Exploiting Domain Knowledge to Improve Biological Significance of Biclusters with Key Missing Genes

Jin Chen #1, Liping Ji #2, Wynne Hsu *3, Kian-Lee Tan *4, Seung Y. Rhee #5

#Department of Plant Biology, Carnegie Institution for Science
260 Panama Street, Stanford, CA 94305
*Department of Computer Science, National University of Singapore
Law Link, Singapore 117590

1,2{chenjin, jiliping}@stanford.edu 3,4{whsu,tankl}@comp.nus.edu.sg 5rhee@acoma.stanford.edu

***Abstract*— In an era of increasingly complex biological datasets, one of the key steps in gene functional analysis comes from clustering genes based on co-expression. Biclustering algorithms can identify gene clusters with local co-expressed patterns, which are more likely to define genes functioning together than global clustering methods. However, these algorithms are not effective in uncovering gene regulatory networks because the mined biclusters lack genes that may be critical in the function but may not be co-expressed with the clustered genes. In this paper, we introduce a biclustering method called *SKeleton Biclustering* (SKB), which builds high quality biclusters from microarray data, creates relationships among the biclustered genes based on Gene Ontology annotations, and identifies genes that are missing in the biclusters. SKB thus defines inter-bicluster and intra-bicluster functional relationships. 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 provides clues for how the processes may be functioning together. Experimental results show that, with SKB, the biological significance of the biclusters is considerably improved.**

I. Introduction

Gene expression clustering is one of the key steps in gene functional analysis. Genes that have similar patterns of expression can be clustered together, and are considered to be functionally related [1]. The gene clusters can thus help formulate new hypotheses from high-throughput experimental data. In many cellular processes, many genes are usually co-expressed only under certain experimental conditions, but behave almost independently under other conditions [2]. Hence, discovering local co-expressed patterns becomes the key in uncovering genetic pathways that are not apparent when clustered globally. Therefore, biclustering algorithms [3, 4] have been proposed to capture a subset of genes that may function together under a specific condition by simultaneously clustering both the genes and experimental conditions together. Bicluster identification is essential in revealing gene regulatory networks [3].

While existing biclustering algorithms can detect biclusters with local co-expressed patterns, they are not effective in uncovering gene regulatory networks or genetic pathways, mainly due to the following two reasons.

First, it is the relationship among clusters and the relationship among the genes within a cluster rather than the sets of clusters that contribute towards better interpretation of the overall picture of genetic pathways and gene regulatory networks [5]. Although hierarchical biclustering algorithms [4] can reveal the inter-bicluster (relationships among biclusters) relationships, existing biclustering algorithms [3, 4], to our knowledge, cannot identify the intra-bicluster relationships. Methods to investigate how the genes are functionally associated within a bicluster and to improve biclustering performance by reinforcing the gene functional associations have not yet been developed.

Second, biclustering methods solely based on microarray data would inevitably miss certain functionally related genes, no matter how well the algorithms are tuned. This is because: 1) not all of the functionally related genes necessarily co-express significantly; 2) unavoidable experimental noises or missing values may occur in microarray data. Hence, certain genes cannot be grouped into a bicluster, although they are functionally similar to the biclustered genes. Without these missing genes, the ability to associate functions among the biclustered genes would be substantially reduced, resulting in weak overall intra-bicluster relationships. For example, transcription factors (TF) are usually not co-expressed substantially with their target genes, and the functional associations among their target genes may not be made if a key TF is missing from the bicluster. The gene pathways or regulatory networks built upon such biclusters would then be incomplete and disconnected. Therefore, a systematic analysis on each bicluster to identify new genes as false negatives would help to better interpret the intra-bicluster relationships, hence improve the biological significance of the biclusters. However, to the best of our knowledge, no algorithm to date exists, which can identify such missing genes and missing gene associations.

In this paper, we propose a novel biclustering algorithm called *SKeleton Biclustering* (SKB) to mine biclusters. 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 biclustered genes, in order to uncover the otherwise hidden gene associations within biclusters. Figure 1 shows the framework of SKB. Overall, SKB has three phases. In phase 1, a hierarchical biclustering method is introduced to generate

1084-4627/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDE.2009.205

biclusters from microarray data. The inter-bicluster relationships are revealed by a hierarchical tree. In phase 2, each bicluster is converted into either one connected graph or a set of separated subgraphs by linking the genes that have similar biological features. In the graph, the distance between any two genes is measured based on biological domain knowledge, including Gene Ontology (GO) annotation [6], cis-elements and others. Connections of the genes within a bicluster based on their functional similarity define the initial intra-bicluster relationships. In phase 3, a graph mining method is proposed to efficiently add new genes, if any, to each bicluster, to reconnect individual 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 [8] show that SKB can build a clear structure of the inter-bicluster and intra-bicluster relationships. Consequently, the mined biclusters have significantly higher biological meaning than existing methods.

Comparing with previous methods, we have made contributions in the following four aspects.

1. New Concept of Biclustering. We are the first to propose the concept to mine the skeleton (*i.e.,* inter-cluster and intra-cluster relationships) of biclusters, rather than to mine the gene biclusters only.

2. New Methodology to Employ Domain Knowledge. It is generally difficult to integrate mining data with domain knowledge in an unsupervised learning model. In this paper, we employ the domain knowledge as an independent data space, and map the mining results from mining data to such a space to further improve the mining performance.

3. Improved Biological Significance. We are the first to improve the biological significance of biclusters by reinforcing the gene functional associations.

4. Efficient Algorithm. Both gene biclustering and missing gene identification are computationally expensive. We introduce an appropriate framework to achieve the goal effectively.

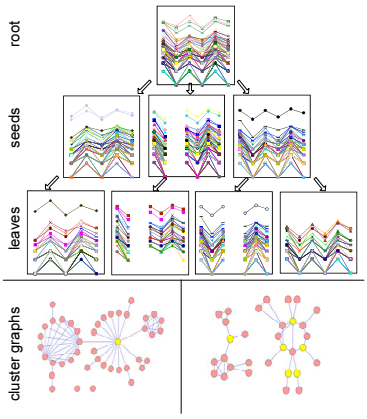

Fig. 1. Framework of SKB. In the upper part, SKB mines biclusters from microarray data and reveals the inter-bicluster relationship with a hierarchical tree. In the tree, each node is a bicluster, shown with gene expression changing trend. In the lower part, SKB reveals the intra-bicluster relationships by identifying missing genes (yellow colored) that could bridge the functionally distinct genes in each bicluster, so as to uncover the otherwise hidden gene associations.

II. Skeleton Biclustering

SKB contains three phases: hierarchical biclustering, cluster graph generation and new gene identification.

PHASE 1, biclustering and inter-cluster relationship identification. We employ an efficient top-down hierarchical biclustering algorithm QHB [4] as the first phase of SKB to build high quality biclusters from microarray data. QHB delivers biclusters with consistent trends and produces a hierarchical tree to reveal the inter-bicluster relationships.

PHASE 2, gene distance measure and cluster graph generation. In general, we consider two genes to be functionally associated if they share at least one biological feature. To model the biological information in different gene sets, we need to take into account that not all the GO terms are equally informative in terms of the biological domain they describe [9]. Therefore, for each gene set, we assign specific weights to the GO terms as it was done in [10].

$$w(t) = \frac{freq(t) + \sum_{d \in D_t} freq(d)}{N} \quad (1)$$

where $freq(x)$ denotes the number of occurrences of GO term x in a given gene set; D_t is the set of descendants of t in T; and N is the total number of term occurrences.

Given two GO terms t_a and t_b, we adopt an enriched GO term comparison method based on the distance to the nearest common ancestor term t_{ab} [11] to assign a term similarity score for t_a and t_b, denoted as $sim(t_a, t_b)$.

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

Let T_{g_i} and T_{g_j} be the set of GO terms annotated to gene g_i and g_j respectively, and let $w(t_{ab})$ be an adjustment factor using the shared information content to avoid the shallow annotation problem [12], we define the gene distance measure $d(g_i, g_j)$ as follows:

$$d(g_i, g_j) = \min_{t_a \in T_{g_i}, t_b \in T_{g_j}} ((1 - sim(t_a, t_b)) \times w(t_{ab})) \quad (3)$$

For a bicluster S, in order to reveal the biological relationships among the biclustered genes, we convert S into a cluster graph. We define a cluster graph and its component as follows:

Definition 2.1: **Cluster Graph.** A cluster graph $G(V, E)$ is an undirected graph obtained from a bicluster S, such that each vertex in V represents a unique gene in S, and edge $e(g_i, g_j) \in E$ if and only if $d(g_i, g_j) < \sigma$, where $g_i,\ g_j \in V$ and σ is a predefined gene distance threshold, $\sigma > 0$.

Definition 2.2: **Component.** Component $C_i(V_i, E_i)$ is a subgraph of cluster graph $G(V, E)$, such that there exists at least one path between any pair of vertices in V_i, and no path exists between any vertex in V_i and any vertex in $V - V_i$.

Based on these definitions, G is either a connected graph or a set of components, depending on the value of σ. Note that genes in a component are usually considered to be functionally related. It is also possible that genes in different components

are functionally related. Biclustering methods solely based on microarray data would inevitably miss certain function-related genes and thus weaken the overall gene relationship. Therefore, if a key gene is missing from a bicluster, the resulting cluster graph may be broken and the functionally associated genes may be separated into different components.

PHASE 3, identifying new genes to improve biological significance of biclusters. Using appropriate new genes to reinforce the biclustered gene functional associations is an effective way of refining biological significance of the biclusters. However, incorporating inappropriate genes will adversely affect the biclustered gene associations. Here, we formulated three rules to discover the appropriate new genes.

Rule 1. Distance along new genes bounded by σ. To increase the overall function similarity of a bicluster by drawing its components closer with new genes, the path between any two components along a new gene should not be longer than the predefined gene distance threshold σ and only one new gene is allowed in the path.

Rule 2. Connect all the connectable components. If any two components can be connected with a set of new genes, one such new gene should be included. Therefore, all the connectable components will be connected. Let I be the whole set of genes in a microarray dataset and S be a bicluster mined from I. Component $C_i(V_i, E_i)$ and $C_j(V_j, E_j)$ are *connectable* if and only if there exists at least one vertex v_x in $I - S$ such that $min\left(d(v_i, v_x) + d(v_j, v_x)\right) < \sigma$, where $v_i \in V_i$ and $v_j \in V_j$.

Rule 3. Minimum number of new genes. Only the set of genes with the minimum size that satisfy the first two rules will be included. First, we notice that owing to incomplete biological knowledge in GO annotation, introducing too many new gene 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.

In summary, the strategy to identify new gene is to search for the smallest set of new genes to connect all the connectable components in a cluster graph with the distance value along the path smaller than σ. Mathematically, given a bicluster S and its cluster graph $G(V, E)$, the new gene set V_x in $I-S$ satisfies two conditions: i) $\forall$ connectable components C_i & $C_j \in G$, $\exists v_x \in V_x$ such that $min(d(v_i, v_x) + d(v_j, v_x)) < \sigma$ ($v_i \in V_i$ and $v_j \in V_j$); ii) V_x is the smallest set satisfying condition i.

We now illustrate the computational difficulties for achieving the goal. A straight-forward method of finding the minimum set of new genes V_x is to test every subset of $I - S$ exhaustively with its size increasing from 1 to $|I - S|$. The computational time is exponential to the size of the search space $|I - S|$. We introduce an efficient algorithm to improve the speed in finding the minimum set of new genes V_x, mainly by i) finding a local optimized V_x with a heuristic algorithm, ii) further optimizing the results with a randomized process. First, for a given bicluster S, instead of testing every new gene candidate in $I - S$, we reduce the search space by considering only the genes that can connect at least two components, denoted as Λ. We also compute the number of components that g can connect, denoted as $N(g)$. Second, for a new gene candidate set Λ, we compute the upper bound κ of the new gene set size with a locally optimized algorithm by sorting the candidates based on $N(g)$ and finding the new genes serially. In detail, we iteratively move gene g from Λ to V_x if g can connect to the most number of components in $\mathbb{C}$. With g, we compose a new component C_x and the number of components is reduced to $|\mathbb{C}| - N(g) + 1$. Then we recompute the $N(g)$ for all the genes in Λ. The iteration stops when Λ is empty. Third, we iteratively call a randomized process to identify the new genes. In the iteration, if the size of the new gene search space is reduced to be smaller than a constant value (in our experiment, 20), we call the exhaustive search to identify the new genes. Otherwise, as long as the search space is large, the probability of randomly hitting the right new genes is pretty small. Hence, we randomly remove $|\Lambda|/(2\times\kappa)$ candidates and perform two recursive calls. In summary, all the three steps of SKB phase 3 have polynomial time complexity.

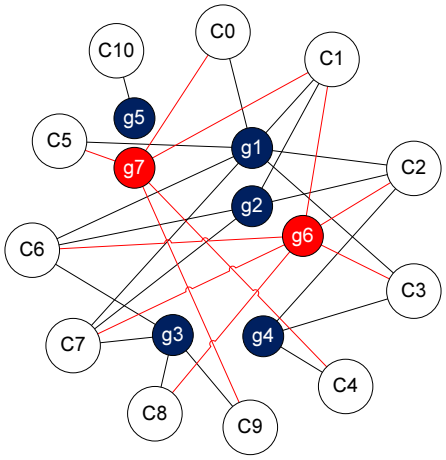

Fig. 2. Illustrative example for new gene identification. $C_0 \ldots C_{10}$ are 11 components in cluster graph G, $g_1 \ldots g_7$ are new gene candidates. The red ones are the new genes identified by SKB.

In the illustrative example in Figure 2, a cluster graph G has 11 components $C_0 \ldots C_{10}$ and 7 new gene candidates $g_1 \ldots g_7$. For simplicity, the distance value for every edge is set to 0. We aim to find the minimum number of new genes to connect all the connectable components. We first filter g_5 since it cannot bridge any component. Next, we obtain the upper bound of new gene set size by compute the values of $N(g)$ for all of the candidates and select new genes serially. g_1 is firstly selected because it can connect 7 components. Then the values of $N(g)$ for the rest of the candidates are recalculated. g_3 and g_7 are able to connect to the most number of the components, thus one of them is selected. In the end, a new gene set $\{g_1, g_3, g_4\}$ could be identified. Therefore, the upper bound κ is 3. Finally,

by iteratively calling the randomized process for a number of times, the final new gene set $\{g_6, g_7\}$ is identified, with which, all the connectable components are connected.

Note that we fix the new gene search space to be all the genes in the microarray data. The search space can be easily changed to any ancestor in the hierarchical biclustering tree. Such extension, depending on user's needs, can find new genes that not only satisfy our searching strategy but are also loosely co-expressed with the biclustered genes. In addition, the search space can be easily changed to include genes from the same organism that are not in the microarray data.

III. Experiments

To test the performance of SKB, we performed experiments on the Yeast cell cycle microarray data[1] in [7] and the Arabidopsis cold-response microarray data collected from NASC, NCBI, TAIR websites[2] and [8], with two GO categories, molecular function and biological process, adopted as domain knowledge. The Yeast microarray dataset contains 2884 Yeast genes whose expression is altered during cell cycle under 17 time points [7]. The Arabidopsis microarray dataset contains 2255 Arabidopsis cold-response genes under 14 time points with cold treatment at 4°C.

In SKB phase 1, 109 Yeast biclusters were mined, each bicluster had on average 56.2 genes; and 59 Arabidopsis biclusters were mined, each bicluster had on average 35.2 genes. In phase 2, we set distance thresholds σ such that its false discovery rate equals 0.01. Note that all of the unannotated genes are removed from the cluster graphs before further processing. In phase 3, a set of new genes are retrieved for each bicluster. On average, 7.3 and 9.6 new Yeast genes were found per cluster with function and process annotations respectively. 1.8 and 3.2 new Arabidopsis genes were found per cluster with function and process annotations respectively.

To test the efficiency of SKB, we compared SKB with the exhaustive search. Experiment shows that SKB is much faster than the exhaustive search, with 100- to 100k-fold speed up.

To qualitatively show the biological significance of the biclusters mined with SKB, we introduce two measures, the Graph Connectivity Score (GCS) and the Overall functional Similarity Score (OSS), to test the cluster performance from the topology and functional perspectives. Let n be the number of components in G and $\Phi(g_i, g_j)$ be the set of paths connecting g_i and g_j, GCS is defined by testing the overall topological connectivity, and OSS is defined by testing the overall functional similarity of a bicluster using GO annotation.

We evaluated the performance of SKB by comparing the GCS and OSS scores with those in the biclusters mined with QHB and a random selection process, in which we randomly choose the same number of new genes as SKB does. Figure 3 shows that by introducing new genes to the Arabidopsis biclusters with process annotations, the mean values of GCS and OSS increase significantly comparing to QHB and the random selection. Similar trends were shown in Yeast dataset or using function annotations.

[1] *http://arep.med.harvard.edu/biclustering*

[2] *http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl; http://www.ncbi.nlm.nih.gov; http://www.arabidopsis.org*

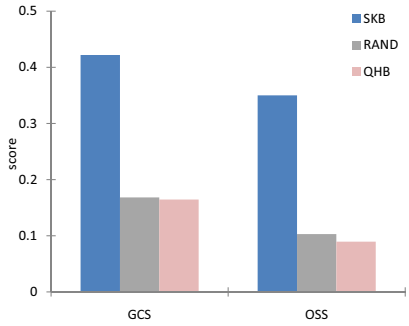

Fig. 3. GCS and OSS scores of Arabidopsis clod-related gene biclusters improved with biological process annotations.

$$GCS(G) = \frac{\sum_{i=1}^{n} |C_i|^2}{\left(\sum_{i=1}^{n} |C_i|\right)^2} \tag{4}$$

$$OSS(G) = 1 - \frac{\sum_{i,j=1}^{|V|} f(g_i, g_j)}{\frac{1}{2} \times |V| \times (|V| - 1)} \tag{5}$$

$$f(g_i, g_j) = \begin{cases} d(g_i, g_j) & \text{if } < g_i, g_j > \in E \\ \min\limits_{\phi \in \Phi(g_i, g_j)} \sum\limits_{(u,v) \in \phi} d(u, v) & \text{if } < g_i, g_j > \notin E \& \Phi \neq \emptyset \\ 1 & \text{otherwise} \end{cases}$$

IV. Conclusion

In this paper, we introduce a biclustering method SKB for gene expression clustering. 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.

References

[1] J. Jun, S. Chung, and D. McLeod, "Subspace clustering of microarray data based on domain transformation," *VLDB Workshop on Data Mining on Bioinformatics*, 2006.

[2] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *PNAS*, pp. 14 863–14 868, 1998.

[3] H. Wang, W. Wang, J. Yang, and P. Yu, "Clustering by pattern similarity in large data sets," *SIGMOD*, pp. 394–405, 2002.

[4] L. Ji, K. Mock, and K. Tan, "Quick hierarchical biclustering on microarray gene expression data," *BIBE*, pp. 110–120, 2006.

[5] Y. Lazebnik, "Can a biologist fix a radio?–or, what i learned while studying apoptosis," *Cancer Cell*, vol. 2, no. 3, pp. 179–182, 2002.

[6] M. Ashburner, C. Ball, J. Blake, *et al.*, "Gene ontology: tool for the unification of biology," *Nat Genet*, vol. 25, no. 1, pp. 25–29, 2000.

[7] Y. Cheng and G. Church, "Biclustering of expression data," *ISMB*, pp. 93–103, 2000.

[8] J. Vogel, D. Zarka, H. Van Buskirk, *et al.*, "Roles of the cbf2 and zat12 transcription factors in configuring the low temperature transcriptome of arabidopsis," *Plant Journal*, vol. 41, pp. 195–211, 2005.

[9] S. Y. Rhee, V. Wood, K. Dolinski, and S. Draghici, "Use and misuse of the gene ontology annotations," *Nature Reviews Genetics*, 2008.

[10] P. Lord, R. Stevens, *et al.*, "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275–1283, 2002.

[11] D. Lin, "An information-theoretic definition of similarity," *ICML*, pp. 296–304, 1998.

[12] J. Sevilla, V. Segura, *et al.*, "Correlation between gene expression and go semantic similarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 4, pp. 330–338, 2005.