Journal of Bioinformatics and Computational Biology
Vol. 7, No. 5 (2009) 853–868

A NOVEL COHERENCE MEASURE FOR DISCOVERING SCALING BICLUSTERS FROM GENE EXPRESSION DATA

ANIRBAN MUKHOPADHYAY
Department of Computer Science and Engineering
University of Kalyani, Kalyani-741235
West Bengal, India
anirban@klyuniv.ac.in

UJJWAL MAULIK
Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032
West Bengal, India
drumaulik@cse.jdvu.ac.in

SANGHAMITRA BANDYOPADHYAY
Machine Intelligence Unit, Indian Statistical Institute
Kolkata-700108, West Bengal, India
sanghami@isical.ac.in

Received 9 February 2009
Revised 17 April 2009
Accepted 26 May 2009

Biclustering methods are used to identify a subset of genes that are co-regulated in a subset of experimental conditions in microarray gene expression data. Many biclustering algorithms rely on optimizing mean squared residue to discover biclusters from a gene expression dataset. Recently it has been proved that mean squared residue is only good in capturing constant and shifting biclusters. However, scaling biclusters cannot be detected using this metric. In this article, a new coherence measure called scaling mean squared residue (SMSR) is proposed. Theoretically it has been proved that the proposed new measure is able to detect the scaling patterns effectively and it is invariant to local or global scaling of the input dataset. The effectiveness of the proposed coherence measure in detecting scaling patterns has been demonstrated experimentally on artificial and real-life benchmark gene expression datasets. Moreover, biological significance tests have been conducted to show that the biclusters identified using the proposed measure are composed of functionally enriched sets of genes.

Keywords: Biclustering; scaling pattern; mean squared residue; scaling mean squared residue.

1. Introduction

Advancement of microarray technology has made it possible to monitor the expression patterns of a huge number of genes in parallel across several experimental conditions. An important computational task in microarray datasets is discovering similarly expressed genes which are expected to be functionally related. Clustering[1] has been widely used in microarrays for the purpose of discovering genes that are co-expressed across all the conditions.[2,3] However, it has been seen that a set of genes can have similar expression profile only for a subset of conditions. Unlike clustering, biclustering algorithms aim to discover a subset of genes that are co-regulated in a subset of conditions. Hence biclustering can be thought as simultaneous clustering from both the dimensions. Biologically, biclusters are more relevant compared to clusters.

In recent years, several studies have been made by researchers in the context of biclustering of microarray data. One of the earlier works on biclustering in the context of microarray data can be found in Ref. 4, where *mean squared residue* (*MSR*) measure was used to compute the coherence among a group of genes. The algorithm developed in Ref. 4 was based on a greedy search technique guided by a heuristic. In Ref. 5, a coupled two-way clustering (CTWC) method has been proposed. An improved version of Cheng and Church's algorithm, called Flexible Overlapped biclustering (FLOC) is proposed in Ref. 6 which deals with the missing values. In Ref. 7, a genetic algorithm (GA)–based biclustering algorithm has been presented that uses mean squared residue as a fitness function to be minimized. A multiobjective fuzzy biclustering method has been proposed in Ref. 8, where three criteria, namely the fuzzy mean squared residue, fuzzy row variance and fuzzy volume are optimized simultaneously. A bipartite graph–based model called Statistical-Algorithmic Method for Bicluster Analysis (SAMBA) has been proposed for biclustering in Ref. 9. In Ref. 10, a simulated annealing–based biclustering technique is presented that minimizes the *MSR* measure.

MSR is a very popular measure and a number of well-known biclustering algorithms have been developed that are based on minimizing *MSR*.[6,7,10] However, a recent study in Ref. 11 shows that *MSR*, which is effective in detecting constant and shifting biclusters, is affected by scaling factors and thus cannot be used to discover biclusters with scaling patterns. In order to overcome this limitation, in this article, a new coherence measure called *scaling mean squared residue* (*SMSR*) is proposed to detect scaling biclusters. The effectiveness of the proposed measure has been established theoretically and through experimentation on both artificial and real-life benchmark gene expression datasets. Finally biological significance tests have been conducted to establish that the scaling biclusters discovered by *SMSR*-based algorithm are composed of functionally enriched sets of genes.

2. Bicluster Models

A microarray dataset can be considered as a $\mathcal{G} \times \mathcal{C}$ matrix $\mathcal{A}$ that represents the expression level of a set of $\mathcal{G}$ genes $G = \{I_1, I_2, \ldots, I_{\mathcal{G}}\}$ over a set of $\mathcal{C}$ conditions

$C = \{J_1, J_2, \ldots, J_{\mathcal{C}}\}$. Each element m_{ij} of matrix $\mathcal{A}$ represents the expression level of the ith gene at the jth condition, where $i \in G$ and $j \in C$.

Definition 1. (Bicluster) A bicluster is a submatrix $\mathcal{M}(I, J) = [m_{ij}], i \in I, j \in J$, of matrix $\mathcal{A}$, where $I \subseteq G$ and $J \subseteq C$.

A bicluster is a submatrix of the whole microarray representing a subset of genes that are similarly expressed over a subset of conditions and vice versa.

2.1. *Types of biclusters*

There are different types of biclusters which are defined as follows[12]:

Definition 2. (Constant Biclusters) A bicluster $\mathcal{M}(I, J) = [m_{ij}], i \in I, j \in J$, is called a constant bicluster if all the elements have a constant value $m_{ij} = \pi$.

Definition 3. (Row Constant Biclusters) A bicluster $\mathcal{M}(I, J) = [m_{ij}], i \in I, j \in J$, is called a row constant bicluster if all the elements of each row of the bicluster have the same value. Hence in a row constant bicluster, each element can be represented using one of the following notations: $m_{ij} = \pi + a_i$ or $m_{ij} = \pi b_i$. Here π is a constant value for a bicluster, a_i is the shifting factor for row i and b_i is the scaling factor for row i.

Definition 4. (Column Constant Biclusters) A bicluster $\mathcal{M}(I, J) = [m_{ij}], i \in I, j \in J$, is called a column constant bicluster if all the elements of each column of the bicluster have the same value. Hence in a column constant bicluster, each element can be represented using one of the following notations: $m_{ij} = \pi + p_j$ or $m_{ij} = \pi q_j$. Here π is a constant value for a bicluster, p_j is the shifting factor for column j and q_j is the scaling factor for column j.

Definition 5. (Perfect Shifting Biclusters) A bicluster $\mathcal{M}(I, J) = [m_{ij}], i \in I, j \in J$, is called a perfect shifting bicluster if each column and row has only some shifting factors. Hence in a perfect shifting bicluster, each element $m_{ij} = \pi + a_i + p_j$.

Definition 6. (Perfect Scaling Biclusters) A bicluster $\mathcal{M}(I, J) = [m_{ij}], i \in I, j \in J$, is called a perfect scaling bicluster if each column and row has only some scaling factors. Hence in a perfect scaling bicluster, each element $m_{ij} = \pi b_i q_j$.

3. Mean Squared Residue

Cheng and Church[4] defined a bicluster as a subset of rows and a subset of columns with a high similarity score. They termed the similarity score as *mean squared residue* (*MSR*), $\mathcal{H}$, which measures the coherence of the rows and columns in the bicluster. In particular, they aim at finding large and maximal biclusters with $\mathcal{H}$ scores below a certain threshold δ (called as δ-bicluster). In a perfect δ-bicluster $\mathcal{M}(I, J) = [m_{ij}]$, each row/column or both rows and columns exhibit an absolutely consistent bias ($\mathcal{H} = \delta = 0$). Thus the *MSR* score becomes zero for a perfect shifting

bicluster, each element of which is of the form: $m_{ij} = \pi + a_i + p_j$. Let us define the row mean of the ith row of $\mathcal{M}(I, J)$ as: $m_{iJ} = \frac{1}{|J|}\sum_{j\in J} m_{ij}$, the column mean of the jth column as: $m_{Ij} = \frac{1}{|I|}\sum_{i\in I} m_{ij}$, and the mean of all the elements of the bicluster as: $m_{IJ} = \frac{1}{|I|\times|J|}\sum_{i\in I, j\in J} m_{ij}$. Now the constant value for the bicluster can be taken as $\pi = m_{IJ}$, the shifting factor for the ith row can be defined as the difference $a_i = m_{iJ} - m_{IJ}$, and the shifting factor for the jth column can be defined as the difference $p_j = m_{Ij} - m_{IJ}$. Therefore each element m_{ij} of a perfect shifting bicluster can be uniquely defined as:

$$m_{ij} = \pi + a_i + p_j = m_{IJ} + (m_{iJ} - m_{IJ}) + (m_{Ij} - m_{IJ}) = m_{iJ} + m_{Ij} - m_{IJ}. \tag{1}$$

Due to the presence of noise in the microarray data, it is almost impossible to find a perfect shifting δ-bicluster of the above form. Hence the concept of residue is introduced to quantify the difference between the actual value of an element m_{ij} and its expected value as found by Eq. (1). Thus the residue r_{ij} of any element m_{ij} of the bicluster is defined as:

$$r_{ij} = m_{ij} - (m_{iJ} + m_{Ij} - m_{IJ}) = m_{ij} - m_{iJ} - m_{Ij} + m_{IJ}. \tag{2}$$

In order to assess the overall quality of a δ-bicluster, the *mean squared residue* (*MSR*) of the bicluster is computed.

Definition 7. (Mean Squared Residue) The mean squared residue [$MSR(I, J)$] of a bicluster $\mathcal{M}(I, J)$ is defined as:

$$MSR(I, J) = \frac{1}{|I||J|}\sum_{i\in I, j\in J} r_{ij}^2 = \frac{1}{|I||J|}\sum_{i\in I, j\in J} (m_{ij} - m_{iJ} - m_{Ij} + m_{IJ})^2, \tag{3}$$

where $|I|$ and $|J|$ denote the number of rows and the number of columns in the bicluster, respectively.

The *MSR* of a bicluster represents the level of coherence among the elements of the bicluster. Lower residue score means larger coherence and thus better quality of the bicluster.

Note that the low residue biclusters should have a sufficient variation of the expression values in each row compared to the row mean value. This is required to avoid the trivial biclusters having almost all constant values. Hence the aim is to find large biclusters that have *MSR* below a threshold δ (δ-biclusters) and relatively high *row variance* which is defined as follows:

Definition 8. (Row Variance) The row variance $var(I, J)$ of a bicluster $\mathcal{M}(I, J)$ is defined as:

$$var(I, J) = \frac{1}{|I||J|}\sum_{i\in I, j\in J} (m_{ij} - m_{iJ})^2. \tag{4}$$

Many biclustering algorithms available in the literature are based on minimizing *MSR* measure. In Ref. 11, it has been proved that *MSR* is invariant to shifting but

is affected by scaling and hence can detect biclusters with shifting patterns only. Therefore the algorithms which rely on *MSR*, are unable to discover significant biclusters that have scaling patterns. This fact motivates us to make an attempt to devise a new residue measure that works for scaling biclusters.

4. Identifying Scaling Pattern: Scaling Mean Squared Residue

In this section, a new coherence measure called Scaling Mean Squared Residue (*SMSR*) that is able to detect biclusters with scaling patterns is developed. Subsequently it is proved that any perfect scaling bicluster will have *SMSR* equal to zero, and global or local scaling do not affect the *SMSR* score of a bicluster.

As discussed in Sec. 2.1, each element of a perfect scaling bicluster $\mathcal{M}(I, J)$ can be represented as $m_{ij} = \pi b_i q_j$, where π is the constant term for the bicluster, b_i is the scaling factor for row i and q_j is the scaling factor for column j.

Now following the derivation of *MSR* in the previous section, *SMSR* can be derived as follows: the constant term of the bicluster can be taken as the bicluster mean m_{IJ}, i.e. $\pi = m_{IJ}$. The scaling factor b_i for each row can be defined by the ratio of m_{iJ} to m_{IJ}, i.e. $b_i = \frac{m_{iJ}}{m_{IJ}}$, $m_{IJ} \neq 0$. Similarly the scaling factor q_j for each column can be defined by the ratio of m_{Ij} to m_{IJ}, i.e. $q_j = \frac{m_{Ij}}{m_{IJ}}$, $m_{IJ \neq 0}$. Therefore each element m_{ij} of a perfect scaling bicluster can be uniquely defined as:

$$m_{ij} = \pi b_i q_j = m_{IJ} \times \frac{m_{iJ}}{m_{IJ}} \times \frac{m_{Ij}}{m_{IJ}} = \frac{m_{iJ} \times m_{Ij}}{m_{IJ}}. \tag{5}$$

Hence for a perfect scaling bicluster, we have:

$$m_{ij} = \frac{m_{iJ} \times m_{Ij}}{m_{IJ}} \quad \text{or,} \quad \frac{m_{ij} \times m_{IJ}}{m_{iJ} \times m_{Ij}} = 1 \quad \text{or,} \quad 1 - \frac{m_{ij} \times m_{IJ}}{m_{iJ} \times m_{Ij}} = 0. \tag{6}$$

A bicluster which is not a perfect scaling one, will not have zero value for the above expression and the scaling residue s_{ij} of any element m_{ij} can now be defined as:

$$s_{ij} = 1 - \frac{m_{ij} \times m_{IJ}}{m_{iJ} \times m_{Ij}} = \frac{1}{m_{iJ}.m_{Ij}}(m_{iJ}.m_{Ij} - m_{ij}.m_{IJ}). \tag{7}$$

Hence we can define the overall scaling mean squared residue as follows:

Definition 9. (Scaling Mean Squared Residue) The Scaling Mean Squared Residue [$SMSR(I, J)$] of a bicluster $\mathcal{M}(I, J)$ is defined as:

$$SMSR(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} s_{ij}^2 = \frac{1}{|I||J|} \sum_{i \in I, j \in J} \frac{1}{m_{iJ}^2.m_{Ij}^2}(m_{iJ}.m_{Ij} - m_{ij}.m_{IJ})^2. \tag{8}$$

Note that Eqs. (5)–(8) hold only if $m_{iJ} \neq 0$, $m_{Ij} \neq 0$ and $m_{IJ} \neq 0$. Hence to avoid accidental *divide-by-zero* conditions, a very small value $\epsilon \approx 0$ can be added with these values whenever they are zero.

Like *MSR*, lower *SMSR* also indicates high coherence in the bicluster. The following theorems prove that a perfect scaling bicluster will have $SMSR = 0$ and *SMSR* is invariant to global or local scaling.

Theorem 1. *A perfect scaling bicluster has SMSR equal to zero.*

Proof. Each element m_{ij} of a perfect scaling bicluster $\mathcal{M}(I, J)$ can be expressed as: $m_{ij} = \pi b_i q_j$. Here, π is a constant value for the bicluster, b_i is the scaling factor of row i and q_j is the scaling factor of column j. Therefore, $\mathcal{M}$ is represented as:

$$\mathcal{M} = \begin{pmatrix} \pi b_1 q_1 & \cdots & \pi b_1 q_{|J|} \\ \pi b_2 q_1 & \cdots & \pi b_2 q_{|J|} \\ \vdots & \vdots & \vdots \\ \pi b_{|I|} q_1 & \cdots & \pi b_{|I|} q_{|J|} \end{pmatrix}. \tag{9}$$

Now the row means m_{iJ}, $i \in I$, the column means m_{Ij}, $j \in J$, and the bicluster mean m_{IJ} can be computed as follows:

$$m_{iJ} = \frac{1}{|J|} \sum_{j \in J} \pi b_i q_j = \pi b_i \frac{1}{|J|} \sum_{j \in J} q_j = \pi b_i \mu_q, \tag{10}$$

where $\mu_q = \frac{1}{|J|} \sum_{j \in J} q_j$, i.e. the mean of the column scaling factors. Similarly,

$$m_{Ij} = \frac{1}{|I|} \sum_{i \in I} \pi b_i q_j = \pi q_j \frac{1}{|I|} \sum_{i \in I} b_i = \pi q_j \mu_b, \tag{11}$$

where $\mu_b = \frac{1}{|I|} \sum_{i \in I} b_i$, i.e. the mean of the row scaling factors, and

$$m_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} \pi b_i q_j = \pi \left(\frac{1}{|I|} \sum_{i \in I} b_i \right) \left(\frac{1}{|J|} \sum_{j \in J} q_j \right) = \pi \mu_b \mu_q. \tag{12}$$

Now putting the values of m_{ij}, m_{iJ}, m_{Ij} and m_{IJ} in Eq. (7), we get

$$s_{ij} = \frac{1}{\pi b_i \mu_q \times \pi q_j \mu_b} (\pi b_i \mu_q \times \pi q_j \mu_b - \pi b_i q_j \times \pi \mu_b \mu_q) = 0. \tag{13}$$

Since scaling residue is zero for each element of the bicluster, the value of *SMSR* will be zero. □

Theorem 2. *Global or local scaling have no effect on SMSR.*

Proof. Let us first consider the effect of global scaling on *SMSR*. Suppose a global scaling factor α is multiplied with each element m_{ij}, $i \in I$, $j \in J$, of a bicluster. Hence $\forall i, j$, the row means m_{iJ}, $i \in I$, the column means m_{Ij}, $j \in J$ and the bicluster mean m_{IJ} are also multiplied by α. Therefore it is evident from Eq. (7) that the scaling residue value for each element m_{ij}, $i \in I$, $j \in J$, does not change. Hence the *SMSR* also does not change in the case of a global scaling.

Now let us consider a local scaling factor β_j for each column $j \in J$. If the β vector is applied to the bicluster, the new value for each element will be $m_{ij} \times \beta_j$, $i \in I$, $j \in J$. The new values for the row means will be $m_{iJ} \times \mu_\beta$, $i \in I$, where $\mu_\beta = \frac{1}{|J|} \sum_{j \in J} \beta_j$. The new values for the column means will be $m_{Ij} \times \beta_j$, $j \in J$.

The new value for the bicluster mean will be $m_{IJ} \times \mu_\beta$. Now putting these values in Eq. (7), it is found that the scaling residue for each element does not change. Hence the *SMSR* also does not change in the case of a local scaling on columns.

Next let us consider a local scaling factor γ_i for each row $i \in I$. If the γ vector is applied to the bicluster, the new value for each element will be $m_{ij} \times \gamma_i$, $i \in I$, $j \in J$. The new values for the row means will be $m_{iJ} \times \gamma_i$, $i \in I$. The new values for the column means will be $m_{Ij} \times \mu_\gamma$, $j \in J$, where $\mu_\gamma = \frac{1}{|I|} \sum_{i \in I} \gamma_i$. The new value for the bicluster mean will be $m_{IJ} \times \mu_\gamma$. Now putting these values in Eq. (7), it is found that the scaling residue for each element does not change. Hence the *SMSR* also does not change in the case of a local scaling on rows.

Finally we consider the local scaling factor β_j for each column $j \in J$ and the local scaling factor γ_i for each row $i \in I$ together. If the β and γ vectors are applied to the bicluster simultaneously, the new value for each element will be $m_{ij} \times \beta_j \times \gamma_i$, $i \in I$, $j \in J$. The new values for the row means will be $m_{iJ} \times \mu_\beta \times \gamma_i$, $i \in I$. The new values for the column means will be $m_{Ij} \times \beta_j \times \mu_\gamma$, $j \in J$. The new value for the bicluster mean will be $m_{IJ} \times \mu_\beta \times \mu_\gamma$. Now putting these values in Eq. (7), it is found that the scaling residue for each element does not change. Hence the *SMSR* also does not change in the case of a local scaling on columns and rows together.

Hence it is proved that global or local scaling have no effect on *SMSR*. □

5. Experiments and Results

Here, three sets of experiments are conducted. First, an artificial dataset consisting of implanted shifting and scaling patterns has been used to show the utility of *SMSR*. Thereafter, two benchmark real-life datasets, that is, Yeast Cell Cycle data[4] consisting of 2884 genes and 17 time points, and Human Large B-cell Lymphoma data[4] consisting of 4026 genes and 96 time points are used to demonstrate the effectiveness of *SMSR*-based biclustering. Both the real-life datasets are available at http://arep.med.harvard.edu/biclustering. Finally, we studied the biological significance of the biclusters obtained from the Yeast and Lymphoma datasets.

For comparison, we implemented three biclustering algorithms: one is exactly the same algorithm proposed by Cheng & Church[4] to search biclusters based on *MSR*. We call this algorithm CC(MSR). The second algorithm is modified CC method where the searching process is exactly the same as CC(MSR), but instead of *MSR*, we use *SMSR* as the filtering strategy. This algorithm is termed as CC(SMSR). The third algorithm is just a combination of both, i.e. in this case, CC(MSR) is executed followed by CC(SMSR) and this strategy is called as CC(MSR + SMSR).

5.1. *Performance measure*

As a performance measure, *match score*[13] has been used which measures the degree of similarity of two sets of biclusters. Let $\mathcal{M}_1(I_1, J_1)$ and $\mathcal{M}_2(I_2, J_2)$ be two biclusters. The gene match score $S_I(I_1, I_2)$ and condition match score $S_J(J_1, J_2)$ are defined as: $S_I(I_1, I_2) = \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|}$ and $S_J(J_1, J_2) = \frac{|J_1 \cap J_2|}{|J_1 \cup J_2|}$, respectively. Note that

gene and condition match scores are symmetric and vary from 0 (when two sets are disjoint) to 1 (when two sets are identical).

In order to evaluate the similarity among two sets of biclusters, the average gene match score and average condition match score can be computed. Let B_1 and B_2 be two sets of biclusters. The average gene match score of B_1 with respect to B_2 can be defined as:

$$S_I^*(B_1, B_2) = \frac{1}{|B_1|} \sum_{(I_1,J_1)\in B_1} \max_{(I_2,J_2)\in B_2} S_I(I_1, I_2). \tag{14}$$

$S_I^*(B_1, B_2)$ represents the average of the maximum gene match scores for all the biclusters in B_1 with respect to the biclusters in B_2. Note that $S_I^*(B_1, B_2)$ is not symmetric and yields different values if B_1 and B_2 are exchanged. Similarly, average condition match score can be defined as:

$$S_J^*(B_1, B_2) = \frac{1}{|B_1|} \sum_{(I_1,J_1)\in B_1} \max_{(I_2,J_2)\in B_2} S_J(J_1, J_2). \tag{15}$$

The overall average match score of B_1 with respect to B_2 can now be defined as:

$$S^*(B_1, B_2) = \sqrt{(S_I^*(B_1, B_2) \times S_J^*(B_1, B_2))}. \tag{16}$$

If B_{im} denotes the set of implanted biclusters and B is the set of biclusters provided by some biclustering method, then the *average module recovery*, $S^*(B_{im}, B)$, represents how well each of the true biclusters is recovered by the biclustering algorithm. This score ranges from 0 to 1 and takes the maximum value of 1, when $B_{im} = B$.

5.2. *Results for artificial data*

A synthetic dataset of size 500×100 has been constructed as follows: first a random 500×100 background matrix is generated. Thereafter, a perfect shifting bicluster and a perfect scaling bicluster of random sizes are implanted in random positions of the background matrix. The shifting and scaling factors for rows and columns of the biclusters are generated randomly with uniform distribution.

For comparing the performances of the algorithms, the overall average module recovery has been computed for different noise levels. Noise is added in the data matrices by adding random values generated from normal distribution. The mean of the normal distribution is fixed to 0 and the standard deviation (noise width) σ is varied from 0 (no noise) to 0.25 (maximum noise). For each value of σ, 20 different random noise matrices are added to the original data matrix and average performance metric values are reported in Table 1 for the three algorithms. It is evident from the table that at $\sigma = 0$, i.e. when there is no noise, the performance scores for CC(MSR), CC(SMSR) and CC(MSR + SMSR) are 0.5, 0.5 and 1.0, respectively. This indicates that CC(MSR) is able to detect the implanted shifting bicluster correctly but fails to detect the scaling bicluster. On the other hand, CC(SMSR) is able to detect the implanted scaling bicluster correctly but fails to detect the shifting bicluster. However, when we combine the algorithms (CC(MSR + SMSR)), the

Table 1. Variation of *average module recovery* $[S^*(B_{im}, B)]$ for different algorithms with respect to noise for the artificial data.

Noise width σ	CC(MSR)	CC(SMSR)	CC(MSR + SMSR)
0.00	0.5000	0.5000	1.0000
0.05	0.4934	0.4839	0.9172
0.10	0.4867	0.4873	0.8506
0.15	0.4799	0.4508	0.8438
0.20	0.4012	0.4025	0.8132
0.25	0.3662	0.3806	0.7354

performance score is 1.0. This signifies that both the shifting and scaling biclusters are properly identified. As the noise width increase, the average module recovery for all the three methods gradually decreases as expected, however, CC(MSR) and CC(SMSR) produce similar values while CC(MSR + SMSR) produces the performance score roughly twice of that produced by the other two algorithms. This indicates that a better performance from CC algorithm can be obtained if both *MSR*- and *SMSR*-based implementations are used one by one rather than using *MSR* or *SMSR* only.

5.3. *Results for real-life data*

In this section, the biclustering algorithms CC(MSR), CC(SMSR) and CC(MSR + SMSR) are applied to the benchmark Yeast and Lymphoma datasets. The *MSR* thresholds for the two datasets are set to 300 and 1200, respectively as in Ref. 4. To set the *SMSR* threshold for the Yeast dataset, the genes of the dataset are clustered into 30 clusters (as in Ref. 14) by K-means clustering with correlation-based distance and the *SMSR* value is computed for each of the clusters. The minimum *SMSR* score is found to be 0.0024. The *SMSR* threshold value of 0.002 is used in the experiment to detect more refined patterns. Similarly, the Lymphoma data are clustered into 40 clusters (as the number of genes in this data is almost 1.4 times of that in the Yeast data) and the minimum *SMSR* value of the clusters is found to be 1.8403. The *SMSR* threshold for the Lymphoma dataset is set to 1.4.

The main objective of the experiments in this section is to demonstrate the utility of the proposed scaling coherence measure and to show that the biclusters detected by CC(SMSR) technique are mostly missed by CC(MSR) algorithm. For this purpose, the three biclustering algorithms considered here are run on the two datasets to extract the first 100 biclusters. Table 2 reports the average gene match scores (S_I^*), average condition match scores (S_J^*) and average overall match scores (S^*) over 10 runs of each of the algorithms along with the standard deviations. The match score values are computed for CC(SMSR) bicluster with respect to CC(MSR) biclusters and vice versa for both the datasets. It can be noticed from the table that the gene match scores (S_I^*) and overall match scores (S^*) are on the lower side (less than 0.2) whereas the condition match scores are on the higher side

Table 2. Average gene match scores (S_I^*), average condition match scores (S_J^*) and average overall match scores (S^*) over 10 runs of CC(MSR) and CC(SMSR) algorithms along with the standard deviations.

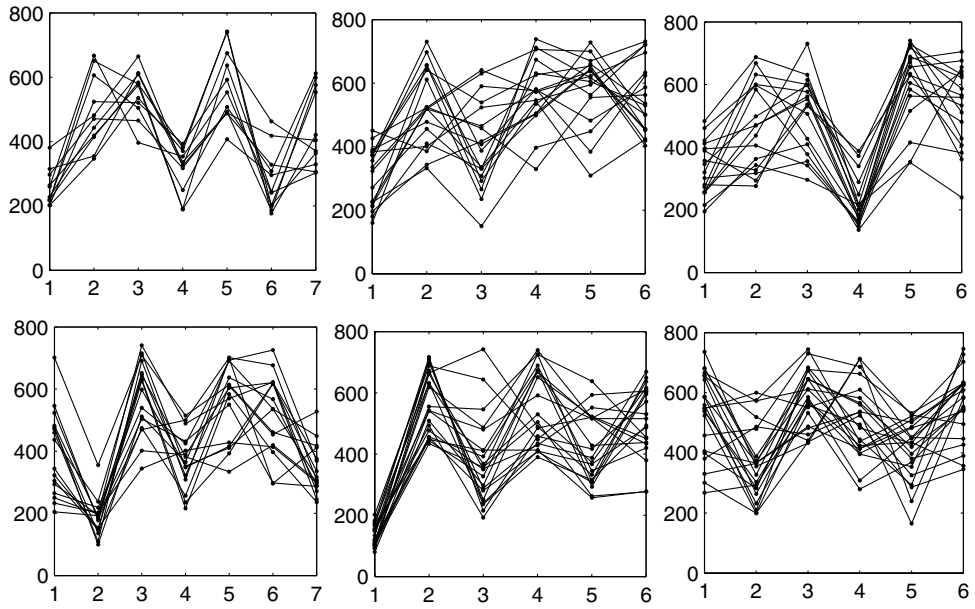
	Yeast	Lymphoma
$S_I^*(CC_{MSR}, CC_{SMSR})$	0.0619 ± 0.0026	0.0830 ± 0.0061
$S_J^*(CC_{MSR}, CC_{SMSR})$	0.6117 ± 0.0104	0.3846 ± 0.0272
$S^*(CC_{MSR}, CC_{SMSR})$	0.1946 ± 0.0034	0.1474 ± 0.0261
$S_I^*(CC_{SMSR}, CC_{MSR})$	0.0366 ± 0.0101	0.0727 ± 0.0072
$S_J^*(CC_{SMSR}, CC_{MSR})$	0.5304 ± 0.0032	0.3236 ± 0.0241
$S^*(CC_{SMSR}, CC_{MSR})$	0.1393 ± 0.0012	0.1363 ± 0.0153

Table 3. Average coverage of the biclusters produced by the algorithms CC(MSR), CC(SMSR) and CC(MSR + SMSR) over 10 runs of the algorithms along with the standard deviations.

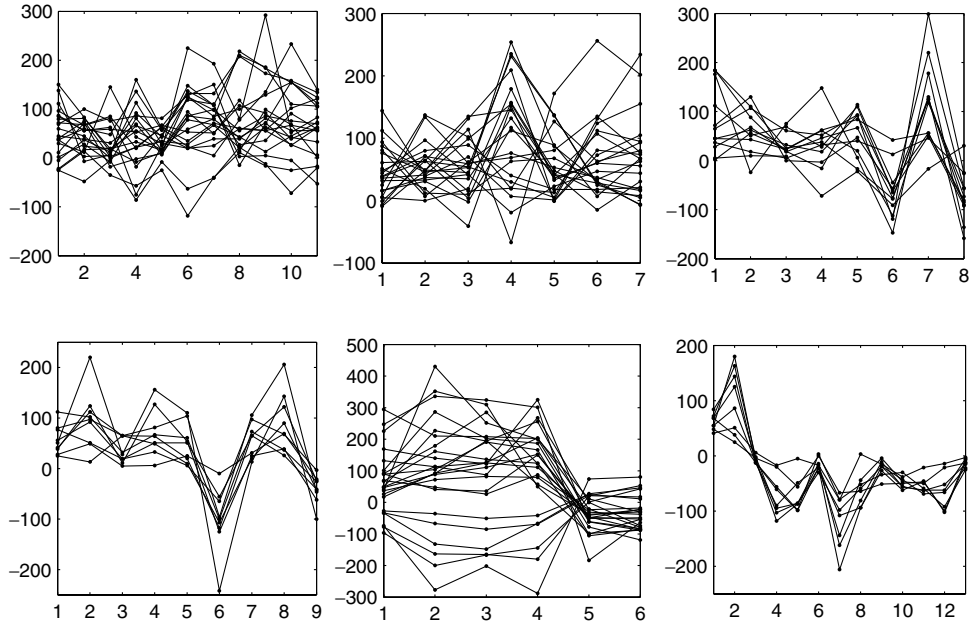
	Yeast	Lymphoma
CC(MSR)	73.3136 ± 0.2718	42.6496 ± 0.1829
CC(SMSR)	69.0372 ± 0.3193	46.8293 ± 0.2816
CC(MSR + SMSR)	92.2943 ± 0.2937	69.3846 ± 0.2422

(greater than 0.5). This implies that the CC(MSR) biclusters (shifting patterns) and CC(SMSR) biclusters (scaling patterns) share many columns (time points) of the datasets, however they share a very small number of rows (genes) and thus a small number of cells in the gene expression matrix. This finding is important since it signifies that the biclusters identified by CC(SMSR), which are having scaling patterns, are mostly missed by the CC(MSR) algorithm. This demonstrates the utility of using *SMSR* as a coherence measure. This is also confirmed by the results in Table 3, where we report the average coverage (percentage of cells of the gene expression matrix covered by a set of biclusters) of the biclusters produced by the algorithms CC(MSR), CC(SMSR) and CC(MSR + SMSR) over 10 runs of the algorithms along with the standard deviations. It is evident from the table that the coverage of the biclusters for CC(MSR) and CC(SMSR) are almost the same, whereas the coverage of the biclusters obtained by CC(MSR + SMSR) is much greater than that. This indicates that the cells covered by the biclusters produced by CC(MSR) and CC(SMSR) are mostly not common, i.e. the biclusters identified by CC(SMSR) are not detected by CC(MSR) method.

For the purpose of illustration, Figs. 1(a) and 1(b) show six biclusters identified by CC(SMSR) method for the Yeast and Lymphoma data, respectively. Evidently, these twelve biclusters have scaling patterns. Such scaling biclusters were not reported in Ref. 4 where CC(MSR) algorithm was used. Moreover, many of these biclusters are having both upregulated and downregulated genes which are interesting from the biological point of view.



(a)



(b)

Fig. 1. Six CC(SMSR) biclusters of the (a) Yeast data and (b) Lymphoma data.

5.4. *Biological significance test*

The biological relevance of the biclusters can be verified based on the GO annotation database. This is used to test the functional enrichment of a group of genes in terms of three structured, controlled vocabularies (ontologies), that is, biological processes, molecular functions and biological components. The *p-value* of a statistical-significance test is used to find the probability of getting the values of a test statistic that are at least equal to in magnitude (or more) compared to the observed test statistic. The degree of functional enrichment (*p-values*) is computed using a cumulative hypergeometric distribution that measures the probability of finding the number of genes involved in a given GO term within a bicluster. From a given GO category, the probability p for getting k or more genes within a cluster of size n, can be defined as[13]: $p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}}$, where f and g denote the total number of genes within a category and within the genome, respectively. If the majority of genes in a bicluster have the same biological function, then it is unlikely that this takes place by chance and the *p-value* of the category will be close to 0.

The biological significance tests for the Yeast and Lymphoma datasets have been conducted at 1% significance level. Among the 100 biclusters produced by CC(MSR) and CC(SMSR) algorithms, the number of biclusters with at least one significant GO term (*p-value* < 0.01) are 11 and 14 for the Yeast data, and 10 and 9 for the Lymphoma data, respectively. Tables 4 and 5 report the top five different biclusters with respect to *p-values* for both the algorithms for the Yeast and Lymphoma datasets, respectively. We have reported the top five different biclusters which have different most significant GO terms. These biclusters are then arranged in ascending order of *p-values* (i.e. descending order of significance) of the most significant GO terms. The corresponding GO terms are also reported. Moreover, the number of genes and the number of conditions for each bicluster are reported in brackets.

The *p-values* reported in Table 4 suggest that the scaling patterns [detected by CC(SMSR)] are at least of equal importance with shifting patterns in analyzing

Table 4. Result of biological significance test: the top five functionally enriched significant biclusters produced by each algorithm for the *Yeast data*. Corresponding GO terms and the *p-values* are reported. The number of genes and conditions in the biclusters are also reported in brackets.

Algorithm	Bicluster 1	Bicluster 2	Bicluster 3	Bicluster 4	Bicluster 5
CC(MSR)	ribosome GO:0005840 p-val: 8.6E-37 (112, 17)	cytosolic part GO:0044445 p-val: 6.0E-24 (74, 15)	ribosome biogenesis and assembly GO:0042254 p-val: 4.9E-11 (619, 17)	translation GO:0006412 p-val: 3.4E-07 (27, 8)	sulfar metabolic process GO:0006790 p-val: 2.7E-07 (28, 6)
CC(SMSR)	cytosolic part GO:0044445 p-val: 2.8E-45 (326, 17)	chromosomal part GO:0044427 p-val: 2.0E-15 (67, 16)	microtubule nucleation GO:0007020 p-val: 1.2E-11 (97, 15)	mitochondrial lumen GO:0031980 p-val: 1.1E-08 (39, 11)	mitosis GO:0007067 p-val: 2.5E-06 (64, 9)

Table 5. Result of biological significance test: the top five functionally enriched significant biclusters produced by each algorithm for the *Lymphoma data*. Corresponding GO terms and the *p-values* are reported. The number of genes and conditions in the biclusters are also reported in brackets.

Algorithm	Bicluster 1	Bicluster 2	Bicluster 3	Bicluster 4	Bicluster 5
CC(MSR)	voltage-gated potassium channel activity GO:0005249 p-val: 3.5E-13 (81, 10)	multicellular organismal development GO:0007275 p-val: 4.6E-11 (198, 63)	cell surface receptor linked signal transduction GO:0007166 p-val: 5.4E-11 (33, 14)	amine receptor activity GO:0008227 p-val: 2.2E-10 (135, 32)	cell-cell signaling GO:0007267 p-val: 2.3E-10 (239, 40)
CC(SMSR)	sequence-specific DNA binding GO:0043565 p-val: 1.4E-11 (231, 35)	ion transport GO:0006811 p-val: 4.8E-11 (104, 7)	extracellular ligand-gated ion channel activity GO:0005230 p-val: 5.6E-10 (110, 71)	potassium channel activity GO:0005267 p-val: 3.2E-09 (12, 13)	multicellular organismal development GO:0007275 p-val: 5.8E-09 (111, 75)

microarray gene expression data. In fact, it is evident from the table that among the total 10 biclusters reported for the Yeast data, the minimum *p-value* (2.8E−45) is obtained for the first bicluster of CC(SMSR) algorithms. It is evident from the table that only one bicluster has common significant GO term (*cytosolic part*) among the top five biclusters produced by CC(MSR) and CC(SMSR) algorithms.

In the case of the Lymphoma data (Table 5), the minimum *p-value* (3.5E−13) is obtained for the first bicluster of CC(MSR) algorithm. For this dataset also, only one bicluster has common significant GO term (*multicellular organismal development*) among the top five biclusters of CC(MSR) and CC(SMSR).

Therefore the biological significance test reveals that the proposed *SMSR*-based CC(SMSR) technique is able to detect biclusters having strong biological significance which are not detected by *MSR*-based CC(MSR) algorithm. This demonstrates the utility of the proposed scaling residue measure.

6. Conclusions

Recent research[11] has revealed that mean squared residue (MSR), a popular metric that is optimized by many biclustering algorithms, is capable of detecting shifting patterns only and fails to capture scaling patterns. Motivated by this, in this article, a new coherence measure called scaling mean squared residue (SMSR) is proposed and we have theoretically proved that the new measure is able to detect the scaling patterns. The effectiveness of the proposed coherence measure has been demonstrated experimentally on one artificial dataset and two benchmark real-life gene expression datasets. Finally biological significance tests have been conducted to establish that the scaling biclusters discovered by SMSR-based algorithm are composed of functionally enriched sets of genes.

As a scope of future research, the new SMSR measure can be incorporated to the other biclustering algorithms which are currently based on MSR.[6,7,10] Moreover, the use of both MSR and SMSR together in a multiobjective framework[15] to detect shifting and scaling patterns simultaneously can also be studied. The authors are working in these directions.

Acknowledgment

Sanghamitra Bandyopadhyay gratefully acknowledges the financial support received from the grant no. DST/SJF/ET-02/2006-07 under the Swarnajayanti Fellowship scheme of the Department of Science and Technology, Government of India.

References

1. Jain AK, Dubes RC, *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
2. Bandyopadhyay S, Mukhopadhyay A, Maulik U, An improved algorithm for clustering gene expression data, *Bioinformatics* **23**(21):2859–2865, 2007.
3. Maulik U, Mukhopadhyay A, Bandyopadhyay S, Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes, *BMC Bioinformatics* **10**:27, 2009.
4. Cheng Y, Church GM, Biclustering of gene expression data, in *Proc Int Conf on Intelligent Systems for Molecular Biology (ISMB'00)*, pp. 93–103, 2000.
5. Getz G, Levine E, Domany E, Coupled two-way cluster analysis of gene microarray data, In *Proc Natl Acad Sci USA* **97**:12079–12084, 2000.
6. Yang J, Wang W, Wang H, Yu P, Enhanced biclustering on expression data, in *Proc 3rd IEEE Conf Bioinformatics and Bioengineering (BIBE'03)*, pp. 321–327, 2003.
7. Bleuler S, Prelic A, Zitzler E, An EA framework for biclustering of gene expression data, in *Proc Congress on Evolutionary Computation*, pp. 166–173, 2004.
8. Maulik U, Mukhopadhyay A, Bandyopadhyay S, Zhang MQ, Zhang X, Multiobjective fuzzy biclustering in microarray data: Method and a new performance measure, in *Proc IEEE World Congress on Computational Intelligence (WCCI 2008)/IEEE Congress on Evolutionary Computation (CEC 2008)*, pp. 383–388, Hong Kong, 2008.
9. Tanay A, Sharan R, Shamir R, Discovering statistically significant biclusters in gene expression data, *Bioinformatics* **18**:S136–S144, 2002.
10. Bryan K, Cunningham P, Bolshakova N, Biclustering of expression data using simulated annealing, in *Proc 18th IEEE Symposium on Computer-Based Medical Systems, (CBMS 2005)*, pp. 383–388, Dublin, Ireland, 2005.
11. Aguilar-Ruiz JS, Shifting and scaling patterns from gene expression data, *Bioinformatics* **21**(20):3840–3845, 2005.
12. Madeira SC, Oliveira AL, Biclustering algorithms for biological data analysis: A survey, *IEEE Trans Comput Biol Bioinform* **1**(1):24–45, 2004.
13. Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E, A systematic comparison and evaluation of biclustering methods for gene expression data, *Bioinformatics* **22**(9):1122–1129, 2006.
14. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM, Systematic determination of genetic network architecture, *Nature Genet* **22**:281–285, 1999.
15. Deb K, *Multi-Objective Optimization Using Evolutionary Algorithms*, John Wiley and Sons, Ltd, England, 2001.

Anirban Mukhopadhyay is currently a faculty member in the Department of Computer Science and Engineering, University of Kalyani, India. He received his B.E., M.E. and Ph.D. degrees in 2002, 2004 and 2009, respectively, all in Computer Science and Engineering. Dr. Mukhopadhyay was the recipient of the University Gold Medal and Amitava Dey Memorial Gold Medal from Jadavpur University in 2004.

He received Erasmus Mundus post-doctoral fellowship in 2009. His biography has been included in the 2009 Edition of *Who is Who in the World*. Dr. Mukhopadhyay has coauthored about 45 research papers in various international journals and conferences. His research interests include soft and evolutionary computing, data mining, bioinformatics, and optical networks.

Ujjwal Maulik is currently a Professor in the Department of Computer Science and Engineering, Jadavpur University, Kolkata, India. Dr. Maulik received his B.S. degrees in Physics and Computer Science in 1986 and 1989, respectively. Subsequently, he received his M.S. and Ph.D. in Computer Science in 1991 and 1997, respectively. He was the recipient of the Govt. of India BOYSCAST fellowship in 2001.

Dr. Maulik has worked in Center for Adaptive Systems Application, Los Alamos, New Mexico, USA in 1997, University of New South Wales, Sydney, Australia in 1999, University of Texas at Arlington, USA in 2001, Univ. of Maryland Baltimore County, USA in 2004, Fraunhofer Institute AiS, St. Augustin, Germany in 2005, Tsinghua University, China in 2007 and University of Rome, Italy in 2008. He has coauthored four books and about 135 papers. His research interests include evolutionary computing, pattern recognition, data mining, bioinformatics, and distributed systems.

Sanghamitra Bandyopadhyay received her B.S. degree in Physics in 1991, and M.S. and Ph.D. degrees in Computer Science in 1993 and 1998, respectively. Currently she is a Professor at Indian Statistical Institute, India. Dr. Bandyopadhyay is the first recipient of Dr. Shanker Dayal Sharma Gold Medal and also the Institute Silver Medal for being adjudged the best all-round post-graduate performer in IIT, Kharagpur, India, in 1994.

She worked in Los Alamos National Laboratory, Los Alamos, USA in 1997, University of New South Wales, Sydney, Australia, in 1999, University of Texas at Arlington, USA in 2001, University of Maryland at Baltimore, USA in 2004, Fraunhofer Institute, Germany in 2005, and Tsinghua University, China from 2006 to 2007 and University of Rome, Italy in 2008. She received the Young Scientist

awards of the Indian National Science Academy (INSA) in 2000, the Indian Science Congress Association (ISCA) in 2000, and Indian National Academy of Engineers (INAE) in 2002. She also received Swarnajayanti Fellowship from the Govt. of India in 2006–07 and Humboldt Fellowship in 2009. Dr. Bandyopadhyay has coauthored six books and more than 150 papers. Her research interests include computational biology and bioinformatics, soft and evolutionary computation, pattern recognition and data mining.