Improving Biological Significance of Gene Expression Biclusters with Key Missing Genes

Shufan Ji  
Department of Computer Science  
Beihang University  
Beijing, China 100191  
jishufan@buaa.edu.cn

Xing Tian  
Department of Computer Science  
Beihang University  
Beijing, China 100191  
txsing@gmail.com

Jin Chen*  
Department of Energy Plant Research Laboratory  
Department of Computer Science and Engineering  
Michigan State University  
East Lansing, MI  
jincheng@msu.edu

ABSTRACT

Identifying condition-specific co-expressed gene groups is critical for gene functional and regulatory analysis. However, given that genes with critical functions (such as transcription factors) may not co-express with their target genes, it is insufficient to uncover gene functional associations only from gene expression data. In this paper, we propose a novel integrative biclustering approach to build high quality biclusters from gene expression data, and to identify critical missing genes in biclusters based on Gene Ontology as well. Our approach delivers a complete inter- and intra-bicluster functional relationship, thus provides biologists a clear picture for gene functional association study. We experimented with the Yeast cell cycle and Arabidopsis cold-response gene expression datasets. Experimental results show that a clear inter- and intra-bicluster relationship is identified, and the biological significance of the biclusters is considerably improved.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Clustering;  
J.3 [Life and Medical Sciences]: Biology and genetics

General Terms

Algorithms

Keywords

bi-clustering, Gene Ontology, gene expression, missing gene, biological network

*contact author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

BCB’15, September 9–12, 2015, Atlanta, GA, USA.
Copyright 2015 ACM 978-1-4503-3853-0/15/09 ...$15.00.
http://dx.doi.org/10.1145/2808719.2808747.

1. INTRODUCTION

Gene expression clustering is a key step towards gene functional and regulatory analysis at the genome level [36]. Genes that have similar expression patterns under certain subsets of experimental conditions can be grouped together, and are considered to be functionally related [14, 8, 41]. Biclustering analysis, that identifies condition-specific co-expressed patterns from gene expression data, is essential to better understanding of the biological interactions between genes that are not apparent when clustered globally [7, 9].

In the last decade, many biclustering algorithms have been proposed to simultaneously cluster genes and experimental conditions together [7, 39, 40, 42, 13, 19, 21, 6, 9]. However, it has been shown that although these biclustering methods can capture gene groups that function together under a subset of experimental conditions, there is a lot of potential to improve towards uncovering gene regulation relationships, mainly because of the following two reasons:

1. Biclustering based on gene expression data would inevitably miss certain functionally related genes, mainly because some functionally related genes do not co-express. For example, transcription factors (TF) usually do not co-express with their target genes, and if a key TF is missing from a bicluster, the functional association among the related genes is difficult to identify.

2. It is the inter- and intra-cluster relationship rather than the clustered genes that contributes towards better interpretation of the overall picture of gene regulatory networks [17, 22, 32, 2]. Although hierarchical biclustering algorithms [13, 21, 6] can reveal the inter-bicluster relationships, to our knowledge no one has extensively studied both the inter- and intra-bicluster relationships.

In summary, some functionally related genes may be missing in a bicluster. As a result, the intra-bicluster connections are weakened with reduced functional association. The gene regulatory networks built upon such biclusters would then be incomplete and disconnected.

In order to generate and interpret biological hypotheses on gene transcriptional regulations, a biclustering algorithm that can 1) build high quality biclusters from gene expression data, 2) reveal inter- and intra-bicluster relationships, and
3) identify missing genes and missing gene associations, is highly expected.

In this paper, we propose to identify gene expression biclusters with a novel biclustering algorithm called SKeleton Biclustering (SKB). SKB not only builds biclusters and reveals bicluster skeleton (inter-bicluster and intra-bicluster relationships), but also identifies relevant missing genes that can bridge the functionally distinct genes within each bicluster, so as to uncover the otherwise hidden gene associations within biclusters. As shown in Figure 1, SKB has three phases. In phase one, a hierarchical biclustering method QHB [13] is adopted to generate biclusters from gene expression data, which reveals the hierarchical inter-bicluster relationship by a dendrogram. The inter-bicluster relationships are revealed by a hierarchical tree. In phase two, each bicluster is converted into a graph (made up of several disconnected subgraphs) by linking the genes based on their functional similarity, which is calculated with Gene Ontology (GO) annotation [1, 10]. Gene connections within the bicluster are defined as the intra-bicluster relationship. In phase three, a graph mining algorithm is proposed to efficiently add new genes, if any, to each bicluster, to re-connect the disconnected subgraphs generated in phase 2, in order to enhance the intra-bicluster relationship. Therefore, the overall functional connectivity of the bicluster is increased.

Experimental results on Yeast cell cycle [7] and Arabidopsis cold-response microarray datasets [37] show that SKB could build a clear structure of the inter- and intra-bicluster relationship, consequently the mined biclusters have higher biological significance than existing methods. Comparing with existing biclustering algorithms, we have made four main contributions:

1. New Concept of Biclustering. We picture the whole skeleton of biclusters. The proposed concept of bicluster skeleton (i.e., inter- and intra-cluster relationships) could reveal the hierarchical subset/superset of a gene bicluster and the relationships among genes within a bicluster. Moreover, we improve the bicluster quality by adding new genes and new connections into biclusters to reinforce the intra-bicluster relationships.

2. Novel Way to Employ Domain Knowledge. It is generally difficult to integrate mining data with domain knowledge in unsupervised learning models. Work to date either combines domain knowledge with mining data to form a complicated feature space, or takes domain knowledge as a post-processing filter [23, 30, 35, 12, 4]. However, the former approach suffers from the complicated combination of heterogeneous features, while the latter approach results in data loss. In this paper, we employ the domain knowledge as an independent data space, onto which the mining results are mapped for further improvement.

3. Improved Biological Significance. SKB improves the biological significance of biclusters by reinforcing the gene functional associations. Different from most existing biclustering algorithms, our method can identify key missing genes that express differently but functionally associatively with biclustered genes, thus delivers a complete skeleton of biclusters.

4. Efficient Algorithm. Both gene biclustering and missing gene identification are computationally expensive. We introduce a novel algorithm to achieve the goal efficiently.

2. RELATED WORK

The biclustering algorithm on gene expression data was first introduced by Cheng and Church 2000 [7] to simultaneously cluster both genes and experimental conditions, which captures the coherence of a subset of genes under a subset of experimental conditions [7]. A two-phase iterative strategy is used to identify biclusters in the data: 1) remove rows and columns from the gene expression matrix to compose a submatrix with the mean squared residue score below a threshold δ; 2) add the deleted rows and columns to the submatrix as along as the mean squared residue score does not exceed the threshold. δ-cluster model [40] was later proposed to accelerate the mining process with a move-based algorithm FLOC, which randomly generates initial biclusters and iteratively optimizes biclusters by adding/deleting genes or conditions. [39] further accelerated the biclustering process by employing a depth-first algorithm to mine pClusters with a user specified minimum pScore. However, those biclustering algorithms fail to generate deterministic mining results.

A deterministic algorithm DBF was proposed [42] to further improve the quality and efficiency of biclustering. DBF
employs a frequent closed pattern mining algorithm to generate “good seeds” with better pattern similarity and then refines the seeds by adding genes and conditions to achieve a low MSR score as well as large bicluster volume. A minimum row variance threshold is set to remove biclusters with trivial changes in trends. In particular, on a binary gene expression matrix, we can view the biclustering formulation as the search for all maximal bicliques in a bipartite graph, where the nodes are genes or conditions and an edge connecting a gene \( g \) and a condition \( c_j \) indicates gene \( g \)’ responses in the condition \( c_j \) [38]. Subsequently, the biclustering problem on a binary matrix can solved by the backtracking Bron-Kerbosch algorithm for maximal clique enumeration [5].

While these algorithms can generate biclusters with similar trends, they are limited in several ways. First, these schemes typically employ a similarity score (e.g., MSR and pScore) to determine the quality of biclusters, which, however, cannot adequately capture the trend consistency of biclusters. Second, these algorithms usually generate biclusters based on selected “seeds” that cover only a small part of the whole dataset. As such, interesting patterns may be missed resulting in loss of relevant information. Third, the seed improvement process follows the hill-climbing paradigm, and hence, is computationally expensive. To overcome these limitations, a quick hierarchical biclustering algorithm called QHB was proposed to generate consistent-trends biclusters (a subsets of genes homogeneously expressed in a subset of conditions) and produces a hierarchical tree to reveal the inter-bicluster relations [13].

More recently, model-based biclustering algorithms [20, 21], which approximate the data density by a mixture of Gaussian distributions, adopt a hierarchical structure of biclusters to define a flexible distribution of data density. In [6], a graphical model was adopted to identify hierarchical biclusters. The resulting tree structure of conditions, where each node corresponds to a bicluster, provides an interpretation basis for the grouping [6]. LateBiclustering identifies biclusters with time lags, thus recovering genes that are missed because they co-expressed with delays [11]. DeBi improves the biological significance of biclusters by discovering maximum size homogeneous biclusters, in which each gene is strongly associated with a subset of samples, providing functionally more coherent gene sets compared to standard clustering or biclustering algorithms [33].

Note that biclustering methods solely based on gene expression data would inevitably leave some functionally related genes out of the biclusters [28], and thus weaken the overall gene relationships. Hence, incorporating biological domain knowledge in biclustering [23, 30, 35] has been recognized as a reliable way to enhance the interpretability of biclustered genes. Gene Ontology (GO) [1, 10] has now been accepted as the de facto language for the attribute description of biological entities in three key domains that are shared by all organisms, namely molecular function, biological process and cellular component [29, 24, 27, 26, 25]. In each of these domains, the corresponding GO is structured as a directed acyclic graph to reflect the complex hierarchy of biological terminologies. To incorporate Gene Ontology to increase biclustering performance, we propose a biclustering algorithm called Skeleton Biclustering (SKB) that could build high quality biclusters with clear inter- and intra-bicluster relationships, as well as identify missing genes and missing gene associations.

### Table 1: Original Data Matrix \( A \)

<table>
<thead>
<tr>
<th>( A )</th>
<th>( c_1 )</th>
<th>( c_2 )</th>
<th>( c_3 )</th>
<th>( c_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_1 )</td>
<td>2.4</td>
<td>2.95</td>
<td>2.45</td>
<td>2.99</td>
</tr>
<tr>
<td>( g_2 )</td>
<td>1.95</td>
<td>1.71</td>
<td>1</td>
<td>0.29</td>
</tr>
<tr>
<td>( g_3 )</td>
<td>0.5</td>
<td>1.1</td>
<td>0.38</td>
<td>1.56</td>
</tr>
</tbody>
</table>

### Table 2: Slope Angle Matrix \( A' \)

<table>
<thead>
<tr>
<th>( A' )</th>
<th>( c_1c_2 )</th>
<th>( c_2c_3 )</th>
<th>( c_3c_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_1 )</td>
<td>28.81°</td>
<td>-26.57°</td>
<td>28.37°</td>
</tr>
<tr>
<td>( g_2 )</td>
<td>-13.50°</td>
<td>-35.37°</td>
<td>-35.37°</td>
</tr>
<tr>
<td>( g_3 )</td>
<td>30.96°</td>
<td>-35.75°</td>
<td>49.72°</td>
</tr>
</tbody>
</table>

### Table 3: Binary Matrix \( A'' \) (\( BinningThreshold = 26.5^\circ \))

<table>
<thead>
<tr>
<th>( A'' )</th>
<th>( (c_1c_2) )</th>
<th>( (c_2c_3) )</th>
<th>( (c_3c_4) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_1 )</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( g_2 )</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( g_3 )</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

### 3. METHOD

In this section, we introduce the three phases of SKB respectively. The overall framework of SKB is in Figure 1.

#### 3.1 SKB phase 1, inter-bicluster relationship identification

We employ an efficient top-down hierarchical biclustering algorithm QHB [13] (without modification) as the first phase of SKB to identify biclusters from gene expression data. We generate biclusters with consistent rising / falling trends and produce a hierarchical tree to reveal the inter-bicluster relationship. Here we adopt QHB algorithm in that it could deliver a well-organized hierarchical tree with each node representing a bicluster, from which we could easily identify the subset/superset and the nearest neighbors of a bicluster, which will facilitate gene family studies. Now that we will briefly describe the algorithm in two steps with a running example.

**Step 1: Trend Separation**

In Step 1, different changing trends of gene expression data are separated. Firstly, the gene expression matrix \( A \) (Table 1), where rows represent genes and columns represent experimental conditions, is transformed into a slope angle matrix \( A' \) (Table 2) to reflect different changing trends when the experimental conditions change. Then \( A' \) is binned into a binary matrix \( A'' \) (Table 3) where the rising(01) and falling(10) trends are separated into two consecutive columns; trends with trivial change(00) are filtered out.

**Step 2: Matrix Partition**

In Step 2, genes with similar changing trends (rising and falling simultaneously) are grouped into biclusters through a hierarchical partition tree in Figure 2. From matrix \( A'' \), the partitioning process constructs a hierarchical tree where all valuable upper level information is kept intact and propagated into the lower level. This helps in preventing any information loss. The binning and partitioning ensures that genes with consistent trends under condition transitions are kept together in the same biclusters while genes with incon-
Figure 2: Matrix ($A'$) Partition. Genes with similar changing trends (rising and falling simultaneously) are grouped into biclusters through a hierarchical partition tree, where all valuable upper level information is kept intact and propagates into the lower level.

3.2 SKB phase 2, gene distance measure and bicluster graph generation

In this phase, each gene expression bicluster (on the leaf level of the hierarchical biclustering tree) is converted into an undirected graph by linking functionally associated genes together. In general, two genes are considered to be functionally associated if they share at least one biological feature. Here we adopt GO annotation for the function association measure, defined as gene distance. Thus, genes satisfying a certain distance threshold are linked together. Clearly, the algorithm could be easily extended to adopt other domain knowledge as gene distance measure.

GO Term Similarity Measure

As the GO terms are not equally informative [29], for each gene set, specific weights are assigned to the GO terms [18]: the weight of a GO term is defined as the occurrence ratio of the GO term (with all its descendants) to the total terms in the gene set. That is, for any GO term $t \in T$, where $T$ is the full set of GO terms, the term weight is defined as $w(t) = \frac{\text{freq}(t) + \sum_{d \in D_t} \text{freq}(d))}{N}$, where $\text{freq}(t)$ denotes the number of occurrences of GO term $t$; $D_t$ is the set of descendants of GO term $t$ in $T$; and $N$ is the total number of term occurrences.

Given two GO terms $t_a$ and $t_b$ and their weights $w(t_a)$ and $w(t_b)$, their term similarity score $\text{sim}(t_a, t_b)$ is defined in Equation 1 [31]. Note that $t_{ab}$ is the nearest common ancestor of $t_a$ and $t_b$, and $w(t_{ab})$ is its weight.

$$\text{sim}(t_a, t_b) = \frac{2 \times \ln w(t_{ab}) \times (1 - w(t_{ab}))}{\ln w(t_a) + \ln w(t_b)}$$ (1)

Gene Distance Measure

Biologically, many genes are involved in multiple cellular processes and therefore labeled with more than one GO term. Thus, the distance between two genes is defined as the minimum dissimilarity between any two GO terms annotating the two genes [34]. Let $T_{g_i}$ and $T_{g_j}$ be the set of GO terms annotating gene $g_i$ and $g_j$, respectively. The gene distance measure is defined as:

$$d(g_i, g_j) = \min_{t_a \in T_{g_i}, t_b \in T_{g_j}} (1 - \text{sim}(t_a, t_b))$$ (2)

Bicluster Graph Generation

To reveal the biological relationship among the biclustered genes, a bicluster $S$ is converted into an undirected bicluster graph $G(V, E)$, with the vertex $g_i \in V$ representing gene $g_i$, and the edge $e(g_i, g_j) \in E$ indicating that the gene distance $d(g_i, g_j)$ is smaller than the predefined distance threshold $\sigma$. The graph $G$ is either connected or made up of several disconnected components, each of which is a connected subgraph of $G$ (see Algorithm 1 line 4-9).

3.3 SKB phase 3, adding new genes to improve biological significance

Without key functional genes, the originally associated genes in resulting bicluster graph might be separated into different components. As for the example in Figure 3, without gene $g_x$ (function $f_1, f_2$), gene $g_y$ (function $f_1, f_3$) is separated with gene $g_b$ (function $f_2, f_4$) due to non-similar functions (long gene-to-gene distance). $g_x$ could bridge the gap of $g_y$ and $g_b$, in that it shares the same function $f_1, f_2$ with $g_y, g_b$ respectively. Therefore, given gene $g_y$ and $g_b$ in a bicluster graph $G(V, E)$, there might exist a third gene $g_x$ such that $d(g_x, g_y) + d(g_y, g_b) < \sigma$ (the gene distance threshold), where $g_x \in V_1$ and $g_y, g_b \in V_j$.

The strategy to identify new genes is to search for the smallest set of new genes that connect all the connectable components in a bicluster graph (see pseudocode in Algorithm 1 line 10-22). We decided to identify the smallest set of new genes because of three reasons. First, we notice that owing to incomplete biological knowledge in GO annotation, introducing too many new genes could cause the bicluster to be biased toward known information. Specifically, if two genes share the same function that is not yet known or annotated, the distance between them will be artificially large. Therefore, we decided to add the minimum number of new genes as a constraint to weaken such bias towards known information. Second, by minimizing the number of new genes to be included, we are maximizing the number of components each new gene connects to. This results in new genes that are more informative. Third, the objective of adding
new genes is to understand the relationships among the biclustered genes, and too many new genes may reduce the gene expression coherence.

Algorithm 1 SKB phase 2 and 3

1: procedure AddNewGenes(T, I, σ)
2: | Input: S - bicluster genes; T - the full set of GO terms; I - the full set of genes in gene expression data; σ - gene distance threshold;
3: | Output: V_s - new gene set; G' - the new cluster graph;
4: | for each gene pair g_i, g_j ∈ S (i ≠ j) do
5: |getDistance(g_i, g_j, T) < σ then
6: | V = V ∪ {g_i, g_j}; E = E ∪ {< g_i, g_j >};
7: | end if
8: |end procedure
9: G = (V, E);
10: C' = getComponent(G);
11: if |C'| = 1 then No need to insert missing genes
12: |for each component pair C_i, C_j ∈ C' do
13: |if (Connectable(C_i, C_j)) = true C = C ∪ {C_i, C_j};
14: |end for
15: (V_s, E_s) = getExtraGenes(C); Identify missing genes
16: G' = (V ∪ V_s, E ∪ E_s);
17: end procedure

Figure 3: Effect of Key Gene. Adding key genes to a bicluster graph may help connect the otherwise disconnected components, leading to the enriched functional associations within biclusters.

Algorithm 2 getExtraGenes(C)

1: | Input: S - bicluster genes; T - the full set of GO terms; I - the full set of genes in gene expression data; σ - gene distance threshold; C - the set of connectable components; c - a constant large number;
2: | Output: V_s - new gene set; E_s - edge set that connect vertices in V_s & V;
3: | V_s = ∅; E_s = ∅; Λ = ∅; C_s = ∅;
4: |for each gene g ∈ I - S do Number of components g is connectable to
5: | N(g) = getConnComp(g, C, σ);
6: |if (N(g) ≥ 2) Λ = Λ ∪ {g};
7: |end for
8: |while Λ ≠ ∅ do
9: |remove g with max N(g) from Λ;
10: |V_s = V_s ∪ {g}; C_s = C_s ∪ {g};
11: |for each component C_i ∈ C that g connects to do
12: |g_i = getMinDist(C_i, g); E_s = E_s ∪ {< g, g_i >};
13: |C_s = C_s ∪ C_i; remove C_i from C;
14: |end for
15: |C = C ∪ {C_s}; C_s = ∅;
16: |for each gene g ∈ Λ do
17: |N(g) = getConnComp(g, C, σ);
18: |if (N(g) < 2) remove g from Λ;
19: |end for
20: end while
21: return V_s and E_s;

For a given bicluster S, instead of testing every subset of I - S exhaustively, which is a NP problem, we develop a heuristic algorithm to efficiently identify the local optimized smallest new gene set. Let V_s be the initial new gene set. The algorithm to identify V_s consists of two steps. First, candidate genes in I - S - V_s are sorted in descending order according to the number of components that it could connect in C. Candidate genes that could connect less than two components are ignored (see Algorithm 2 line 4-7). Second, the top candidate gene is added to V_s, connecting as many components in C as possible. Then the whole process is repeated until all components in C are connected or no more candidate gene exists (see Algorithm 2 line 8-20). This heuristic algorithm has polynomial time complexity.

For example, given a cluster graph with 9 disconnected components C_0, C_1, ..., C_8 in Figure 4, and five new gene candidates g_1, g_2, ..., g_5. In each iteration, the numbers of components that each gene could connect are listed in Table 4. As for this example, g_1 is firstly ignored since it cannot connect any two components. Meanwhile, g_1 is added to V_s in that it connects the most number of the components, i.e., 6. In the second iteration, the number of connections for each remaining gene candidate is recalculated after g_1 connects C_0, C_1, C_2, C_3, C_5, C_6 together. Similarly, g_2 is added to V_s, connecting together two more components C_7, C_8 with C_5. And finally, g_3 is added to connect C_4 with the others. As such, the whole process terminates with V_s = {g_1, g_2, g_4} to connect all the components together.

4. EXPERIMENTAL RESULTS

We performed experiments on two datasets, the Yeast cell cycle microarray data [7] and the Arabidopsis cold-response microarray data [37], with two GO categories, molecular
function (MF) and biological process (BP) adopted as domain knowledge [10]. Note that the proportion of functional annotations, esp. BP, are not complete for most genomes. For example, only 37% of the Arabidopsis genes have annotations with experimental evidence [15], leading to challenges in using GO in gene clustering. We tested the performance of SKB on both rich annotated genome (yeast) and sparsely annotated genome (Arabidopsis). The results showed that the performance of SKB is robust.

The Yeast microarray dataset contains 2884 genes whose expressions are altered during cell cycle under 17 time points. Among them, 1649 genes have at least one MF annotation, and 1973 genes have at least one BP annotation. The Arabidopsis microarray dataset contains 2255 cold-response genes under 6 time points with cold treatment at 4°C. Among them, 1142 genes have at least one MF annotation, and 975 genes have at least one BP annotation. Only non-IEA GO annotations are used in the experiments.

### 4.1 Results of SKB

In SKB phase 1 of hierarchical biclustering, we mined biclusters and inter-cluster relationships, requiring that each bicluster has at least 50 genes and 5 conditions. For the Yeast dataset, 114 biclusters were mined with average 56.7 genes on each bicluster. For the Arabidopsis dataset, 44 biclusters were mined with average 55.0 genes on each bicluster. By drawing the hierarchical biclustering trees, the inter-cluster relationships are revealed.

In SKB phase 2 of bicluster graph generation, we set distance thresholds $\sigma$ for each gene expression dataset under different GO categories, such that for any edge in a bicluster graph, the false discovery rate of its distance value is smaller than 0.01. Hence, all the edges in each bicluster graph are statistically sound. With the given distance thresholds, we convert all of the 114 Yeast and 44 Arabidopsis biclusters into bicluster graphs. Note that all of the un-annotated genes are removed from the bicluster graphs before further processing.

In SKB phase 3 of new gene identification, a set of new genes are identified for each bicluster (see details in Table 5). On average, for Yeast biclusters, 7 and 10 new genes were found per cluster with MF and BP annotations. For Arabidopsis biclusters, 3 and 5 new genes were found per cluster with MF and BP annotations.

### 4.2 Performance evaluation

The performance of SKB was tested on two ways: 1) a GO based measure on the identification of functionally related missing genes, and 2) network connectivity on the identification of hidden key genes. In addition, the biological significance of the new genes were tested with functional characterization.

**Effectiveness of SKB on the identification of functionally related missing genes**

To qualitatively show the biological significance of a gene cluster graph, we define the Functionally Overall Similarity measure (FOS) in Equation 3:

$$FOS(G) = 1 - \frac{1}{2|V|}|\sum_{i,j=1}^{V} f(g_i, g_j)| - \frac{1}{2|V|(|V| - 1)}$$  (3)
Figure 6: Case study. Red colored are clustered genes, and yellow colored are new genes. To study why the different processes co-express under cold treatment and how these processes may work together, SKB introduces 7 new genes, in which MPK3 acts as a central node to bridge C1 with other components, providing clues for how the processes may be functioning together in cold conditions.

The score of FOS indicates the overall functional similarity of a bicluster using GO annotation. The value close to 1 means that, for most of the genes in cluster graph $G$, the overall pair-wise similarity of the connected is high, or the similarity value along the path to connect the gene pair is high. In contrast, the value of $FOS(G)$ close to 0 means most of the genes are dissimilar from each other.

To evaluate the effectiveness of SKB in identifying functionally related missing genes, we compared SKB against QHB on FOS scores of the resulting biclusters. We chose QHB to compare because that i) QHB performs well in mining biclusters with consistent trends and trends with similar degrees of fluctuations; ii) such comparison shows the ef-
fectiveness of the new genes since SKB generates the same biclusters as QHB. To evaluate SKB’s strategy to identify new genes, we also compared SKB with a random selection process, in which we randomly choose the same number of new genes as SKB does from the candidate gene set and add them to each bicluster. Experimental results in Table 5 show that the FOS scores of the biclusters are considerably increased compared to QHB and the random selection. For example, by introducing an average of 5.1 new genes to the Arabidopsis biclusters with BP annotations, comparing to QHB and the random selection, the FOS score is increased by 2.67 and 2.51 times, respectively.

Furthermore, in order to verify the performance of SKB independent to GO annotation, we adopt 11 known cold related motifs [3] and test their existence in the Arabidopsis cold-response genes. Experiment shows that 80.0% new genes share the same cold motifs with the biclustered genes with BP annotation. In contrast, randomly picked new genes only share 47.2% same cold motifs with the biclustered genes. The results correlate well with our biological knowledge that the new genes should have the same cold motifs as the biclustered genes that they connect, since all the biclustered genes are co-expressed under cold treatment.

Graph connectivity on the identification of hidden key genes

We further tested whether the selection of new genes is significantly different from random with a leave-k-out approach. The SKB phase 2 on the Arabidopsis dataset generated a bicluster graph with 975 nodes. In the graph, we identified and then deleted 255 key genes with the minimal cut algorithm [16]. In addition, we removed 56 marginal genes with the lowest node betweenness scores. This leave-k-out process separated the whole network into 25 disconnected subnetworks. Let the candidate new genes be all the removed genes, we tested whether the SKB phase 3 is able to re-discover any bridge genes to connect the disconnected components in the resulting graph. The results showed that SKB added 13 missing genes, all of which are the bridge genes, and thus reduced the number of disconnected subgraphs from 25 to 7. In contrast, a random addition of the same amount (13) of new genes from the candidate gene set results in 22.0 disconnected subgraphs (the average of 1000 repetitions, with standard deviation at 1.1). The results indicate that the SKB results are far from random (t-test p-value < 0.001).

4.3 Functional characterization of the missing genes in Arabidopsis

The functional characterization of the new genes for Arabidopsis shows in Figure 5 that there is an enrichment of TFs in the new genes compared to all the genes in the microarray data. TFs are regulate transcriptions and are generally located in nucleus. A TF usually does not co-express substantially with its target genes. Hence, such TFs are usually missing from biclusters based on expression data. The significant and consistent rise of TFs in the new gene set from the highly noisy background on all the three GO categories indicates that SKB is biologically sound.

A case study on Arabidopsis dataset is shown in Figure 6. An Arabidopsis bicluster has 26 annotated genes (red). Using BP annotation, the cluster graph is formed with one large component belonging to the cold regulatory process, denoted as $C_1$, and the other 10 small components. These 11 components are disconnected because they are conceptually dissimilar based on GO annotation. To study why the different processes co-express under cold treatment and how these processes may work together, SKB introduces 7 new genes (yellow) that can connect the biclustered genes, in which MPK3 acts as a central node to bridge $C_1$ with other components, providing clues for how the processes may be functioning together in cold conditions.

5. DISCUSSION AND CONCLUSIONS

In this paper, we introduce SKB, a method to effectively improve the biological significance of gene expression biclustering. The delineation of functional relationships and incorporation of such missing genes may help biologists to discover biological processes that are important in a given study and provide clues for how the processes may function together. Experimental results show that SKB can reveal the inter- and intra-bicluster relationships efficiently and accurately, and greatly improve the biological significance of the biclusters. Note that, SKB achieves better performance on yeast than on Arabidopsis, probably because of the different ratio of Gene Ontology annotations of the two organisms.

It may be argued that instead of designing a new algorithm to add new genes, a connected cluster graph may be easily obtained by simply increasing the distance threshold. To address this question, we increased the distance thresholds until the same number of components as those connected by SKB was reached. To reach this point, the distance threshold was increased on average from 0.122 to 0.665 for the Arabidopsis dataset with BP annotations. Consequently, the resultant cluster graphs are cliques or clique-like dense graphs, in which the edges are not statistically significant and numerous weak gene associations are present. By contrast, the distance threshold in SKB is constantly low and all the edges are guaranteed to be statistically significant, revealing that SKB can not only recognize functional modules in biclusters, but also increase the intra-cluster relationships and filter weak gene associations effectively. Therefore, SKB is significantly better on graph representation and noise control than simply increasing the distance threshold.

It may also be argued that the effort to introduce new gene can be replaced by simply joining two related biclusters because the biclusters are highly overlapped and new genes found for one bicluster may already exist in another bicluster. To address this question, we tested how many new genes identified by SKB were already included in the biclusters. For Arabidopsis, only 12 out of 50 new genes are overlapped with biclustered genes using BP annotation. This indicates that the majority of the new genes cannot be captured by joining related biclusters. Therefore, SKB is more effective than joining existing biclusters because it finds considerably more new genes.

In the framework of SKB, we chose QHB because it is effective in identifying hierarchical inter-bicluster relationships, suitable for our skeleton bicluster expectation. With simple extension, SKB can adopt any biclustering method as its first step. We will extend the capability of SKB in two ways in the future. First, heterogeneous domain knowledge, including protein motifs, cis-elements, sequence alignment and others, will be utilized to improve the bicluster performance in addition to the GO annotations. Second, SKB will
be further extended to enhance the performance of a gene search engine by grouping functionally similar genes, recommending additional genes that are not in the search result, and visualizing the search result graphically.

**Competing interests**

none.

**Authors’ contributions**

JC conceived the approach. SJ and XT implemented the algorithm and conducted all the experiments. All wrote the manuscript.

**Acknowledgement**

This research was supported by Chemical Sciences, Geosciences and Biosciences Division, Office of Basic Energy Sciences, Office of Science, U.S. Department of Energy (award number DE-FG02-91ER20021).

**REFERENCES**


[28] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and


