A Biclustering Method to Discover Co-regulated Genes Using Diverse Gene Expression Datasets*

Doruk Bozdağ [1,2], Jeffrey D. Parvin [1], and Umit V. Catalyurek [1,2]

[1] Biomedical Informatics, The Ohio State University
[2] Electrical and Computer Engineering, The Ohio State University
{bozdagd,umit}@bmi.osu.edu, jeffrey.parvin@osumc.edu

Abstract. We propose a two-step biclustering approach to mine co-regulation patterns of a given reference gene to discover other genes that function in a common biological process. Currently, several successful methods utilize Pearson Correlation Coefficient (PCC) based gene expression analysis across all samples in datasets. However, microarray datasets are fraught with spurious samples or samples of diverse origin, and many genes/proteins that function in the same biological pathway may be missed. The novel PCC based biclustering algorithm introduced in this paper identifies subsets of genes with high correlation by stringently filtering the data and reducing false negatives due to spurious or unrelated samples in a dataset. Then, correlation information extracted from resulting biclusters are synthesized. We applied our method using the breast cancer associated tumor suppressors, BRCA1 and BRCA2, as the reference proteins to reveal genes and proteins important in the complex process of breast tumor formation. Experiments on 20 very large datasets showed that the top-ranked genes were remarkably enriched for genes that regulate the mitotic spindle and cytokinesis. The results imply that BRCA1 and BRCA2 proteins, which are considered to be DNA repair factors, have critical function regarding the mitotic spindle as well. Initial biological verification reveal that this identified factor function to control both centrosome dynamics, and also, surprisingly, DNA repair. Thus, this biclustering approach is successful at identifying proteins with highly related function from extremely complex datasets, and permits novel insights into gene function.

1 Introduction

Proteins that function in concert in a given cellular process often have their encoding mRNA co-expressed [1]. Therefore, examining transcription levels of genes under different conditions provides insight about functions of genes, and eventually development and treatment of complex diseases. DNA microarray technology has become the central enabling technology in genomic research by allowing measurement of expression levels of thousands of genes in parallel. In a microarray experiment, expression levels of genes in various samples are arranged in a matrix called *gene expression data*.

* This work was supported in parts by the U.S. DOE SciDAC Institute Grant #DE-FC02-06ER2775; the U.S. National Science Foundation Grants #CNS-0643969, #CCF-0342615, and #CNS-0426241, #CNS-0403342; Ohio Supercomputing Center Grant #PAS0052.

S. Rajasekaran (Ed.): BICoB 2009, LNBI 5462, pp. 151–163, 2009.

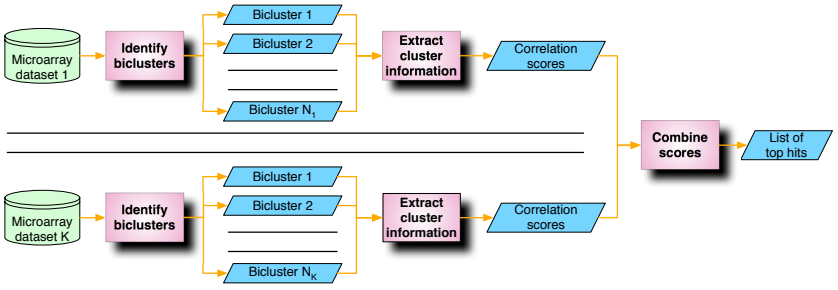

Fig. 1. Overview of the proposed approach

Samples are usually collected from different individuals and may correspond to different environmental conditions. Mining gene expression data to discover biologically relevant knowledge is a challenging task and has been the focus of many research efforts [2,3,4,5].

In this work, our objective is to develop a method that utilizes multiple gene expression datasets to identify genes exhibiting co-regulation with respect to a reference gene. Identifying genes co-regulated with a gene of important function is crucial to understand biochemical and genetic pathways in which the gene participates. A straightforward approach towards this aim is to cluster genes in each dataset using a correlation or similarity metric such as *Pearson Correlation Coefficient (PCC)* [6]; then count the number of times each gene co-occurs in the same cluster with the reference gene over all datasets. PCC is a very effective and widely used metric in this type of analysis to quantify co-regulation between pairs of genes [3,5].

A major drawback in this approach is that the entire set of samples in a dataset are used to decide cluster membership or correlation with the reference gene. Since samples are usually collected from diverse sources, genes and proteins that function together may only be similarly expressed in a subset of the samples. Moreover, most clustering techniques generate exclusive partitions of genes, therefore disregard the fact that a single gene may be involved in more than one biological pathway. To overcome these limitations we propose a new biclustering algorithm, called *Correlated Pattern Biclusters (CPB)*, that identifies groups of genes highly correlated with a given reference gene in empirically defined subsets of samples. We introduce novel techniques in CPB to address two important issues in biclustering of gene expression data: (1) mining datasets only to discover correlated patterns that contain the given reference gene, (2) extension of the use of PCC in biclustering context. In addition, CPB algorithm allows overlapping clusters and also captures negative correlation through use of PCC.

To reach our ultimate goal of identifying genes that consistently exhibit correlation with the reference gene, we also propose a method to extract correlation information from identified biclusters in an intuitive way. The proposed method evaluates uniqueness of information captured in each bicluster and computes a *correlation score* for each gene based on how frequently and in how distinct biclusters it co-occurs with the reference gene. Then, correlation scores from all datasets are combined to filter out inconsistent information. The overview of our approach is illustrated in Figure 1.

Our motivating application was from breast cancer research, where there are two important reference proteins, BRCA1 and BRCA2, highly penetrant breast cancer specific

Table 1. A sample dataset and biclusters identified by several methods from this dataset. (a) Sample matrix (b) Additive model (c) Multiplicative model (d) proposed CPB algorithm (e) OPSM.

1	2	3	8	95
2	3	4	9	21
5	6	7	12	51
2	4	6	16	18
3	6	9	24	30
15	14	13	8	7

(a)

1	2	3	8
2	3	4	9
5	6	7	12

(b)

1	2	3	8
2	4	6	16
3	6	9	24

(c)

1	2	3	8
2	3	4	9
5	6	7	12
2	4	6	16
3	6	9	24
15	14	13	8

(d)

1	2	3	8	95
2	3	4	9	21
5	6	7	12	51
2	4	6	16	18
3	6	9	24	30

(e)

tumor suppressors. Both of these proteins function in the repair of DNA damage. In addition, BRCA1 also functions at an organelle called centrosome, which is critical for cell division. To determine genes co-regulated with BRCA1 and BRCA2 we applied the method proposed in this paper on very large datasets publicly available at Gene Expression Omnibus (GEO) database [7]. The results are given in Section 5.

2 Background

Biclustering was first introduced to gene expression data analysis by Cheng and Church [8]. This is followed by numerous biclustering algorithms to identify additive, multiplicative [9,10], or even more complex relationships [2,11,12,13,14] between the rows and columns of a data matrix. In additive (multiplicative) models, the difference (ratio) between corresponding elements of any two rows and the difference (ratio) between corresponding elements of any two columns in a bicluster are constants. In general, additive models are useful to capture shifting patterns, whereas multiplicative models are useful to capture scaling patterns in the data. However, neither of them can identify shifting and scaling patterns simultaneously. Furthermore, these models are too restrictive in the sense that constant difference (ratio) constraints are applied on both row and column dimensions. In Table 1b and 1c, example biclusters that can be identified respectively by additive and multiplicative models from the sample matrix in Table 1a are shown.

In this work, we propose the CPB algorithm that utilizes statistical co-expression measure PCC as a similarity metric between rows of a bicluster. PCC is a strong metric to evaluate positive as well as negative co-regulation between rows, and is commonly used in clustering gene expression data [3,5] due to its power in capturing both shifting and scaling patterns. In Table 1d, an example bicluster identified by the CPB algorithm, where there is perfect correlation (or negative correlation) between each pair of rows is given. As shown in this figure, PCC allows capturing both shifting and scaling patterns that would be separately identified by additive and multiplicative models, respectively.

Application of PCC in biclustering context is not a trivial task and requires overcoming two challenges. Firstly, PCC lacks transitivity property. Therefore, instead of measuring closeness to a reference pattern, one has to compute all pairwise PCC values between rows in the same bicluster to measure quality. To tackle this problem, we empirically show that if two rows have a sufficiently high correlation with a reference

pattern, there is a lower bound for PCC between each these two rows. The second challenge is that, PCC is only meaningful to measure coherence between rows but is too restrictive if it is used to measure coherence between columns simultaneously. For instance, in the example in Table 1, if high PCC between each pair of columns was also enforced, only the biclusters that were identified by additive and multiplicative models would be found to match the ensuing criteria. In CPB algorithm, we enforce the coherence between columns by including a column in a bicluster only if it does not decrease correlation among the rows in the bicluster. To estimate the impact of including a column, we map columns to real numbers and capture tendency of gene expression changes in the bicluster. Then, we compute root mean squared error (RMSE) for each column to evaluate the fit of the column to this tendency pattern.

Mapping columns to real numbers induces an ordering of the columns similar to OPSM [2] and OP-cluster [11] algorithms. In a bicluster identified by OPSM algorithm, the direction of expression level change between any two columns is the same for all rows in the bicluster. OP-cluster is an extension to OPSM such that equivalence levels are defined to tolerate small differences between expression levels. Example of biclusters that would be identified by these algorithms for the example matrix in Table 1a is illustrated in Table 1e. In OPSM, the coherence between columns is defined in a more loose sense than CPB, and results in inclusion of a relatively less related column (column 5) in the bicluster. In addition to considering the direction of change, using PCC in CPB algorithm allows considering magnitude of change as well to eliminate inclusion of such columns. Moreover, it allows capturing negative correlation which is not handled by these algorithms. To the best of our knowledge, our work is the first work that uses PCC as an objective function for biclustering.

3 Correlated Pattern Biclusters Algorithm

Let R and C denote the set of rows and columns of a data matrix A, respectively, and each element a_{rc} represents the relation between row r and column c. A bicluster $B = (X, Y)$ can be defined by a subset of rows $X = \{x_1, \ldots, x_n\}$ and a subset of columns $Y = \{y_1, \ldots, y_m\}$, where $n \leq N$, and $m \leq M$ [4]. In our algorithm, we use PCC metric to decide membership of a row to a bicluster $B = (X, Y)$. We denote absolute value of PCC between rows $r, s \in R$ with respect to columns in Y by $pcc(r, s, Y)$. For a row r to be included in X, we require $pcc(r, x_i, Y)$ to be greater than a threshold for all $x_i \in X$. We also impose a constraint on the minimum size of Y to avoid getting large PCC values merely by chance. The objective of the proposed CPB algorithm can be formally defined as follows. Given a data matrix A, reference row r_r, PCC threshold ρ and minimum number of columns γ, identify a set of biclusters $B = (X, Y)$ such that $r_r \in X$, $m \geq \gamma$ and $pcc(x_i, x_j, Y) \geq \rho$ for all rows $x_i, x_j \in X$.

3.1 The Algorithm

Algorithm 1 outlines the proposed biclustering algorithm CPB. The algorithm starts with an initial bicluster $B = (X, Y)$ and improves it by iteratively moving rows and columns in and out of the bicluster using a search technique similar to mean-shift [15]. In mean-shift, the goal is to find the densest region with a certain radius (window size)in

Algorithm 1. Correlated Pattern Biclusters.

function CPB(A, r_r, w, γ, ρ')
  $step \leftarrow 1$; $B = (X, Y)$ where $X = \{r_r\}$ and Y is a random subset of columns of A.
  repeat ▷ Outer loop
    $B_{save} \leftarrow B$; $\rho'_c \leftarrow 2/3\rho'$; $\rho'_\Delta = 1/12\rho'$; $\gamma = m$; $\gamma_\Delta = \frac{m-\gamma}{4}$
    repeat
      Compute reference vector T and normalization parameters
      if $step$ mod $2 = 1$ **then**
        Update X such that $pcc(x_i, T, Y) > \rho'_c$ for all $x_i \in X$
      else
        Let r be the row with smallest $pcc(r, T, Y) > \rho'_c$
        Update Y such that $RMSE(y_k) > RMSE(r)$ for all $y_k \in Y$
      $\rho'_c \leftarrow \rho'_c + \rho'_\Delta$; $\gamma_c \leftarrow \gamma_c - \gamma_\Delta$
    until $\rho'_c > \rho'$
    $step \leftarrow step + 1$
  until $step > 20$ **or** $B = B_{save}$
  return $B = (X, Y)$

the search space. At each iteration, the center of mass of the points that are at a distance smaller than the given radius to the center of the current solution is computed. Then, the center of the solution is moved to this computed center of mass and the process is repeated until convergence. Similarly in CPB algorithm, we compare PCC between each row and a reference vector $T =< t_1, \ldots, t_m >$ that represents general tendency of rows in X with respect to the columns in the bicluster while deciding which rows to move. Vector T is analogous to cluster center in k-means or mean-shift techniques. If $pcc(r, T, Y)$ for a row r is above a certain threshold, we include r into set X and update T by only considering the rows in X. On the other hand, using a similar criterion for columns is too restrictive for our objective as explained in Section 2. Instead, a good criterion for inclusion of a column c into Y should measure the impact of c on PCC between rows $x_i \in X$. For this purpose, we use *root mean squared error (RMSE)* to evaluate similarity of tendencies of rows in X with respect to column c.

In each iteration of CPB, first, reference vector T and parameters related to normalization of data values are computed; then, either set X or set Y are updated. We do not update both sets simultaneously to avoid large fluctuations in the bicluster structure, that may slow down or prevent convergence. In the spirit of the mean-shift technique, while updating X, we include into X each row r that has $pcc(r, T, Y)$ above PCC threshold ρ'_c. While updating Y, we first determine row r that has the smallest $pcc(r, T, Y)$ above threshold ρ'_c. Then we include each column c into Y that has smaller RMSE than row r. Iterations to update bicluster end when neither X nor Y changes at an iteration or after 20 iterations (convergence is usually achieved in 5-10 iterations). We use the CPB algorithm with different parameters and initializations to discover possibly overlapping clusters that contain rows correlated with the reference row.

3.2 Computing Normalization Parameters and the Reference Vector

In order to make tendency of rows in X comparable, we apply normalization to account for different scaling and shifting patterns of rows in the bicluster. We compute a

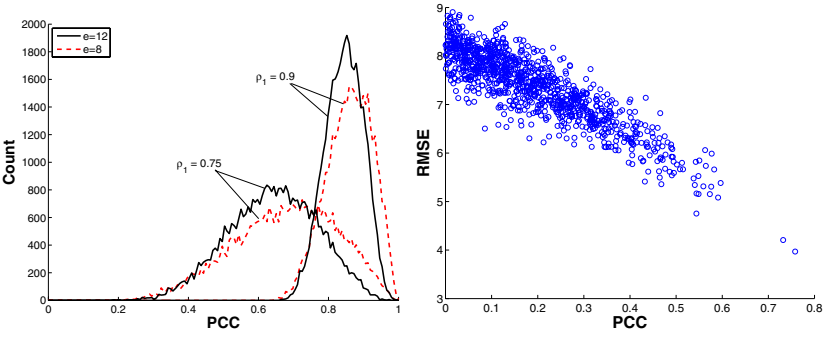

Fig. 2. (a) Distribution of PCC between pairs of 200 random vectors with e elements that have PCC with reference vector greater than a threshold ρ_1. (b) Relationship between PCC and RMSE on random vectors.

normalized data value $\hat{a}_{x_i y_k} = \frac{a_{x_i y_k} - \alpha_{x_i}}{\beta_{x_i}}$ for each $x_i \in X$ and $y_k \in Y$, where α_{x_i} and β_{x_i} are shifting and scaling parameters associated with row x_i, respectively. Then, each element t_k of reference vector T is computed as the arithmetic mean of $\hat{a}_{x_i y_k}$ on all rows $x_i \in X$. We compute T, α_{x_i} and β_{x_i} using an iterative process. Initially we set $\alpha_{x_i} = 0$ and $\beta_{x_i} = 1$, and compute T. Then, we apply least squares fitting on pairs $\{(t_1, a_{x_i y_1}), \ldots, (t_m, a_{x_i y_m})\}$ to obtain the best shifting and scaling parameters that maximize alignment of each row x_i with the reference vector T. We assign intercept and slope obtained in least squares fitting to α_{x_i} and β_{x_i}, respectively. T is updated using these parameters, and the process iterates until convergence.

3.3 Updating Rows of a Bicluster

For a row r to be a member of set X, we require $pcc(r, x_i, Y) > \rho$ for all $x_i \in X$. To avoid testing this condition against all $x_i \in X$, we utilize the reference vector T, and only test whether $pcc(r, T, Y)$ is greater than another threshold ρ' instead. ρ' is selected such that $pcc(r, T, Y) > \rho'$ must ensure $pcc(r, x_i, Y) > \rho$ for all $x_i \in X$. However, PCC lacks transitivity property [16] and has a fairly complex formula that strongly depends on the values and the length of the vectors. Therefore, it is difficult, if not impossible, to analytically compute a lower bound for ρ' as a function of ρ. To empirically determine the value of ρ' for a given ρ, we designed a simple experiment. First, we generated a reference random vector with e elements. Then we generated more random vectors and kept only those having absolute value of PCC with the reference vector greater than ρ'. After generating 200 such vectors we plotted the distribution of the absolute value of PCC between each pair of these vectors (see Figure 2a). The distributions verify that a lower bound for ρ' exists and increases with ρ.

In Algorithm 1, we start with a relaxed threshold ρ' and slowly tighten it at Line 12. While tightening ρ', we relax the constraint on minimum number of columns. This allows sweeping the search space between two extreme combinations of these parameters. In our code we use 5 tightening steps and initial values for ρ'_c and γ_c are set to $2/3\rho'$, and the number of columns in the initial bicluster, respectively (Line 4).

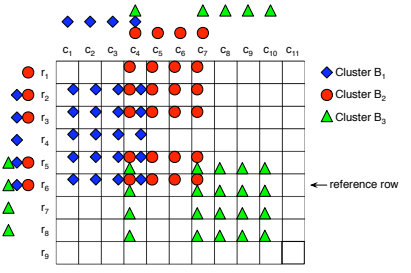

Fig. 3. Example biclusters on an example data matrix with reference row $r_r = r_6$

3.4 Updating Columns of a Bicluster

We use $RMSE$ to assess coherence of tendencies of rows $x_i \in X$ in a given column. $RMSE(y_k)$ for a column $y_k \in Y$ is computed as $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{a}_{x_i y_k} - t_k)^2}$. For a column $c \notin Y$, we compute $RMSE(c)$ in a similar way, by using a value t_c analogous to t_k that quantifies tendency of rows $x_i \in X$ in column c.

In CPB, only the columns having $RMSE$ below a threshold ϵ are included in the bicluster. In order to have control on the ratio of the number of rows to the number of columns in the bicluster, we select ϵ in relation to ρ'. To establish this relation, first we note that $RMSE$ can also be computed for rows, and it is a comparable metric for rows and columns. For a row $x_i \in X$, $RMSE(x_i)$ is computed as $\sqrt{\frac{1}{m}\sum_{k=1}^{m}(\hat{a}_{x_i y_k} - t_k)^2}$. Then, we observe that $RMSE(r)$ generally implies a high $pcc(r, T, Y)$ (see Figure 2b). Therefore, by setting ϵ to the $RMSE$ of row r that has the smallest $pcc(r, T, Y)$ above threshold ρ'_c (Line 10), we expect that the ratio n/m in the resulting bicluster is close to the ratio N/M. In order to obtain biclusters with different n/m ratios, we use parameter κ. Then, when updating set Y, κ times the number of columns with $RMSE$ above the threshold are included into Y.

To ensure that the reference row r_r has a larger impact in decision mechanisms of the algorithm, we assign a larger weight to the reference row when computing the vector T and $RMSE$ values. Total contribution from rows except r_r is multiplied by $(1-\omega)$ and contribution from r_r is multiplied by ω, where ω is an input parameter. Large values for ω allows discovering patterns that more closely resemble r_r; whereas small values increase sensitivity, hence offers higher tolerance to noise.

4 Combining Correlation Information

In this section, we explain our method to extract correlation information from identified biclusters. For this purpose, first we quantify uniqueness of information captured by each bicluster. Then, for each row we compute a *correlation score* based on co-occurrence frequency and uniqueness information associated with the row with respect to the reference row. Finally, we combine correlation scores from different datasets.

If two biclusters $B_v = (X_v, Y_v)$ and $B_w = (X_w, Y_w)$ do not overlap except for the reference row r_r, then these two biclusters represent two distinct relationships between

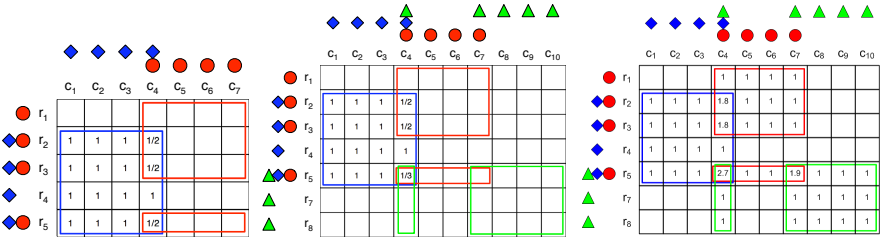

Fig. 4. $1/|IR(r,x_i) \cap IC(c,y_k)|$ values for each $x_i \in X_1$ and $y_k \in Y_1$ for (a) $r = r_2$, $c = c_4$, (b) $r = r_5$, $c = c_4$. (c) $OS(r,c)$ for each row r and column c.

rows and columns of the data matrix. In the context of gene expression, this may correspond to two different biological functions associated with the reference gene. On the other hand, if $X_w \subseteq X_v$ and $Y_w \subseteq Y_v$ the relationship in B_w is already captured by B_v. In the latter case we discard B_w from the result set.

Let $IR(\ldots)$ denote the set of biclusters that contain all rows specified in the argument list. Similarly, let $IC(\ldots)$ denote the set of biclusters that contain all columns specified in the argument list. Consider a row r, a column c and a bicluster $B_v = (X_v, Y_v)$ such that $r \in X_v$ and $c \in Y_v$. To measure uniqueness of information in B_v with respect to other biclusters in set $IR(r) \cap IC(c)$ on the relationship between r and c, we define a *bicluster uniqueness* measure $BU(B_v, r, c)$ as follows.

$$BU(B_v, r, c) = \frac{\sum_{x_i \in X_v - \{r_r\}} \sum_{y_k \in Y_v} \frac{1}{|IR(r,x_i) \cap IC(c,y_k)|}}{(|X_v| - 1)|Y_v|} \tag{1}$$

If B_v does not overlap with any other bicluster at row r and column c, then $IR(r,x_i) \cap IC(c,y_k)$ only contains B_v for all $x_i \in X$ and $y_k \in Y_v$ in (1). In this case $BU(B_v,r,c)$ takes its maximum possible value of 1. This means that B_v captures the relationship between row r and column c exclusively. $BU(B_v,r,c)$ decreases as overlap between B_v and biclusters in $IR(r) \cap IC(c)$ increases. In the case that B_v completely overlaps with all clusters in $IR(r) \cap IC(c)$, information on the relationship between row r and column c is shared between all of these clusters. Then $BU(B_v,r,c)$ takes its minimum value of $1/|IR(r) \cap IC(c)|$ (note that this case is not actually possible since we remove biclusters that are subsets of other biclusters beforehand). Computing cluster uniqueness as given in (1) is useful to avoid some relationships to be over-emphasized due to convergence of biclustering algorithm to solutions close to each other in the search space.

An example matrix and three biclusters are shown in Figure 3. Consider bicluster $B_1 = (X_1, Y_1)$ and row r_2 and column c_4 of the matrix . Since $IR(r_2) \cap IC(c_4) = \{B_1, B_2\}$, overlaps between B_1 and B_2 need to be considered when computing $BU(B_1, r_2, c_4)$. Figure 4a shows values $\mathbf{1}/|IR(r_2,x_i) \cap IC(c_4,y_k)|$ for each $x_i \in X_1$ and $y_k \in Y_1$. Applying these values to (1) gives $BU(B_1, r_2, c_4) = 0.91$. Corresponding values to compute $BU(B_1, r_5, c_4)$ are given in Figure 4b. Here $IR(r_5) \cap IC(c_4) = \{B_1, B_2, B_3\}$, thus $BU(B_1, r_5, c_4) = 0.9$.

Using bicluster uniqueness measure, we compute an *overlap score* $OS(r, c)$ for every row-column pair (r, c) to quantify the amount of different relationships identified between r and c. We compute $OS(r, c)$ by summing $BU(B_v, r, c)$ for all biclusters in $IR(r) \cap IC(c)$. In other words, $OS(r, c) = \sum_{B_v \in IR(r) \cap IC(c)} BU(B_v, r, c)$. Then, the we compute a *correlation score* $CS(r)$ for each row r by summing overlap scores of the pairs (r, c) across all columns, i.e. $CS(r) = \sum_{c \in C} OS(r, c)$. Summing overlap scores across columns gathers total evidence on how frequently and in how distinct relationships row r is correlated with the reference row r_r.

In Figure 4c, $OS(r, c)$ is given for pair (r, c) Summing these values across columns gives $CS(r_1) = CS(r_4) = 4$, $CS(r_7) = CS(r_8) = 5$, $CS(r_2) = CS(r_3) = 7.8$ and $CS(r_5) = 12.6$. As expected, rows that appear in larger number of biclusters and in more diverse relationships together with the reference row have larger correlation score.

To increase significance and consistency of our findings, we apply our method on different datasets separately and combine correlation scores. To achieve this in a meaningful way, we require datasets to have the same row labels. In gene expression data analysis, this requirement can be met by merging results only from datasets obtained using the same microarray chip. Even though such datasets could be combined into a single data matrix, this approach requires undoing any normalization previously carried out on each dataset. Since data are collected from different sources, this approach may not be practical or even possible if information about the normalization procedures are unavailable. As an alternative approach, we use the following three-step method: First, for each dataset, we divide correlation score of each row by that of the reference row in the same dataset. Then, in order to make contribution from each dataset equal, we scale correlation scores such that sum of the scores in each dataset is the same. Finally, we sum the scaled scores across datasets to compute a total score for each row.

5 Experimental Results

5.1 Experiments on Synthetic Data

To demonstrate the effectiveness of CPB, we generated datasets with embedded biclusters and applied CPB to find these biclusters. We first generated a 10000×100 matrix and a reference row vector of length m, filled with random real numbers between 0 and 100. Then, we generated $n - 1$ additional vectors, each having perfect positive or negative correlation with the reference vector. These vectors together represent an $n \times m$ bicluster. Next, we added a random number between 0 and K chosen from normal distribution to each entry in the bicluster to simulate noise in the data. Finally, we embedded the bicluster into randomly selected n rows and m columns of the dataset.

As with most clustering algorithms, there is no single set of parameter values of CPB that will suit to all datasets. Therefore, when using CPB, we consider a range of values for each parameter to scan the search space thoroughly. In our experiments on synthetic datasets we generated 10 datasets for every combination of $n = \{30, 60, 90, 120, 150\}$, $m = \{30, 60, 90\}$ and $K = \{0, 1, 2\}$. First, we applied the CPB algorithm with row column ratio parameter $\kappa = 1$, $\rho' = 0.9$ and relative weight ω of reference gene selected from $\{0.1, 0.25, 0.5, 0.75\}$. For each value of ω we applied CPB 21 times

Table 2. Datasets we used in our experiments from GEO [7] database

GDS dataset ID	534	596	715	1067	1209	1220	1284	1375	1479	1615
Number of samples	75	158	87	52	54	54	50	70	60	127
GDS dataset ID	1781	1815	1956	1975	2113	2190	2255	2362	2373	2643
Number of samples	104	100	121	85	76	61	58	71	130	56

Table 3. Intersection of top-25 lists of BRCA1 and BRCA2 reference probe sets

Affymetrix probe set ID	Associated protein	Affymetrix probe set ID	Associated protein
201292_at	TOP2A	210052_s_at	TPX2
202095_s_at	BIRC5	214710_s_at	CCNB1
202705_at	CCNB2	218009_s_at	PRC1
204962_s_at	CENPA	218039_at	NUSAP1
209642_at	BUB1	218355_at	KIF4A

using different initial clusters. The value of threshold ρ corresponding to $\rho' = 0.9$ was 0.65. This value is obtained by the method explained in Section 3. We consider an embedded bicluster identified, if the returned bicluster consists of at least half of the rows and half of the columns of the embedded bicluster. If all rows and columns of the embedded bicluster are returned, we call the bicluster perfectly identified. When there was no noise in the data, CPB algorithm perfectly identified 148 of the 150 embedded biclusters. When some noise is added, the bicluster structure is more difficult to discover due to reduced PCC values between the rows. Furthermore, it is likely that some of the rows will no longer have PCC above 0.9 with the reference row. In our experiments with $K = 1$, CPB was able to identify 145 of the 150 biclusters. Of these, 140 were perfectly identified, and the for the remaining ones, all rows and at least 90% of the columns were returned by the CPB algorithm. Finally, when K was 2, there was a more pronounced impact of noise resulting in much reduced PCC between the rows. Still, CPB algorithm successfully identified 131 of the 150 embedded biclusters. For 91 of these biclusters, CPB returned at least two thirds of the columns and two thirds of the rows.

Next, we applied a PCC based clustering approach to identify rows having PCC greater than 0.65 with the reference row over all columns. Even in the best case, at most 27% of the columns in a bicluster were successfully identified by this approach. This shows that considering all columns to compute PCC prevents detection of biclusters.

5.2 Identifying Genes Co-regulated with BRCA1 and BRCA2

For real data experiments we selected 20 large datasets each obtained using Affymetrix HG U133 GeneChip Array and having at least 50 samples (Table 2). This array has 22,215 probe sets including two probe sets for each of BRCA1 (204531_s_at, 211851_x_at) and BRCA2 (208368_s_at, 214727_at). For each run of CPB, κ is selected from $\{1, 3, 5, 7, 9\}$; ω from $\{0.25, 0.5, 0.75\}$; and ρ' was set to 0.9. We executed the algorithm for every combination of these values, and for each parameter set

Table 4. GO term enrichment results using 90 genes obtained by intersecting top-500 lists of reference genes. n and N represent the number of genes associated with the GO term in the set of identified genes and in the Affymetrix chip, respectively.

GO term ID	GO term	p-value	n	N
GO:0000910	cytokinesis	$< 1.0 \times 10^{-12}$	8	40
GO:0007049	cell cycle	$< 1.0 \times 10^{-12}$	38	720
GO:0007067	mitosis	$< 1.0 \times 10^{-12}$	34	205
GO:0031577	spindle checkpoint	$< 1.0 \times 10^{-12}$	3	3
GO:0040001	establishment of mitotic spindle localization	$< 1.0 \times 10^{-12}$	4	4
GO:0045842	positive regulation of mitotic metaphase/anaphase transition	$< 1.0 \times 10^{-12}$	3	3
GO:0051301	cell division	$< 1.0 \times 10^{-12}$	29	266
GO:0051303	establishment of chromosome localization	$< 1.0 \times 10^{-12}$	3	3
GO:0007051	spindle organization and biogenesis	2.9×10^{-12}	5	9
GO:0031503	protein complex localization	7.6×10^{-12}	6	8
GO:0031536	positive regulation of exit from mitosis	1.0×10^{-11}	4	7

we generated 21 random initial clusters. We applied the analysis four times using one of the BRCA1 or BRCA2 probe sets as the reference each time. In Table 3, we present genes that appeared in top-25 highest correlated gene list of each of the four reference probe sets.

There are 90 genes that were common in top-500 list for all four reference probe sets. Analysis of Gene Ontology (GO) terms associated with these 90 genes statistically supports the extraordinary clustering of proteins that function in mitosis. The top-ranked genes are remarkably enriched for genes that regulate the mitotic spindle and cytokinesis. As given in Table 4, of these 90 genes, 38 control the cell cycle, 34 relate to mitosis and 29 involved in cell division. The enrichment of cell cycle, mitosis, and cellular assembly are exactly what would be predicted for control by the centrosome. DNA replication and repair would be predicted to be a part of the BRCA1 and BRCA2 module, and this pathway would also impact the centrosome.

The results show that our algorithm is successful at identifying from extremely complex datasets proteins with highly related function. While our results did reveal known factors for the repair of DNA damage as expected, the most significant results were enriched for centrosome and mitotic spindle related processes. This implies that BRCA1 and BRCA2, which are considered to be DNA repair factors, also have critical function regarding the mitotic spindle. Biological testing of this point is in progress, but in initial tests a gene of unknown function identified by our method is found to control centrosome[1]. If confirmed, this will imply that control of the mitotic spindle is a critical control element in breast cancer. In addition, several of the identified proteins that function to control the centrosome were found to also control a DNA repair assay. This was an unanticipated finding. Thus, biological validation in progress is revealing that this biclustering tool both reveals proteins that function together to control centrosomes and also to participate in a second process of DNA damage repair.

We have also tested this method for false positives by applying the biclustering tool on seven more genes to verify that our method avoids systematic errors. Two of the

[1] A recent work of J. D. Parvin, unpublished.

genes we used for this analysis are RB1 and TP53, tumor suppressor genes involved in many cancers. The other five genes were CCNB2, FGD2, TAF7, SRP54 and CHPF, which were ranked 1^{st}, 5000^{th}, 10000^{th}, 15000^{th} and 20000^{th}, respectively, when correlation scores of four reference probe sets are combined. We used each of these selected genes as anchors and applied our analysis to determine top-25 lists for high correlation for each of these genes. This analysis verified that BRCA1, BRCA2, as well as the number one hit CCNB2 are correlated with a similar set of genes. For each pair of these genes there were 16 to 20 genes at the intersection of top-25 lists. On the other hand, the genes we selected for verification had at most one gene in common in the top-25 list of either of BRCA1, BRCA2 or CCNB2.

6 Conclusion and Future Work

In this work, we proposed a two-step approach to mine co-regulation patterns, relative to a set of reference genes, that may only exist in a subset of samples. First, co-regulation patterns in microarray datasets are discovered using a novel PCC-based biclustering algorithm. Then, correlation information is combined to compute a correlation score with respect to the reference gene. In our experiments we used BRCA1 and BRCA2 as our reference genes. Analysis of the top-ranked genes using GO terms revealed an extraordinary clustering of proteins that function in mitosis. In the future, we plan to compare the CPB algorithm with other biclustering algorithms in terms of both objective functions and optimization techniques. Furthermore, we will evaluate significance of our findings by testing the algorithm on various real datasets.

References

1. Agrawal, H.: Extreme self-organization in networks constructed from gene expression data. Phys. Rev. Lett. 89(26), 268702 (2002)
2. Ben-Dor, A., Chor, B., Karp, R., Yakhini, Z.: Discovering local structure in gene expression data: The order-preserving submatrix problem. Int'l Conf. Comput. Biol., 49–57 (2002)
3. Jiang, D., Pei, J., Zhang, A.: DHC: A density-based hierarchical clustering method for time series gene expression data. In: IEEE Symp. Bioinform. and Bioeng., pp. 393–400 (2003)
4. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Comput. Biology Bioinform. 1(1), 24–45 (2004)
5. Pujana, M.A., et al.: Network modeling links breast cancer susceptibility and centrosome dysfunction. Nature Genetics 39(11), 1338–1349 (2007)
6. Devore, J.L.: Probability and Statistics for Engineeringand Sciences. Brook/Cole Publishing Company (1991)
7. Edgar, R., Domrachev, M., Lash, A.E.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucl. Acids Res. 30(1), 207–210 (2002)
8. Cheng, Y., Church, G.M.: Biclustering of expression data. Int'l Conf. Intelligent Systems for Molecular Biology, 93–103 (2000)
9. Segal, E., Taskar, B., Gasch, A., Friedman, N., Koller, D.: Rich probabilistic models for gene expression. Bioinformatics 17(suppl. 1), S243–S252 (2001)
10. Wang, H., Wang, W., Yang, J., Yu, P.S.: Clustering by pattern similarity in large data sets. In: ACM SIGMOD (2002)

11. Liu, J., Wang, W.: Op-cluster: Clustering by tendency in high dimensional space. In: IEEE Int'l. Conf. Data Mining, p. 187 (2003)
12. Murali, T., Kasif, S.: Extracting conserved gene expression motifs from gene expression data. In: Pac. Symp. Biocomp., vol. 8 (2003)
13. Tanay, A., Sharan, R., Shamir, R.: Discovering statistically significant biclusters in gene expression data. Bioinformatics 18(suppl. 1), 136–144 (2002)
14. Kluger, Y., Basri, R., Chang, J.T., Gerstein, M.: Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res. 13(4), 703–716 (2003)
15. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Analysis and Machine Intelligence 24(5), 603–619 (2002)
16. Casella, G., Wells, M.T.: Is Pitman closeness a reasonable criterion?: Comment. Journal of the American Statistical Association 88(421), 70–71 (1993)