

Querying Co-regulated Genes on Diverse Gene Expression Datasets Via Biclustering

Mehmet Deveci, Onur Küçüktunç, Kemal Eren, Doruk Bozdağ,
Kamer Kaya, and Ümit V. Çatalyürek

Abstract

Rapid development and increasing popularity of gene expression microarrays have resulted in a number of studies on the discovery of co-regulated genes. One important way of discovering such co-regulations is the query-based search since gene co-expressions may indicate a shared role in a biological process. Although there exist promising query-driven search methods adapting clustering, they fail to capture many genes that function in the same biological pathway because microarray datasets are fraught with spurious samples or samples of diverse origin, or the pathways might be regulated under only a subset of samples. On the other hand, a class of clustering algorithms known as *biclustering* algorithms which simultaneously cluster both the items and their features are useful while analyzing gene expression data, or any data in which items are related in only a subset of their samples. This means that genes need not be related in all samples to be clustered together. Because many genes only interact under specific circumstances, biclustering may recover the relationships that traditional clustering algorithms can easily miss. In this chapter, we briefly summarize the literature using biclustering for querying co-regulated genes. Then we present a novel biclustering approach and evaluate its performance by a thorough experimental analysis.

Keywords: Biclustering, Microarray, Gene expression, Clustering

“What we call chaos is just patterns we haven’t recognized. What we call random is just patterns we can’t decipher.”

— Chuck Palahniuk, Survivor

1 Introduction

The microarray technology enables large-scale genomic research by allowing the measurement of the expression levels of thousands of genes in parallel. Expression levels of genes in various samples are collected and stored in a gene expression matrix. Mining these gene expression matrices can provide insights into gene functions and aids in the development and treatment of complex diseases. The discovery of related genes is a challenging task and has been the focus of many research studies [1–4] that search for more sophisticated analysis methods. Most of the time, however, researchers

focus on a specific gene or a gene set rather than exploring the whole dataset. Query-based search algorithms [5–10] are proven to be very useful when the objective is to rank the genes according to how strongly they are correlated with the queried gene(s). For example, several genes in *S. cerevisiae* database are categorized and annotated by Hibbs et al. [6]. Similarly, top-ranked genes co-regulated with breast cancer associated tumor suppressors, BRCA1 and BRCA2, are found to be regulating the mitotic spindle and cytokinesis by Bozdağ et al. [9]. In analyzing this torrent of new data, unsupervised learning methods such as clustering are important as the first step. In particular, a class of clustering algorithms known as *biclustering* is useful for analyzing gene expression data, or any data whose items are related in only a subset of their samples. Biclustering methods cluster both the items and their features simultaneously. In gene expression context, this means that genes need not be related in all samples to be clustered together. Because many genes only interact under specific circumstances, biclustering may recover relationships that traditional clustering algorithms can miss.

In this chapter, first, we briefly survey the literature on biclustering and proposed algorithms. Then we introduce a novel biclustering algorithm, Correlated Patterns Biclustering (CPB), which attempts to find genes that are related on a subset of their features with a query gene. As mentioned above, identifying the genes co-regulated with a gene of important function is crucial to understand biochemical and genetic pathways in which the gene participates. To quantify gene relationships, CPB uses the Pearson correlation coefficient (PCC), an effective and widely used metric in this type of analysis to quantify co-regulation between pairs of genes [2, 4]. CPB's novel approach avoids costly pairwise correlation calculations in a manner that also increases its accuracy. It also allows assigning genes to multiple biclusters, because many genes participate in multiple biological pathways. We further introduce a unique method for combining results from multiple datasets, which is important for uncovering uncommon genetic relationships. Initial testing on artificial data shows that CPB outperforms other biclustering methods in finding multiple types of biclusters. CPB's performance for querying the microarray data is similarly promising: it was able to find many genes that have high correlation with BRCA1, BRCA2, and p53. Of those genes, half are already known to be involved in cancer processes, and the others are promising new candidates for further investigation. The source code of the framework, documentation, and sample datasets is available at <http://bmi.osu.edu/hpc/software/cpb/>.

The methods in this chapter extend the framework proposed by Bozdağ et al. [9] to increase the efficiency of the algorithms as well as the consistency and relevancy of the results. The novelty of the proposed algorithm can be summarized as follows:

- A grid-based method is used for generating initial biclusters, which covers the whole dataset.
- Results are investigated and the statistically insignificant biclusters are filtered out with a non-parametric scheme.
- The biclustering method is tested on various models, noise levels, and overlap ratios; compared with other techniques.
- Correlation scores of the genes are computed and combined more efficiently.

The key advantages of the proposed query-based search framework are:

- It finds co-regulated genes with a given reference gene on a number of diverse microarray datasets having the same genes. This is the case for data obtained from a single microarray.
- PCC-based biclustering technique is able to discover constant-row, shift, scale, and shift-scale models with positive and negative correlations.
- CPB is extremely efficient compared to other PCC-based methods because of a novel correlation calculation.
- Filtering step increases the relevance of the results while eliminating insignificant and overlapping biclusters.

The rest of the chapter is organized as follows: In Section 2, biclustering algorithms from the literature are briefly surveyed. Section 3 describes the CPB algorithm. The results of the proposed algorithm and framework's experimental evaluation are given in Section 4. Section 5 concludes the chapter.

2 Biclustering of the Microarray Data

Biclustering refers to a class of methods that perform simultaneous clustering of both rows and columns of a data matrix. It was first introduced to gene expression data analysis by Cheng and Church [11]. This initial algorithm was followed by numerous biclustering algorithms to identify additive, multiplicative [12, 13], or more complex relationships [14–22] between the rows and columns of a data matrix that correspond to genes and samples, respectively.

A straightforward two-phase approach to identify the biclusters is applying standard clustering algorithms to the genes and samples separately in the first step, and combine the results in the second one [23]. However, the research on biclustering is focused to a more integrated approach in which the genes and samples are analyzed simultaneously. Several randomized or deterministic algorithms based on both novel and existing techniques from various domains, such as independent component analysis, singular value decomposition, simulated annealing, and local search, have been

proposed, i.e., [24–33], and evaluated on the gene expression data for many diseases including the complex ones such as cancer. Some algorithms in this set use greedy techniques, i.e., [11, 34], and some employ evolutionary techniques [35–40]. In addition, graphs, modeling pairwise gene–gene interactions, have also been employed to design novel biclustering methods. For example, a local, correlated structure in the graph obtained by the gene expression data is shown to be promising to be used as a bicluster [41].

In the literature, bicluster models that a biclustering algorithm seeks for can be divided into two categories. Global biclusters are defined by comparing a metric within the bicluster to the outside of the bicluster. Up-regulated biclusters with higher expression values compared to background, and down-regulated biclusters with lower expression values than the background are examples of global biclusters. Many algorithms have been proposed to capture the global biclusters such as SAMBA [42], ISA [43], Spectral [44], BiMax [45], QUBIC [46], COALESCE [47], BBK [48]. On the other hand, local biclusters can be defined by the relationships within the bicluster columns and rows such as constant, additive, and multiplicative biclusters. Additive models are useful for capturing shifting patterns (see Fig. 1b), whereas multiplicative models are useful for capturing scaling patterns (see Fig. 1c) in the data. However, neither can simultaneously identify the shifting and scaling patterns. In this chapter, we will seek biclusters fitting the shift-scale model (see Fig. 1d) which covers both additive and multiplicative patterns as special cases.

How to evaluate the quality of the biclusters is also an important problem: for example, the classical mean squared residue (MSR) has been shown to be successful at finding constant, and additive biclusters, while it is not suitable for multiplicative biclusters. In addition, it is claimed to be biased toward flat biclusters with low row variance, and hence, different scoring schemas have been proposed Bryan and Cunningham [49]. Cheng and Cherk propose a deterministic greedy algorithm that seeks to find the biclusters with low variance, as defined by the MSR [11].

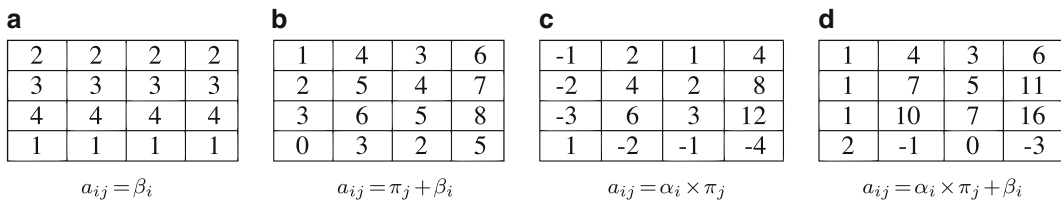


Fig. 1 Sample biclusters with various models: (a) constant-row, (b) shift, (c) scale, and (d) shift-scale. In pattern expressions, a_{ij} represents expression level of gene i in sample j , π_j a base value, α_i scaling, and β_i shifting patterns. The parameters are selected as $\alpha_i = [1, 2, 3, -1]^T$, $\beta_i = [2, 3, 4, 1]^T$, $\pi_j = [-1, 2, 1, 4]$. Shift-scale is the most general model, as it has shift and scale models as special cases and can represent both positive and negative correlation

Similarly, the xMOTIFs algorithm has been proposed to capture conserved gene expression motifs that are the biclusters with conserved rows in discretized dataset [50]. A more complex relationship among the genes has been later studied in order-preserving submatrix problem (OPSM) [1, 51]. The authors propose a deterministic greedy algorithm that seeks biclusters for which the columns can be sorted in increasing order for all rows in the bicluster. Although additive and multiplicative biclusters can be captured by OPSM algorithm, it fails to capture constant biclusters. Surveys on biclustering of gene expression data, the proposed algorithms, and their evaluation via bicluster validation from a biological point of view can be found in [3, 45, 52–56].

2.1 PCC-Based Biclustering

PCC is a measure that evaluates positive and negative linear relationships between vectors. It is commonly used in clustering gene expression data [2, 4] due to its power in capturing both shifting and scaling patterns. For a PCC-based biclustering on gene expression dataset, the correlation of two genes is calculated on some specified columns since those genes may or may not be correlated on every experiment. Therefore, our PCC-based similarity measure between rows r and s on selected columns \mathcal{Y} is calculated with:

$$pcc(r, s, \mathcal{Y}) = \frac{\left| \sum_{i \in \mathcal{Y}} (r_i - \bar{r})(s_i - \bar{s}) \right|}{\sqrt{\sum_{i \in \mathcal{Y}} (r_i - \bar{r})^2 \sum_{i \in \mathcal{Y}} (s_i - \bar{s})^2}}, \quad (1)$$

where the equation runs on select columns, and the absolute value of the expression gives a result in $[0, 1]$ interval.

PCC-based biclustering was recently proposed in [9, 57]. In [57], the authors present the bi-correlation clustering algorithm (BCCA), which tries to find biclusters using Pearson correlation. They also discuss the complexity of computing pairwise PCCs, and the inefficiency of the method. Bozdağ et al. [9] discuss potential complexity issues of an exhaustive search using PCC, and propose that, instead of computing all pairwise PCC values, a center-like vector (*tendency vector*) is sufficient and more efficient at finding correlated rows.

3 Correlated Pattern Biclusters

Given a query gene and a set of microarray datasets, we compute a ranked list of co-regulated genes in three steps. Here we give the details of these steps: In the first step, the CPB algorithm recovers a set of biclusters (Section 3.1). In the next step, we filter out statistically insignificant biclusters (Section 3.2). Finally, the correlation scores gathered from different datasets (Section 3.3). The overview of the framework is given in Fig. 2.

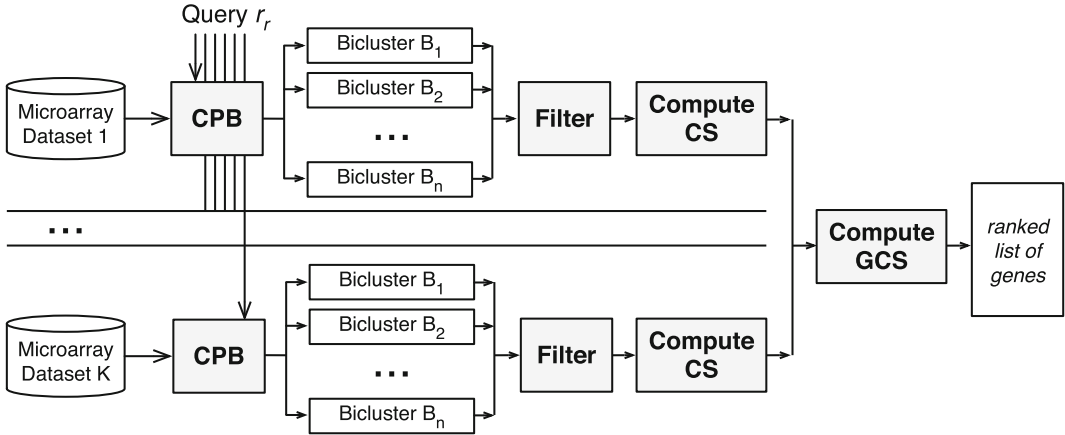


Fig. 2 Overview of the proposed framework

3.1 The CPB Algorithm

Let R and C denote the set of rows and columns of a data matrix \mathbf{A} , respectively. Each element $a_{rc} \in \mathbf{A}$ represents the relation between row r and column c . A bicluster $B = (X, \Upsilon)$ is a subset of rows $X = \{x_1, \dots, x_n\}$ and a subset of columns $\Upsilon = \{y_1, \dots, y_m\}$, where $n \leq N$, and $m \leq M$.

Definition 1 (Correlated Pattern Biclusters Algorithm). *Given a data matrix \mathbf{A} , reference row r_r , PCC threshold ρ , and minimum number of columns γ , CPB finds a bicluster $B = (X, Y)$ such that $r_r \in X$, $m \geq \gamma$, $\forall_{x_i, x_j \in X} pcc(x_i, x_j, Y) \geq \rho$.*

CPB starts with an initial bicluster $B = (X, \Upsilon)$ and improves it by iteratively moving rows and columns in and out of the bicluster using a search technique similar to local search methods. Algorithm 1 outlines the proposed biclustering algorithm. Important steps, i.e., generation of the initial biclusters, computing tendency vector T and normalization parameters, updating rows and columns are described in detail in the following subsections.

Algorithm 1 Correlated Pattern Biclusters

```

1: function CPB( $\mathbf{A}, B, r_r, w, \gamma, \rho'$ )
2:    $B = (X, Y)$  is an initial bicluster s.t.  $r_r \in X$ 
3:    $\rho'_c \leftarrow 2/3\rho'$ ;  $\rho'_\Delta = 1/12\rho'$ ;  $\gamma_c = m$ ;  $\gamma_\Delta = \frac{m-\gamma}{4}$ 
4:   repeat
5:     step  $\leftarrow 0$ 
6:     repeat
7:       step  $\leftarrow$  step + 1;  $B_{save} \leftarrow B$ 
8:       Compute  $T, \alpha_i, \beta_i$ 
9:       if step mod 2 = 1 then
10:        Update  $X$  such that
11:         $\forall x_i \in X, pcc(x_i, T, Y) > \rho'_c$ 
12:       else
13:        Find row  $r$  with smallest  $pcc(r, T, Y) > \rho'_c$ 
14:        Update  $Y$  such that
15:         $\forall y_k \in Y, ERROR(y_k) > ERROR(r)$ 
16:       end if
17:     until step > 20 or  $B = B_{save}$ 
18:      $\rho'_c \leftarrow \rho'_c + \rho'_\Delta$ ;  $\gamma_c \leftarrow \gamma_c - \gamma_\Delta$ 
19:   until  $\rho'_c > \rho'$ 
20:   return  $B = (X, Y)$ 
21: end function

```

3.1.1 Generating Initial Biclusters

Selecting the rows and columns of the initial bicluster is important since the algorithm converges to a more stable one by adding and removing rows and columns to this bicluster. In [9], initial biclusters were chosen randomly, and the algorithm runs efficiently when discovering small number of biclusters embedded in synthetic datasets. However, we observe that when there are multiple biclusters this approach does not provide a consistent mechanism to return multiple biclusters with good coverage of the whole dataset.

In CPB, we generate initial biclusters with a grid-based approach. We first shuffle the row and column numbers of the dataset, and then partition the dataset into a coarse-grain grid of 10×2 initial biclusters. The query gene r_r is inserted into each bicluster, if necessary. At the end, all genes and conditions in the dataset are assigned to at least one initial bicluster. Repeating the process gives us enough initial biclusters to find co-regulated genes and corresponding conditions. In addition, different runs obtain more than 75 % of the top-ranked co-regulated genes with the grid-based initialization, even though the generation of the initial biclusters is randomized.

3.1.2 Computing Normalization Parameters and Tendency Vector

In order to avoid making pairwise comparisons of all rows, we compute a *tendency vector* that represents an average of the rows of the bicluster. We compute a normalized data value

$$\tilde{a}_{x_i y_k} = \frac{a_{x_i y_k} - \alpha_{x_i}}{\beta_{x_i}}$$

for each $x_i \in X$ and $y_k \in \mathcal{Y}$, 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 tendency vector T is computed as the arithmetic mean of $\tilde{a}_{x_i y_k}$ on all rows $x_i \in X$.

To ensure that the reference row r_r has a larger impact on decision mechanisms of the algorithm, we assign a larger weight, ω , to the reference row when computing the vector T . 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 allow discovering patterns that resemble r_r more closely, whereas small values reduce sensitivity, hence offer a higher tolerance to noise. Therefore, if a reference row and ω specified, the elements are calculated with

$$t_k = \frac{\omega \times \tilde{a}_{x_i y_{r_r}} + (1 - \omega) \times \sum_{k \in X - \{r_r\}} \tilde{a}_{x_i y_k}}{|X|}. \quad (2)$$

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 $\{(t_1, a_{x_i y_1}), \dots, (t_m, a_{x_i y_m})\}$ to obtain the best shifting and scaling parameters that maximize alignment of

each row x_i with the tendency 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.

3.1.3 Updating the Rows of a Bicluster

For a row r to be included in X , we require $pcc(r, x_i, \mathcal{Y}) > \rho$ for all $x_i \in X$. To avoid testing this condition against all $x_i \in X$, we utilize the tendency vector T , and only test whether $pcc(r, T, \mathcal{Y})$ is greater than another threshold ρ' instead. ρ' is selected such that $pcc(r, T, \mathcal{Y}) > \rho'$ must ensure $pcc(r, x_i, \mathcal{Y}) > \rho$ for all $x_i \in X$. However, PCC lacks transitivity property [58] and has a complex formula that strongly depends on the values and the length of the vectors. Although it is analytically difficult to compute a lower bound for ρ' , it was empirically shown that there exists a lower bound proportional to ρ [9].

In Algorithm 1, we start with a relaxed threshold and slowly tighten it at Line 18. 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. The algorithm uses five tightening steps and initial values of $\rho'_c = 2/3\rho'$ and $\gamma_c = |\mathcal{Y}|$ (Line 3).

3.1.4 Updating the Columns of a Bicluster

Using PCC to measure the coherence between the columns is too restrictive. For example, although the rows in Fig. 1d are perfectly correlated, Pearson correlation between columns is less than 1. Therefore, we use root mean square error to assess the coherence of the columns. It is computed as:

$$ERROR(y_k) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\tilde{a}_{x_i y_k} - t_k)^2}, \quad (3)$$

where $y_k \in \mathcal{Y}$ and $n = |\mathcal{Y}|$. For a column $c \notin \mathcal{Y}$, we compute $ERROR(c)$ 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 $ERROR$ below a threshold ε are included in the bicluster. In order to have comparable $ERROR$ threshold for the column selection with respect to row addition, we select ε in relation to ρ' . To establish this relation, first we note that $ERROR$ can also be computed for rows, and it is a comparable metric for rows and columns. For a row $x_i \in X$, $ERROR(x_i)$

is computed as $\sqrt{\frac{1}{m} \sum_{k=1}^m (\tilde{a}_{x_i y_k} - t_k)^2}$. Then, it is observed that

$ERROR(r)$ generally implies a high $pcc(r, T, \mathcal{Y})$ [9]. Therefore, by setting ε to the $ERROR$ of row r that has the smallest $pcc(r, T, \mathcal{Y})$ above threshold ρ'_c (Line 13), we prevent the algorithm from returning imbalanced biclusters (i.e., very small or very high number of columns).

3.2 Filtering Biclusters Found by Random Chance

Any dataset contains small biclusters with a high Pearson correlation value by random chance. Although we specify a lower bound for PCC ρ' and minimum number of columns γ , especially when γ is small, in addition to larger biclusters, CPB recovers such small biclusters. To eliminate randomly found biclusters in a non-parametric fashion, we developed following method. Suppose $\mathbb{B} = \{B_1, B_2, \dots, B_z\}$ be the set of biclusters found by different runs of CPB on a data matrix \mathbf{A} . We first generate \mathbf{A}' by shuffling the elements of \mathbf{A} . Then, we find the bicluster B_{max} with the highest number of rows in \mathbf{A}' , and use its dimension n' as a threshold to filter biclusters in \mathbb{B} . Algorithm 2 summarizes the filtering process. Note that the parameter n' is unique for each dataset, but this method empirically finds a lower bound for n' . The more biclusters generated from the shuffled dataset, the better the estimate of n' .

Algorithm 2 Filter Random Biclusters.

```

1: function FILTERRANDOMBICLUSTERS( $\mathbf{A}, \mathbb{B}, \gamma, \rho'$ )
2:    $\mathbf{A}' \leftarrow \text{SHUFFLE}(\mathbf{A}); \mathbb{B}' \leftarrow \{\}$ 
3:   for each row  $r_i$  in  $\mathbf{A}'$  do
4:      $\mathbb{B}' \leftarrow \mathbb{B}' \cup \text{CPB}(\mathbf{A}', r_i, 0.5, \gamma, \rho')$ 
5:   end for
6:    $n' \leftarrow \text{argmax}_{B_i \in \mathbb{B}'} n_i$ 
7:   for each bicluster  $B_i$  in  $\mathbb{B}$  do
8:      $(n_i, m_i) \leftarrow \text{size}[B_i]$ 
9:     if  $n_i \leq n'$  then
10:       $\mathbb{B} \leftarrow \mathbb{B} \setminus \{B_i\}$ 
11:     end if
12:   end for
13:   return  $\mathbb{B}$ 
14: end function

```

In addition to filtering out the statistically insignificant biclusters, we also remove those that have substantial overlaps. For any bicluster pair that has an overlap of 75 % or more, we remove the smaller bicluster.

3.3 Combining Correlation Information

CPB often produces different resulting biclusters due to the random selection of initial biclusters. Information from these biclusters, each including the reference row r_r , is merged to score each row's relationship with r_r .

In [9], *bicluster uniqueness* (BU) measure was proposed to calculate the correlation score of the genes. Although BU is able to capture the information redundancy caused by overlapping biclusters, we present a similar but more efficient scoring function to be used instead.

Let $\mathbb{B} = \{B_1, B_2, \dots, B_z\}$ be the set of biclusters found by different runs of CPB on a data matrix \mathbf{A} , and with reference row r_r . Suppose $IR(r)$ and $IC(c)$ denote the maximal subset of \mathbb{B} that contain the given row and column, respectively.

Definition 2 (Correlation Score (CS)). A score is assigned to a row r based on the number of experiments in which r is co-regulated with r_r by:

$$CS(r) = \sum_{c \in C} |IR(r) \cap IC(c)|. \quad (4)$$

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.

Definition 3 (Gene Correlation Score (GCS)). Let $\mathbb{A} = \{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_p\}$ be the set of (microarray) datasets with the same row labels (genes). Given a reference row r_r and datasets \mathbb{A} , gene correlation score GCS of a row r is calculated with

$$GCS(r, r_r) = \sum_{\mathbf{A}_p \in \mathbb{A}} \frac{CS(r)}{CS(r_r)}. \quad (5)$$

4 Experimental Results

We test CPB on the probes of three tumor suppressor genes (i.e., BRCA1, BRCA2, and p53) as queries to reveal co-regulated genes involved in the complex process of tumor formation. Experiments on 40 large datasets, each with 22,283 probe sets, show that the results are remarkably enriched for genes that have a role in cancer progression, tumor growth and metastasis regulation, and DNA degradation and repair. We also compare our results with another query-based framework to see how successfully each method finds *known* and *unexplored* genes co-regulated with tumor suppressors. While the ratios of *known* cancer-related genes are similar, the proposed framework finds more unexplored genes that are likely to be missed by earlier clustering-based methods. CPB's performance for querying the microarray data is promising: it was able to capture many genes that are highly correlated with BRCA1, BRCA2, and p53. We observed that of those genes, half are already known to be involved in cancer processes, and the others are promising new candidates for further investigation.

We first define some evaluation metrics and test CPB on synthetic datasets generated with biclusters with (1) different models, (2) increasing noise levels, and (3) increasing overlaps between embedded biclusters. We selected four other biclustering algorithms; δ -biclusters [11], OPSM [1], BBC [17], and BCCA [57], for comparison, due to their success at capturing shift-scale biclusters.

We test CPB on a number of human microarray datasets using four probes of breast cancer associated BRCA1 and BRCA2 genes,

and two probes of p53 tumor suppressor gene as queries. The correlation scores of the genes are combined, and the top-ranked genes are further studied. We also compare our results with MEM framework [8] in terms of the algorithms' effectiveness of retrieving *undiscovered* cancer-related genes.

4.1 Experiments on Synthetic Datasets

We first define recovery and relevance metrics to evaluate the results of biclustering algorithms. For each experiment, a synthetic dataset is generated with 1000 rows and 200 samples. Then two 60×60 biclusters with the given model are embedded into the dataset. The average score of 100 replication of the same experiment is reported.

4.1.1 Evaluation Metrics

Similar to recall and precision metrics, recovery and relevance scores are proposed to evaluate the biclustering results. These measures can be defined to compare a single found bicluster against an expected one, as well as a set of found biclusters against a set of expected ones.

Let e and f be expected and found biclusters, respectively. The recovery score of a found bicluster against an expected one is calculated by dividing the intersection area by the area of the expected bicluster:

$$rec(e, f) = \frac{|e \cap f|}{|e|}, \tag{6}$$

where the recovery score reaches to 1 if and only if $e \subseteq f$.

Similarly, relevance score is calculated by dividing the intersection area by the area of the found bicluster:

$$rel(e, f) = \frac{|e \cap f|}{|f|}, \tag{7}$$

where the relevance score reaches to 1 if and only if $f \subseteq e$. Examples of how these scores are computed are given in Fig. 3.

Using these rec and rel measures, we define recovery and relevance scores to compare two sets. Let E and F be a set of expected

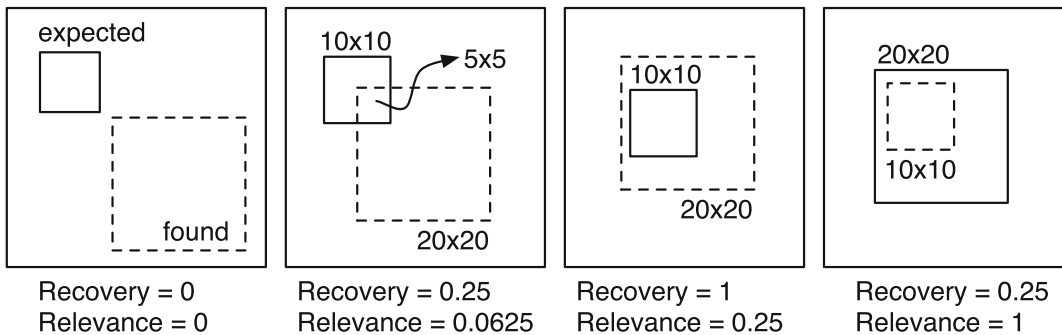


Fig. 3 Example expected/found biclusters with their recovery and relevance scores

and found biclusters, respectively. The set-based recovery score is calculated by taking the mean of the maximum recovery score for each expected bicluster. An equivalent approach is used for relevance.

$$REC(E, F) = \frac{1}{|E|} \sum_{e \in E} \max_{f \in F} rec(e, f) \quad (8)$$

$$REL(E, F) = \frac{1}{|F|} \sum_{f \in F} \max_{e \in E} rel(e, f) \quad (9)$$

4.1.2 Effects of the Bicluster Model

Biclustering methods often focus on detecting specific types of biclusters, as mentioned in Background section. In this experiment, we compare the success rate of CPB with other algorithms on detecting biclusters generated with various models. Constant-row, shift, scale, and shift-scale models were chosen for this experiment. Examples of these models are given in Fig. 1.

The resulting recovery and relevance scores (*see* Fig. 4a–d) show that CPB is the only algorithm that can fully recover biclusters generated with all four models with a high relevance score. BBC was able to find shifted and constant-row biclusters with a slightly lower relevance score. BCCA was expected to display similar results to CPB since they both use Pearson correlation; however, it was only able to fully recover shift biclusters. OPSM could not identify shift, scale, or shift-scale biclusters although they are all valid order-preserving submatrices. Our experiments show that when a base row is scaled with a value between -1 and 1 , the expression rankings of the columns of a bicluster row lie in a narrow range along the row; therefore, OPSM fails to discover it. Despite this limitation, OPSM was able to identify one of the shifted biclusters, since it reports a single bicluster for each size of column. The δ -biclusters algorithm performed poorly on all the datasets, among which it can partially recover only constant-row biclusters. The other models are not captured by the metric, which was previously discussed in [53].

For the noise and overlap experiments, CPB and other methods were compared on shift biclusters since the shift model is successfully recovered by most of the algorithms (*see* Fig. 4b).

4.1.3 Effects of the Noise

Microarrays results are perturbed by many sources of noise. In order to measure the sensitivity of CPB to noise, an error value ϵ was added to each element of the synthetic datasets. The experiments were run with various noise levels: each error value was drawn from a normal distribution with zero mean and variance equal to the chosen noise level.

Figure 4e shows the recovery and relevance scores of the algorithms on datasets with varying noise levels. OPSM is dramatically affected by noise since it may violate the order-preserving structure.

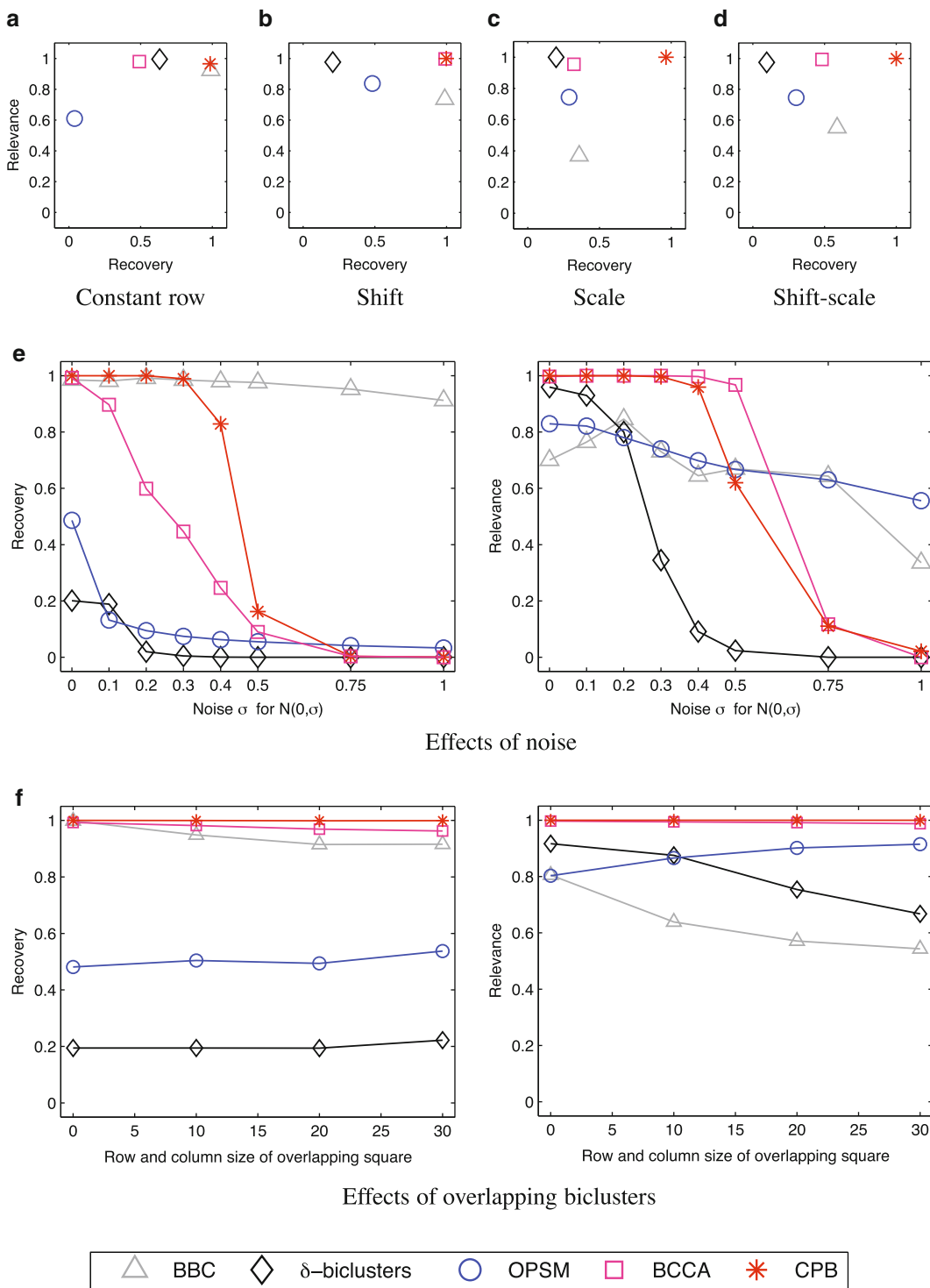


Fig. 4 Experiments on synthetic datasets: (a–d) with different bicluster models, (e) under noise, and (g) with overlapping biclusters

Since BCCA checks for pairwise correlation score for each row, this method is more likely to be affected by the increasing noise. Although BBC seems to be insensitive to noise in recovery plot, its relevance score drops slightly with noise addition (*see* Fig. 4e). CPB is the second best algorithm that is resistant to noise even though it has a linear metric. Moreover, the noise resistance of CPB can be improved by adjusting a better PCC threshold. We fixed $\rho = 0.9$ in order to be consistent with the rest of the experiments.

We also experimented with relative noise, in which the noise is added to each element with respect to its expression value, i.e., element x becomes $x + x \epsilon$. We observed results similar to the previous experiment.

4.1.4 Effects of the Overlap

A gene may take roles in several functions in a cell, each of which may be occurring simultaneously in a given sample; therefore, there might be overlaps between biclusters. In this experiment we test how CPB and other algorithms perform with increasing overlaps of biclusters. The datasets are generated with two overlapping biclusters. The overlapping regions of these biclusters are increased by 10 rows and 10 columns at each step. The expression values in these regions are not assumed to be additive; instead, shift values for rows and base vector are chosen in a way to allow both of the biclusters to have the same expression value at overlapping regions.

Figure 4f shows the results of the overlap test. We observe that CPB and BCCA are both insensitive to increasing overlap, while BCCA fails to recover a very small portion of the biclusters. BBC is affected more than BCCA in terms of recovery; also, its relevance score drops with increasing overlap. Although OPSM recovers only one of the biclusters, it increases its recovery score by including more of the overlapping region with increasing overlap.

4.2 Identifying Genes Co-regulated with BRCA1, BRCA2, p53

In this experiment, we employ CPB to identify the most correlated genes with BRCA1, BRCA2, and p53, which are highly penetrant cancer specific tumor suppressors. CPB was run on 40 different datasets obtained from the GPL96 series (GDS{1064, 1284, 1615, 2113, 2362, 2649, 2954, 3116, 3312, 3716, 1067, 1329, 1815, 2190, 2373, 2736, 3057, 3128, 3471, 534, 1209, 1375, 1956, 2255, 2519, 2767, 3096, 3233, 3514, 596, 1220, 1479, 1975, 2297, 2643, 2771, 3097, 3257, 3517, 987}), all of which have the same set of probes. The results of each dataset are then combined with *Gene Correlation Score* function.

Table 1 gives the top-ranked genes for the probes of BRCA1, BRCA2, and p53. We observe that more than 50 % of the genes found by our framework are already investigated in cancer research, suggesting that CPB is indeed finding genes involved with cancer.

Table 1
Associated top-ranked genes for 6 probes of BRCA1, BRCA2, and p53

	BRCA1		BRCA2		p53	
	204531_s_at	211851_x_at	208368_s_at	214727_at	201746_at	211300_s_at
1	C1orf105	H49077	CHRNA4	C1orf105	C1orf105	C1orf105
2	GPR98	C1orf105	C1orf105	ACRV1	H49077	PCNXL2
3	H49077	ARID4B	MTMR8	GFRA4	GFRA4	GFRA4
4	CHRNA4	MTMR8	ACRV1	MTMR8	ARID4B	U88898
5	SLC17A1	MBD2	GFRA4	AGTR2	MTMR8	H49077
6	GPX5	AK022006	H49077	PRO2958	AK022006	CHRNA4
7	PCNXL2	SLC17A1	PCNXL2	H49077	UBQLN3	ACRV1
8	MTMR8	CSRP3	PRO2958	ACRV1	ACRV1	MTMR8
9	ARID4B	PRO2958	U88898	U88898	GNRHR	PPP3CC
10	GFRA4	GFRA4	SLC17A1	SLC17A1	CHRNA4	NKX3-1
11	MBD2	NOS1	NKX3-1	GNRHR	GPR98	GNRHR
12	IL17A	PPP3CC	ACRV1	PPP3CC	RNF185	ACRV1
13	AK022006	CHRN3	PPP3CC	MBD2	SLC17A1	MBD2
14	NOS1	MAPK11	AGTR2	SPINLW1	KLK10	GNPTAB
15	ALPI	ACRV1	FAM55D	GNPTAB	U05589	CHRN3
16	KLK10	GNPTAB	SNX1	AK000787	GPX5	ACRV1
17	PRO2958	IL17A	TREX2	AK023690	OR7C2	AK000787
18	CSRP3	NEK1	GPR98	OR5I1	P2RY4	GPR98
19	PPP1R3A	U05589	LEP	BTNL8	PPP3CC	SPINLW1
20	ACRV1	IFNA5	M78162	EPAG	PCNXL2	AL162044
21	GNPTAB	GPX5	CSRP3	C22orf33	AK023690	ARID4B
22	AL117549	GNA11	KRT38	RPS6KA6	TREX2	ACRV1
23	NEK1	KLK10	TBR1	GPR98	PRO2958	AW139195
24	AK023690	GJB3	EPAG	MYL1	RBM2Y2FP	GPX5
25	MYL1	SLC7A11	MBD2	PCNXL2	IL17A	AK022006

Genes that have cancer-related studies found in PubMed are highlighted

Among those, for example, MBD2 is shown to have a role in cancer progression and can be therapeutically targeted in aggressive breast cancers [59]. KLK10 provides important prognostic information in early breast cancer patients [60]. CHRNA4 polymorphisms are found to activate factors that participate in DNA

degradation and repair, specifically the level of p53 participating in DNA repair [61].

Some genes, highly correlated with p53 in the list, are also investigated in other cancer types. For instance, 4 messenger RNA biomarkers, including ACRV1, may differentiate pancreatic cancer patients from noncancer subjects [62]. GFRA4 is predominantly expressed in normal and malignant thyroid medullary cells [63]. Of those, which are also correlated with BRCA1 and BRCA2 genes, can be further studied to see if they are up- or down-regulated in cancer patients.

The genes that have not already appeared in the literature may play an as-yet unknown role in cancer-related processes. These genes are open for further research.

4.3 Comparison with Other Query-Based Frameworks

There are a limited number of studies on query-based discovery of co-regulated genes in the literature. Gene recommender [5] analyzed Rb protein complex to find new co-regulated genes in worms (specifically *C. elegans*) using a technique similar to biclustering. SPELL [6], a PCC-based clustering framework was tested on *S. cerevisiae* datasets, where several genes were categorized and annotated. A probabilistic biclustering framework, QDB [7], was tested on some synthetic and yeast microarray datasets. Adler et al. [8] propose a query engine (MEM) to search for correlated genes across many datasets. Zhao et al. proposed ProBic [10], a probabilistic biclustering algorithm, and tested on *E. coli* to detect high quality biclusters in the presence of noise.

Among those query-driven search methods, we could compare our framework on cancer-related genes with only MEM [8], because the other studies are either specialized in non-human organisms, or resource is not accessible. Using MEM framework, we retrieved the genes correlated with the selected probes of BRCA1, BRCA2, and p53 genes with Pearson correlation. Top-ranked genes are then investigated to find whether the gene is claimed to be cancer-related in a research study in medical literature.

In Table 2 we compare our results with MEM's based on how successful each method is on finding *known* and *unexplored* genes. Since all samples (columns) of a microarray dataset were included in similarity calculations before biclustering, the top-ranked genes discovered by MEM are expected to be investigated before. While the ratios of *known* cancer-related genes are similar, we argue that our framework finds more unexplored genes that are likely to be missed by earlier clustering-based methods.

Although PCC is the default similarity measure, we also run MEM with absolute Pearson correlation, which is expected to capture negative correlations as in our *pcc* function (see Eq. (1)). However, the results on six probes between two runs of MEM framework are 87 % overlapping. Absolute PCC could only

Table 2

Ratio of known and unexplored genes found within top-25 results of MEM (with Pearson and absolute Pearson correlation) and our framework

		BRCA1		BRCA2		p53		Average (%)
		204531 _s_at (%)	211851 _x_at (%)	208368 _s_at (%)	214727 _at (%)	201746 _at (%)	211300 _s_at (%)	
MEM Pearson	Known	56	64	64	64	68	68	64.0
	New	32	16	28	32	24	24	26.0
	Duplicate	12	20	8	4	8	8	10.0
MEM Pearson	Known	64	64	72	68	76	72	69.3
	New	32	16	24	32	20	20	24.0
	Duplicate	4	20	4	0	4	8	6.6
CPB	Known	60	64	52	32	48	48	50.6
	New	40	36	44	64	52	40	46.0
	Duplicate	0	0	4	4	0	12	3.3

Only the unique gene names are considered. Probes of the same gene after the first one are counted as redundant (duplicate) information

introduce 20 new genes, where 60 % of them are already found to be cancer-related. We conclude that altering the similarity measure (in this case, taking the absolute value to capture negative correlations) is not as effective for finding undiscovered correlated genes as applying biclustering.

5 Conclusion

In this chapter, we briefly survey the biclustering algorithms in the literature and introduce a method for querying co-regulated genes using a novel biclustering method, the CPB. Initial testing on artificial data confirms that CPB is capable of finding such biclusters and that it outperforms other biclustering methods in finding multiple types of biclusters. CPB's performance for querying the microarray data is promising: it finds many genes that have high correlations with BRCA1, BRCA2, and p53. Of those genes, half are already known to be involved in cancer processes, and the others are promising new candidates for further investigation.

There are many possible extensions to CPB that may yet be explored. For instance, PCC is only one of the well-known metrics for evaluating similarity. CPB approach may be extended to use other metrics and benefit from their unique properties. CPB's iterative optimization process may likewise be improved by

choosing initial biclusters differently or using a mathematical optimization method to avoid the local maximas.

6 Acknowledgments

This work was supported in parts by National Institutes of Health/ National Cancer Institute (grant number R01CA141090); by the Department of Energy (grant number DE-FC02-06ER2775); and by the National Science Foundation (grants numbers CNS-0643969, OCI-0904809, OCI-0904802).

References

1. Ben-Dor A, Chor B, Karp R, Yakhini Z (2002) Discovering local structure in gene expression data: The order-preserving submatrix problem. In: Proceedings of the International Conference on Computational Biology, pp 49–57
2. Jiang D, Pei J, Zhang A (2003) DHC: a density-based hierarchical clustering method for time series gene expression data. In: Proceedings IEEE Symposium on BioInformatics and Bioengineering, pp 393–400
3. Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform* 1 (1):24–45
4. Pujana MA, Han J-DJ, LM Starita, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, Assmann V, ElShamy WM, Rual J-F, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, Sole X, Hernandez P, Lazaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, Livingston DM, Gruber SB, Parvin JD, Vidal M (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet* 39(11):1338–1349
5. Owen AB, Stuart J, Mach K, Villeneuve AM, Kim S (2003) A gene recommender algorithm to identify coexpressed genes in *C. elegans*. *Genome Res* 13(8):1828–1837
6. Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* 23:2692–2699
7. Dhollander T, Sheng Q, Lemmens K, De Moor B, Marchal K, Moreau Y (2007) Query-driven module discovery in microarray data. *Bioinformatics* 23:2573–2580
8. Adler P, Kolde R, Kull M, Tkachenko A, Peterson H, Reimand J, Vilo J (2009) Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biol* 10:R139
9. Bozdağ D, Parvin JD, Çatalyürek ÜV (2009) A biclustering method to discover co-regulated genes using diverse gene expression datasets. In: Proceedings of 1st International Conference on Bioinformatics and Computational Biology, pp 151–163
10. Zhao H, Cloots L, Van den Bulcke T, Wu Y, De Smet R, Storms V, Meysman P, Engelen K, Marchal K (2011) Query-based biclustering of gene expression data using probabilistic relational models. *BMC Bioinf* 12(Suppl 1):S37
11. Cheng Y, Church GM (2000) Biclustering of expression data. In: Proceedings of International Conference on Intelligent Systems for Molecular Biology, pp 93–103
12. Segal E, Taskar B, Gasch A, Friedman N, Koller D (2001) Rich probabilistic models for gene expression. *Bioinformatics* 17(suppl_1):S243–S252
13. Wang H, Wang W, Yang J, Yu PS (2002) Clustering by pattern similarity in large data sets. In: Proceedings of ACM SIGMOD
14. Lazzeroni L, Owen A (2000) Plaid models for gene expression data. Tech. Rep., Stanford University
15. Mitra S, Banka H (2006) Multi-objective evolutionary biclustering of gene expression data. *Pattern Recognit* 39(12):2464–2477
16. Mejía-Roa E, Carmona-Saez P, Nogales R, Vicente C, Vázquez M, Yang XY, García C, Tirado F, Pascual-Montano A (2008) bioNMF: a web-based tool for nonnegative matrix factorization in biology. *Nucleic Acids Res* 36(suppl 2):W523–W528

17. Gu J, Liu JS (2008) Bayesian biclustering of gene expression data. *BMC Genomics* 9(Suppl 1):S4
18. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, Kasim A, Khamiakova T, Van Sanden S, Lin D, Talloen W et al (2010) Fabia: factor analysis for bicluster acquisition. *Bioinformatics* 26(12):1520–1527
19. Painsky A, Rosset S (2012) Exclusive row biclustering for gene expression using a combinatorial auction approach. In: *Proceedings of the 2012 I.E. 12th International Conference on Data Mining*, pp 1056–1061. IEEE Computer Society
20. Joung J-G, Kim S-J, Shin S-Y, Zhang B-T (2012) A probabilistic coevolutionary biclustering algorithm for discovering coherent patterns in gene expression dataset. *BMC Bioinf* 13(Suppl 17):S12
21. Flores JL, Inza I, Larrañaga P, Calvo B (2013) A new measure for gene expression biclustering based on non-parametric correlation. *Comput Methods Prog Biomed* 112(3):367–397
22. Sun P, Speicher NK, Röttger R, Guo J, Baumbach J (2014) Bi-force: large-scale bicluster editing and its application to gene expression data biclustering. *Nucleic Acids Res.* doi:10.1093/nar/gku201
23. Chakraborty A (2005) Biclustering of gene expression data by simulated annealing. In: *Proceedings of Eighth International Conference on High-Performance Computing in Asia-Pacific Region, 2005*, pp 627–632
24. Liew AW-C, Law N-F, Yan H (2011) Recent patents on biclustering algorithms for gene expression data analysis. *Recent Pat DNA Gene Seq* 5(2):117–125
25. Hussain SF (2011) Bi-clustering gene expression data using co-similarity. In: *Proceedings of the 7th International Conference on Advanced Data Mining and Applications - Volume Part I, ADMA'11*, pp 190–200. Springer, Berlin/Heidelberg
26. An J, Liew AW-C, Nelson CC (2012) Seed-based biclustering of gene expression data. *PLoS ONE* 7:e42431, 08
27. Kiraly A, Abonyi J, Laiho A, Gyenesi A (2012) Biclustering of high-throughput gene expression data with bicluster miner. In: *IEEE 12th International Conference on Data Mining Workshops (ICDMW), 2012*, pp 131–138
28. Liu J, Wang J, Wang W (2004) Biclustering in gene expression data by tendency. In: *Proceedings of IEEE Computational Systems Bioinformatics Conference*, pp 182–193. IEEE Computer Society
29. Liu J, Wang J, Wang W (2004) Gene ontology friendly biclustering of expression profiles. In: *Proceedings of IEEE Computational Systems Bioinformatics Conference*, pp 436–447. IEEE Computer Society
30. Madeira S, Oliveira A (2005) A linear time biclustering algorithm for time series gene expression data. In: Casadio R, Myers G (eds) *Algorithms in bioinformatics. Lecture Notes in Computer Science*, vol 3692, pp 39–52, Springer, Berlin/Heidelberg
31. Pontes B, Giraldez R, Aguilar-Ruiz JS (2013) Configurable pattern-based evolutionary biclustering of gene expression data. *Algorithms Mol Biol* 8:4
32. Yang W-H, Dai D-Q, Yan H (2011) Finding correlated biclusters from gene expression data. *IEEE Trans Knowl Data Eng* 23:568–584
33. Yoon S, Nardini C, Benini L, De Micheli G (2005) Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams. *IEEE/ACM Trans Comput Biol Bioinf* 2:339–354
34. Angiulli F, Cesario E, Pizzuti C (2008) Random walk biclustering for microarray data. *Inf Sci* 178(6):1479–1497
35. Bryan K (2005) Biclustering of expression data using simulated annealing. In: *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems, CBMS'05*, (Washington, DC, USA), pp 383–388. IEEE Computer Society
36. Bryan K, Cunningham P, Bolshakova N (2006) Application of simulated annealing to the biclustering of gene expression data. *Trans Inf Tech Biomed* 10:519–525
37. Bleuler S, Prelic A, Zitzler E (2004) An EA framework for biclustering of gene expression data. In: *Congress on Evolutionary Computation, 2004 (CEC2004)*, vol 1, pp 166–173
38. Divina F, Aguilar-Ruiz J (2006) Biclustering of expression data with evolutionary computation. *IEEE Trans Knowl Data Eng* 18:590–602
39. Nepomuceno JA, Troncoso A, Aguilar-Ruiz JS (2010) Correlation-based scatter search for discovering biclusters from gene expression data. In: *Proceedings of the 8th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO'10*, pp 122–133. Springer, Berlin/Heidelberg
40. Nepomuceno JA, Troncoso A, Aguilar-Ruiz JS (2011) A comparative analysis of biclustering algorithms for gene expression data. *BioData Mining* 4:3

41. Erten C, Sözdinler M (2009) Biclustering expression data based on expanding localized substructures. In: Rajasekaran S (ed) *Bioinformatics and computational biology. Lecture Notes in Computer Science*, vol 5462, pp 224–235. Springer, Berlin/Heidelberg
42. Tanay A, Sharan R, Shamir R (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 18(Supplement 1):136–144
43. Bergmann S, Ihmels J, Barkai N (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys Rev E Stat Nonlinear Soft Matter Phys* 67:031902
44. Kluger Y, Basri R, Chang JT, Gerstein M (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res* 13(4):703–716
45. Prelic A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 22:1122–1129
46. Li G, Ma Q, Tang H, Paterson AH, Xu Y (2009) QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res* 37(15):e101
47. Huttenhower C, Mutungu KT, Indik N, Yang W, Schroeder M, Forman JJ, Troyanskaya OG, Collier HA (2009) Detailing regulatory networks through large scale data integration. *Bioinformatics* 25:3267–3274
48. Voggenreiter O, Bleuler S, Gruissem W (2012) Exact biclustering algorithm for the analysis of large gene expression data sets. *BMC Bioinf* 13 (Suppl 18):A10
49. Bryan K, Cunningham P (2006) Bottom-up biclustering of expression data. In: *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, 2006 (CIBCB '06), pp 1–8
50. Murali T, Kasif S (2003) Extracting conserved gene expression motifs from gene expression data. *Pac Symp Biocomput* 8:77–88
51. Liu J, Wang W (2003) Op-cluster: clustering by tendency in high dimensional space. In: *Proceedings of IEEE International Conference on Data Mining*, p 187
52. Freitas AV, Ayadi W, Elloumi M, Oliveira J, Oliveira J, Hao J-K (2013) Survey on biclustering of gene expression data, pp 591–608. Wiley, New York
53. Bozdağ D, Kumar A, Çatalyürek ÜV (2010) Comparative Analysis of Biclustering Algorithms. In: *ACM International Conference on Bioinformatics and Computational Biology*
54. Chia BKH, Karuturi RKM (2010) Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. *Algorithms Mol Biol* 5(1):8
55. Eren K, Deveci M, Küçüktunç O, Çatalyürek ÜV (2012) A comparative analysis of biclustering algorithms for gene expression data. *Brief Bioinform*
56. Oghabian A, Kilpinen S, Hautaniemi S, Czeizler E (2014) Biclustering methods: Biological relevance and application in gene expression analysis. *PLoS one* 9(3):e90801
57. Bhattacharya A, De RK (2009) Bi-correlation clustering algorithm for determining a set of co-regulated genes. *Bioinformatics* 25 (21):2795–2801
58. Casella G, Wells MT (1993) Is Pitman closeness a reasonable criterion: comment. *J Am Stat Assoc* 88(421):70–71
59. Mian O, Wang S, Zhu S, Gnanapragasam M, Graham L, Bear H, Ginder G (2011) Methyl-binding domain protein 2-dependent proliferation and survival of breast cancer cells. *Mol Cancer Res* 9(8):1152–62
60. Kioulafa M, Kaklamanis L, Stathopoulos E, Mavroudis D, Georgoulas V, Lianidou ES (2009) Kallikrein 10 (KLK10) methylation as a novel prognostic biomarker in early breast cancer. *Ann Oncol* 20:1020–1025
61. Dorszewska J, Florczak J, Rozycka A, Jaroszewska-Kolecka J, Trzeciak WH, Kozubski W (2005) Polymorphisms of the CHRNA4 gene encoding the alpha4 subunit of nicotinic acetylcholine receptor as related to the oxidative DNA damage and the level of apoptotic proteins in lymphocytes of the patients with Alzheimer's disease. *DNA Cell Biol* 24:786–794
62. Zhang L, Farrell JJ, Zhou H, Elashoff D, Akin D, Park N-H, Chia D, Wong DT (2010) Salivary transcriptomic biomarkers for detection of resectable pancreatic cancer. *Gastroenterology* 138(3):949–957, e1–7
63. Lindahl M, Poteryaev D, Yu L, Arumae U, Timmusk T, Bongarzone I, Aiello A, Pierotti MA, Airaksinen MS, Saarma M (2001) Human glial cell line-derived neurotrophic factor receptor alpha 4 is the receptor for persephin and is predominantly expressed in normal and malignant thyroid medullary cells. *J Biol Chem* 276:9344–9351