seed we selected the one with the maximum score. Thus we got one bicluster for each gene seed. Barkow et al. in [26] had implemented both ISA [8] and BIMAX [9] and had obtained 100 and 72 biclusters, respectively, on *Arabidopsis thaliana* gene expression data. We downloaded these biclusters for comparison with ours.

5.2. Validation of biclusters

A major task for all clustering algorithms is to evaluate the quality and reliability of the clusters obtained. Biclusters have different set of conditions and may also overlap both on genes and conditions therefore it is not clear how to extend internal measures (measures which rely only on the input data) like Rand index or Jacard index (used to define the homogeneity and the separation of the clusters) to the concept of biclustering. To the best of our knowledge no general internal index has been developed for biclustering solutions. Prelic et al. in [9] endorse this fact. External measures use additional data to validate the obtained biclustering results. According to Prelic et al. [9] and Gat-Viks et al. [27] biological merit is the main criterion for validation of biclusters. Thus additional data like already known biological knowledge has been used for assessing the quality of biclusters. Also, biological validation allows us to draw conclusions about the biological usefulness of the biclusters. Most of the biclustering algorithms use external measures like GO annotation term [7], metabolic pathways [8], protein–protein interaction network [9], patterns in promoter regions [28], etc. to assess the quality of the biclusters. Reliability of the biclusters, i.e. the likelihood that the bicluster is not found by chance can be measured by calculating statistical measures like p values. We used different methods to find the biological significance of the obtained biclusters. The methods used were based on the fact that a group of related genes are responsible for some biological activity. The first method involved the use of GO database to check whether the obtained biclusters were functionally enriched with GO terms. The other method was to find common patterns in the promoter regions of the genes belonging to a bicluster.

5.2.1. Functional enrichment term

We investigated whether the biclusters obtained show significant enrichment with respect to a specific gene ontology annotation using the AMIGO tool [14]. The term enrichment tool of AMIGO finds significant shared GO terms of each GO category (biological process, cellular component and molecular function) used to describe the genes in the input set. It calculates the p value for each GO term within the genes of a bicluster. p value is the probability of seeing x number of genes from the input list of n genes annotated to a particular GO term, given the proportion of genes in the whole genome annotated to that GO term. That is, the GO terms shared by the genes in the user's list are compared to the background distribution of annotation. The closer the p value is to zero, the more significant is the association of the particular GO term with the group of genes (i.e. it is less likely that the observed annotation of the particular GO term to a group of genes occurs by chance).

There may be several GO terms with different p values associated with an input set of genes belonging to a bicluster. For each bicluster we record the best p value of the GO term for each GO category (biological process, cellular component and molecular function). Table 1 summarizes the best and the worst p value over all biclusters for the three algorithms. From the table we see that these values are better for the biclusters of MIB than those of ISA and BIMAX. Thus we can say that our biclusters are more biologically significant than those of ISA and BIMAX.

We also used the web tool Funcassociate [15] to evaluate the biclusters. Funcassocite computes the adjusted significant score of the genes in the bicluster in two steps. First it uses Fisher's Exact test to compute the hyper geometric functional score of the gene

Table 1
P values of GO terms associated with the genes of the biclusters.

Method	Biological process		Cellular component		Molecular function	
p value	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
MIB	$3.00e^{-95}$	$2.01e^{-19}$	$4.91e^{-24}$	$1.46e^{-06}$	$3.77e^{-202}$	$2.51e^{-33}$
ISA	$4.43e^{-51}$	$1.05e^{-18}$	$2.07e^{-13}$	$1.88e^{-03}$	$1.01e^{-69}$	$5.14e^{-33}$
BIMAX	$6.10e^{-23}$	$1.26e^{-05}$	$9.49e^{-16}$	$3.09e^{-03}$	$2.68e^{-33}$	$3.37e^{-11}$

set. It finds the probability of finding atleast m genes with attribute A in a query list of length q if the null hypothesis is true, where the null hypotheses is that the query list is independent of the attribute. Next Funcassociate uses the Westfall and Young procedure [29] to compute the adjusted significant score of the gene set. It estimates the adjusted p value from the results of 1000 simulated null hypothesis. All our biclusters were found to be statistically significant with adjusted p values < 0.001.

5.2.2. Detection of common motifs

Transcription factors are proteins that are responsible for gene regulation. They bind to a sequence site in the promoter region of the genes called the transcription factor binding site (TFBS) or a motif. When a transcription factor binds to its TFBS the transcription of a gene is initiated. Thus the genes showing dependencies in expression data are expected to have common patterns in their promoter regions. We studied the promoter regions of the genes belonging to a bicluster for such common patterns.

To simplify the study, smaller biclusters with more than 50% genes in common with some larger bicluster were filtered out for all the three algorithms. We were left with 50 out of 100, 22 out of 72 and 15 out of 50 distinct biclusters of BIMAX, ISA and MIB, respectively. For each of these remaining biclusters, we retrieved the promoter sequences of its genes using regulatory sequence analysis (RSA) tools [30]. We further used the tool MEME [31] (multiple expectation maximization for motif elicitation) to find the common motifs in their promoter regions. The width of the motif was set in the range from 5 to 10 and the number of motifs was set to 10.

For each bicluster we got 10 consensus sequence patterns of the motifs. Along with each motif, MEME calculates an E value of the motif which is the measure of the statistical significance of the motif. Lower the E value more significant is the motif. E value is an estimate of the expected number of motifs with the given (or higher) log likelihood ratio with the same width and number of occurrences, that one would find in a similarly sized set of random sequences. The minimum (best) E value of the motifs of the MIB biclusters ($5.9e^{-010}$) was better than the minimum E value of the motifs extracted from ISA ($1.8e^{+001}$) and BIMAX ($6.1e^{+003}$), once again endorsing the fact that the biclusters of MIB are biologically more significant than the ones obtained by ISA and BIMAX.

We further checked the existence of these motifs in the motif database of *Arabidopsis thaliana* using PLACE [32]. At least one of these motifs from every bicluster (except for one) belonged to the motif database of *Arabidopsis thaliana* thereby endorsing the good quality of the biclusters. Some of the other motifs were found in other organisms but not in *Arabidopsis thaliana*. Patterns like *CGGCGACGAG* that did not match in the *Arabidopsis thaliana* motif database but were found in other organisms like rice (*Oryza sativa*) and *Chlamydomonas reinhardtii* can be targets for further research by biologists.

5.3. Overlap with other biclusters

We also calculated the overlap between the gene set of two biclusters A and B having N_A and N_B number of genes, respectively, as

$$O_{A,B} = \frac{N_{A\cap B}}{(N_A + N_B)/2} * 100 \quad (1)$$

where $N_{A\cap B}$ is the number of genes belonging to both the biclusters A and B. We found that the overlap between biclusters of MIB and those of ISA and BIMAX is $< 30\%$. Thus we claim that biclusters of MIB using MI as a similarity measure are different from those of ISA and BIMAX.

5.4. Effect of noise

We used synthetic dataset to study the performance of our algorithm in the presence of noise. We implanted ten 10×10 biclusters in a 100×100 matrix and superimposed this matrix on another 100×100 matrix of noise generated by drawing random numbers from a normal distribution. The minimum value in the implanted bicluster was 0.65. Noise level was varied from 0.001 to 0.003. We were able to extract all the 10 implanted biclusters from the synthetic data. We also extracted the rows and column of a bicluster detected by MIB from the expression data of *A. thaliana*. To add noise to this bicluster of size 138×4, it was perturbed by adding random numbers generated from normal distribution stored in a matrix of size 150×50. We were able to extract the implanted bicluster from the data using MIB.

6. Conclusion and future work

In our algorithm we have given a new approach to bicluster high dimensional gene expression data using mutual information. As the mutual information captures more general relationships as compared to traditional similarity measures like Euclidean distance and correlation coefficient our algorithm is able to discover different and biologically significant biclusters.

In the present work mutual information is computed with randomly chosen gene seeds. We are working on a method to find the gene seeds intelligently.

Acknowledgments

We wish to thank Ishan Qureshi and Surubhi Bajaj for their help in biological validation of the biclusters. We also wish to thank Chintalapati Janaki from CDAC for her useful suggestions and timely help.

References

[1] Y. Cheng, G.M. Church, Biclustering of gene expression data, system molecular biology, System Molecular Biology 8 (2000) 1–93.
[2] J. Ihmels, G. Friedlander, S. Bergmann, Y. Ziv, O. Sarig, N. Barkai, Revealing modular organization in the yeast transcription network, Nature Genetics 31 (2002) 1–370.
[3] H. Wang, W. Wang, J. Yang, P.S. Yu, Clustering by pattern similarity in large data sets, Bulletin of Mathematical Biology 46 (1984) 515–527.
[4] G. Getz, E. Levine, E. Domany, Coupled two-way clustering analysis of gene microarray data, PNAS 97 (2000) 12079–12084.
[5] Y. Kluger, R. Basri, J.T. Chang, M. Gerstein, Spectral biclustering of microarray data: coclustering genes and conditions, Genome Research 13 (2003) 703–716.
[6] M. Kloster, C. Tang, N.S. Wingreen, Finding regulatory modules through large-scale gene-expression data analysis, Bioinformatics 21 (2005) 1172–1179.
[7] X. Liu, L. Wang, Computing the maximum similarity bi-clusters of gene expression data, Bioinformatics 23 (2007) 50–56.
[8] S. Bergmann, J. Ihmels, N. Barkai, Iterative signature algorithm for the analysis of large-scale gene expression data, Physical Review 67 (2003) 1–18.
[9] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, E. Zitzler, A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics 22 (2006) 1122–1129.
[10] N. Gupta, S. Aggarwal, MIB: using mutual information for biclustering high dimensional data, in: IADIS European Conference on Data Mining, 2008.
[11] I.S. Dhillon, S. Mallela, D.S. Modha, Information theoretic CO-clustering, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, August 24–27, 2003.
[12] A. Banerjeee, I. Dhillion, D.S. Modha, J. Ghosh, S. Merugu, A generalized maximum entropy approach to bergman co-clustering and matrix approximation, Journal of Machine Learning Research 8 (2007) 1919–1986.

[13] M. Ashburner, C.A. Ball, J.A. Blake, D. Bolstein, H. Butler, M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The gene ontology consortium, Nature Genetics 25 (2000) 25–29.
[14] S. Carbon, A. Ireland, C.J. Mungall, S. Shu, B. Marshall, S. Lewis, AMIGO Hub, Web Presence Working Group, AmiGO: online access to ontology and annotation data, Bioinformatics 25 (2) (2009) 288–289.
[15] G.F. Berriz, Characterizing gene sets with funcassociate, Bioinformatics 19 (2003) 2502–2504 ⟨http://llama.med.harvard.edu/cgi/func1/funcassociate/⟩.
[16] R. Steuer, J. Kurths, C.O. Daub, J. Weiseand, J. Selbig, The mutual information: detecting and evaluating dependencies between variables, Bioinformatics 18 (Suppl 2) (2002) S231–S240.
[17] A.J. Butte, I.S. Kohane, Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements, PSB 5 (2000) 415–426.
[18] G.S. Michaels, D.B. Carr, M. Askenazi, S. Fuhrman, X. Wen, R. Somogyi, Cluster analysis and data visualization of large scale gene expression data, PSB 3 (1998) 42–53.
[19] X. Zhou, X. Wang, E.R. Dougherty, D. Russ, E. Suh, Gene clustering based on clusterwide mutual information, Journal of Computational Biology 11 (1) (2004) 147–161.
[20] N. Slonim, G.S. Atwal, G. Tkacik, W. Bialek, Information based clustering, PNAS 102 (2005) 18297–18302.
[21] I. Priness, O. Maimon, I. Ben-Gal, Evaluation of gene expression clustering via mutual information distance measure, BMC Bioinformatics 8 (2007) 111.
[22] T. Kohonen, Self Organizing Maps, Springer, Berlin, 1997.
[23] R. Shamir, R. Sharan, Algorithmic approaches to clustering gene expression data, in: J. Tao, X. Ying, Q.Z. Michael (Eds.), Current Topics in Computational Biology, MIT Press, Cambridge, USA, 2002.
[24] N. Slonim, The information bottleneck: theory and applications, Ph.D. Thesis, Tel-Aviv University, Computer Science Department, 2002.
[25] S. Haykin, Neural Networks—A comprehensive Foundation, second ed., Prentice Hall of India Ltd, New Delhi, 2007.
[26] S. Barkow, S. Bleuler, A. Prelic, P. Zimmermann, E. Zitzler, BicAT, a biclustering analysis toolbox manual, Bioinformatics 22 (2006) 1282–1283 ⟨http://www.tik.ee.ethz.ch/sop/bicat⟩.
[27] I. Gat-Viks, R. Sharan, R. Shamir, Scoring clustering solutions by their biological relevance, Bioinformatics 19 (2003) 2381–2389.
[28] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, G.M. Church, Systematic determination of genetic network architecture, Nature Genetics (1999) 281–285.
[29] P.H. Westfall, S.S. Young, Resampling Based Multiple Testing, Wiley, New York, 1993.
[30] ⟨http://rsat.ulb.ac.be/rsat⟩.
[31] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, in: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, California, 1994, pp. 28–36.
[32] K. Higo, Y. Ugawa, M. Iwamoto, T. Korenaga, Plant cis-acting regulatory DNA elements (PLACE) database, Nucleic Acids Res. 27 (1999) 297–300 ⟨http://www.dna.affrc.go.jp/htdocs/PLACE⟩.

About the Author—NEELIMA GUPTA graduated with a B.Sc. in Mathematics from the University of Delhi, India in 1985. She then went on to complete her M.Sc. in Mathematics in 1987 and M.Tech. in Computer Science in 1989 from the Indian Institute of Technology, Delhi (IITD), India. She received her Ph.D. in Computer Science from IITD in 1998, where she worked on designing randomized parallel algorithms for a number of problems in computational geometry. Earlier in her career, after her M.Tech., she briefly held the position of a Software Engineer at HCL Technologies Pvt. Ltd., in New Delhi, India in 1989. She then joined HansRaj college at University of Delhi in 1989. She taught a number of courses in Computer Science at HansRaj college earlier as a Lecturer and later as a Reader from 1989 to 2002. Simultaneously during this period, she pursued her Research Work in the area of parallel algorithms and computational geometry and published a number of papers in various journals and conferences, including her Ph.D. thesis. She presently holds the position of a Reader in the Department of Computer Science at University of Delhi, which she joined in the year 2002. Her research interests include approximation algorithms for network design problems, data mining and bioinformatics.

About the Author—SEEMA AGGARWAL graduated with a B.Sc. in Physics from the University of Delhi, India in 1989. She then went on to complete her Masters in Computer Applications in 1992. She worked as Software Engineer at DCM Datasystems, in New Delhi, India in 1992. She then joined Miranda House, at University of Delhi in 1995 and is currently a Reader there. She is involved in teaching a number of courses in Computer Science at Miranda House. Currently, she is pursuing her Ph.D. in Computer Science at Department of Computer Science, University of Delhi in the area of Algorithms in Computational Biology.