PARALLEL BICLUSTERING OF GENES WITH COHERENT EVOLUTIONS: ALGORITHM AND BIOLOGICAL SIGNIFICANCE OF THE BICLUSTERS

A. H. Tewfik, A. B. Tchagang, and L. Vertatschitsch. *{tewfik, tcha0003}@umn.edu*
Electrical and Computer Engineering, University of Minnesota

ABSTRACT

Uncovering genetic pathways is equivalent to finding clusters of genes with expression levels that evolve coherently under subsets of conditions. This can be done by applying a biclustering procedure to gene expression data. Given a microarray data set with *M* genes and *N* conditions, we define a bicluster with coherent evolution as a subset of genes with expression levels that are non-decreasing as a function of a particular ordered subset of conditions. We propose a new biclustering procedure that identifies all biclusters with a specified number of *K* conditions in parallel with *O(MK)* complexity. Unlike almost all prior biclustering techniques, the proposed approach is guaranteed to find all biclusters with a specified minimum numbers of genes and conditions in the data set. All of the biclusters it identifies have no imperfection, i.e., the evolutions of the genes in each bicluster will be coherent across all conditions in the bicluster. Furthermore, the complexity of the proposed approach is lower than that of prior approaches. We also discuss the biological significance of the biclusters computed by the algorithm for a set of yeast gene microarray data.

1. INTRODUCTION

One of the major goals of gene expression data analysis is to uncover genetic pathways, i.e., chains of genetic interactions. For example, a researcher may be interested in identifying the genes that contribute to a disease. This task is difficult because subgroups of genes display similar activation patterns *only* under certain experimental conditions. Genes that are co-regulated or co-expressed under a subset of conditions will behave differently under other conditions. Finding genetic pathways therefore could be aided by identifying clusters of genes that are co-expressed under subsets of conditions as opposed to all conditions. A high degree of correlation between the activity levels of subsets of genes under subsets of conditions does not of course necessarily imply causality relations. Further biological analysis would be required to find actual genetic pathways.

Gene expression data is typically arranged in an *M by N* data matrix $A=[a_{mn}]$, with rows corresponding to genes and columns to experimental conditions. The $(m, n)^{th}$ entry of the gene expression matrix represents the expression level of the gene corresponding to row *m* under the specific condition corresponding to column *n*. By simultaneously clustering the rows and columns of the gene expression matrix, one can identify candidate subsets of conditions that may be associated with cellular processes that exhibit themselves only or subsets of genes that potentially play a role in a given biological process. Biological analysis and experimentation could then confirm the biological significance of the candidate subsets.

Cheng and Church were the first to apply biclustering to DNA microarray data analysis [1]. They introduced the term biclustering to denote simultaneous row-column clustering of gene expression data. Many other approaches were proposed thereafter in the literature, e.g., [2]-[8]. The reader is referred to [9] for an extensive development and comparison of prior approaches.

Most of these previous techniques search for one or two types of biclusters among four that have been identified in the literature [9]: biclusters with constant values, biclusters with constant values on rows or columns, biclusters with coherent values, and biclusters with coherent evolution. Most previous techniques are also greedy and will miss meaningful biclusters. Many of these pioneering approaches used a cost function to define biclusters. In many cases, the cost function will measure the square deviation from the sum of the mean value of expression levels in the entire bicluster, and the mean values of expression levels along each row and column in the bicluster.

Our objective here is to develop a biclustering algorithm that is able to discover *all* biclusters with coherent evolutions in a given data set in a timely manner. The proposed biclustering algorithm approach is different from previous ones in several ways. First, the proposed approach will identify all valid perfect biclusters with coherent evolutions and any given number of conditions. Unlike our prior work [8] or those of [1]-[6], none of the biclusters it identifies will have imperfections, i.e., one or more gene that does not behave coherently with the remaining genes for one or more condition. Secondly, the proposed approach uses basic linear algebra and arithmetic tools and avoids the need for heuristic cost functions of prior approaches that can miss some pertinent biclusters. Third, the complexity of our approach is lower than that of prior biclustering techniques, including that of our prior approach [7].

2. PROBLEM STATEMENT AND DISCUSSION

As mentioned earlier, our goal here is to identify all biclusters of genes that behave coherently across a number of conditions. Biclusters with coherent evolutions identify groups of genes that are co-regulated up or down coherently across a subset of conditions *with no reference to their actual expression levels.* In this paper, we shall define a bicluster with coherent evolution as a subset of genes that have expression levels that are non-decreasing as a function of an ordered subset of conditions. Note that, according to our definition, another subset of genes with expression levels that are decreasing as a function of the ordered subset of conditions in a bicluster that satisfies our definition will constitute a *different* bicluster. This follows from the fact that their expression levels will be non-decreasing as a function of the reverse of the given ordered subset of conditions. If desired, the two biclusters can be merged in a post-processing step to identify all genes that are co-regulated up or down coherently across the subset of conditions.

To avoid the identification of possibly trivial biclusters, we allow the user to pre-specify the minimum size of a valid bicluster, that is the minimum number G_{min} of genes and conditions K_{min} that appear in a given bicluster.

The reader will immediately notice that traditional clustering techniques cannot be used to solve the problem of identifying biclusters with coherent evolutions. Straightforward application of traditional clustering techniques would require that we look for meaningful patterns on arbitrary subsets of conditions amongst several conditions and a large number of genes. This is difficult and computationally very complex as illustrated in Figs 1 and 2. These two figures refer to a subset of results of the analysis of 2882 yeast genes under 17 conditions using the proposed technique. Fig. 1 displays the expression levels of a subset of 9 genes under all conditions. It is very difficult to identify any coherent evolution from Fig. 1. Yet, the proposed approach is able to identify a subset of 4 conditions under which the gene expression levels evolve coherently, as shown in Fig. 2. We will have more to say about the data set and the results we obtained in Section 4.

The second challenge that we encounter when trying to apply traditional clustering techniques to the problem of identifying biclusters with coherent evolutions is that, by definition, we are not interested in the actual expression levels of the genes. We are solely interested in how they are co-regulated across subsets of conditions. This implies that traditional distance measures cannot be used to identify patterns or clusters. Clearly, the expression levels shown in Fig. 2 are not close in a Euclidean, Manhattan or other distance measure.

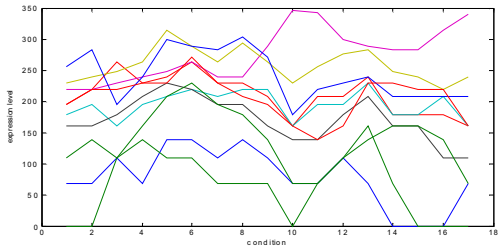

Fig. 1. Expression levels of a subset of 9 genes across all conditions in the yeast microarray data. Different lines correspond to the expression levels of different genes.

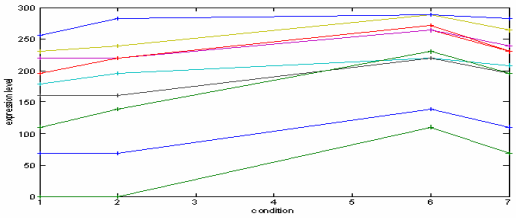

Fig. 2. Expression levels of the subset of 9 genes in Fig. 1 across 4 conditions: *{1, 2, 6, 7}*, in the yeast microarray data. Different lines correspond to the expression levels of different genes.

3. THE ALGORITHM

Our proposed algorithm consists of two steps: a pre-processing step followed by a bicluster identification step.

The pre-processing step in particular, starts with a data conditioning routine that strictly speaking is not part of the proposed algorithm. Its main purpose is to deal with the noise in the DNA microarray data as well as missing values.

The actual bicluster identification step consists of two sub-steps. For all valid numbers K of conditions, where $K \geq K_{min}$, and K_{min} is the pre-specified minimum number of conditions in a valid bicluster, the procedure will enumerate all combinations of K conditions from the given N conditions in the DNA microarray data that could potentially appear in a valid bicluster. For each subset of K conditions, it then uses a row sort procedure that allows us to focus on the coherent evolutions of gene expression levels, rather than the raw or processed expression levels. The output of this step is a matrix that contains the rank of each of the K conditions for each row (gene) when the expression levels of each gene are ordered in a non-decreasing manner. We shall refer to this matrix as the *condition rank matrix* and will use it as the input to the main bicluster identification routine. Finally, the main bicluster identification routine identifies all valid coherent evolution patterns involving all genes and a set of K conditions *simultaneously* through a fast row sorting procedure. Note that this allows the algorithm to identify all the possible valid biclusters *without* an exhaustive enumeration of all possible $K!$ permutations of the K conditions. The procedure will also yield biclusters of genes where a subset of genes are coherently up-regulated and another subset coherently down-regulated across the K conditions.

a. Data pre-processing

Many techniques to recover missing expression level values have been developed in the literature, e.g., [10]. Obviously, the choice of technique for dealing with missing values will impact the output of the bicluster identification approach. In our experiments, we have typically removed any gene from consideration when examining subsets of conditions for which the gene expression level is missing.

Several techniques have also been proposed in the literature to deal with noise in DNA microarray data, e.g. [11]-[14]. In our work, we have relied on two approaches based on a quantization of the gene expression values. We have used approaches similar to those of [15] wherein a clustering technique, such as K-means or fuzzy c-means, is used to cluster the expression levels across the entire DNA microarray data set, or on a gene by gene basis. Each gene expression level is then replaced by the codeword corresponding to the cluster to which it belongs prior to analysis with the bicluster identification procedure.

b. Bicluster identification

Let us now discuss the three main sub-steps of the biclustering procedure.

(i). Enumeration of subsets of conditions

Suppose that the user determines that all valid biclusters must have K_{min} or more conditions. Let K be the possible number of conditions in a valid bicluster, where $K_{min} \leq K \leq N$. The algorithm proceeds in one or two ways to enumerate all possible subsets of conditions that may appear in a valid bicluster.

The first approach consists of using any algorithm, such as the one in [16] or [17], to enumerate all possible $C(N,K)$,

combinations of K conditions from the given N conditions in the DNA microarray data, where $C(N,K)=N!/((N-K)!\,K!)$, regardless of their likelihood of appearing in a valid bicluster.

The main drawback of this enumeration approach is its potentially high complexity. For example, with the yeast gene expression data that we discuss later in the paper, $N=17$. For $K=8$, the algorithm will generate 24,310 subsets of 8 conditions to be tested. A preferred approach consists of generating the subsets recursively. Specifically, we note that a subset of 8 conditions that leads to valid biclusters will include 2 subsets of 4 conditions each. Therefore, we can recursively form subsets of 2^m conditions by concatenating two non overlapping subsets of 2^{m-1} conditions that actually generated valid biclusters.

Subsets with K conditions where K is not a power of 2, can be similarly computed by merging several subsets of 2^{k_i} conditions that actually generated valid bicluster, where $1 \le i \le I$ and K equals the sum of 2^{k_i}. In particular, the values of the powers k_i can be computed by expressing K in binary form.

(ii). Data sorting

For each given subset of K conditions, the algorithm proceeds to sort the rows of a submatrix of the gene expression matrix $\boldsymbol{A}$ defined above. Let vector $\boldsymbol{n} = [n_1\ n_2\ \cdots\ n_K]$ denote a row vector with entries n_i equal to the column positions of the K conditions in the given subset of K conditions. The algorithm proceeds to sort the entries in each row of $\boldsymbol{A(:,n)}$, where ":" stands for all rows, in non-decreasing order of gene expression level. For each row (gene), the algorithm sorts the expression levels of the gene, from smallest to largest, across the K conditions corresponding to the columns indexed by $\boldsymbol{n}$. The indices of the sorted list are then stored in the corresponding row of the condition rank matrix $\boldsymbol{B}$. In particular, let $\boldsymbol{y}$ be the sorted list corresponding to $\boldsymbol{x= [}x(1)\ x(2)\ \cdots x(K)\boldsymbol{]}$ the lth row of $\boldsymbol{A(:,n)}$, i.e., $\boldsymbol{x=A(l,n)}$. Further, let $\boldsymbol{m =[}m_1\ m_2\ \cdots m_K\boldsymbol{]}$ be the vector of indices of the sorted list $\boldsymbol{y}$. Then, $\boldsymbol{x(m) = y}$ and $\boldsymbol{B(l,n) = m,}$ where $\boldsymbol{x(m)=[}x(m_1)\ x(m_2)\ \cdots x(m_K)\boldsymbol{]}$. Thus, a {3} in cell (4,5) of the gene condition rank matrix $\boldsymbol{B}$ indicates that the expression level for gene 4 under the condition corresponding to n_3 is the 5th smallest (possibly quantized) expression level for gene 4 across all conditions corresponding to $\boldsymbol{n}$.

(iii). Bicluster identification

Note that the rows of the condition rank matrix B exhibit all possible ordered patterns of the K conditions across all rows. Hence, to identify all possible ordered sets of conditions that are exhibited by rows of A, we need to find all distinct patterns in B. This can be done by sorting the rows of matrix B and identifying "edges" between patterns.

The algorithm therefore begins by sorting the rows of B according to the entries along its columns, starting with the first column and proceeding to the last. The output of this row sorting procedure is a matrix C in which rows are grouped by similarity.

Note that the algorithm also needs to keep track of the original index of the sorted rows to correctly identify the genes in any given bicluster. Let *mindex* be the column vector indicating the original position of all rows in C. The index in the kth row of *mindex* indicates the position of the kth row of C in B.

To identify the unique patterns in B, we apply a simple differencing procedure to the columns of C. specifically; we subtract from each row the values in the preceding row, column by column. A non-zero output in any row indicates the position of a distinct pattern. Hence, we can identify the row indices of all distinct patterns by looking for rows with any non-zero element. Let *dindex* be the column vector indicating the row position of the unique patterns in C. The difference between two consecutive entries in *dindex* gives us the number of genes that display a given pattern. The exact identity of these genes is found by picking the rows of *mindex* with indices starting at an entry of *dindex* and proceeding to the next consecutive entry minus one.

Denote by $[m_{1i}\, m_{2i}\ \cdots\ m_{Mi}]$ the list of genes that exhibit the ith unique pattern in matrix C. The pair $([m_{1i}\, m_{2i}\ \cdots\ m_{Mi}], [n_1\ n_2\ \cdots\ n_K])$ is a valid bicluster if $G \ge G_{min}$. Hence, the algorithm discards any unique ordering of the K conditions that is not displayed by at least the desired minimum number of genes, G_{min}.

The complexity of the proposed approach can be shown to be $O(MK)$, which is lower than that of prior techniques, even though most prior approaches are not guaranteed to find all valid biclusters with coherent evolutions. For example, the approaches of [1] and [2] have complexities of $O(MN(M+N)I)$ and $O((M+N)^2 I)$ respectively to identify I biclusters.

4. RESULTS AND THEIR BIOLOGICAL SIGNIFICANCE

We conclude the paper by analyzing the yeast gene microarray data that can be found at [18]. The data contains the expression levels of 2884 genes under 17 conditions. We eliminated two genes from the discussion below because they had missing data. We also eliminated three other genes from the original data set. These genes contained all zeros as expression levels. Obviously any way you order the conditions, the expression levels of these genes stay constant. These genes will therefore be picked up in every bicluster because as mentioned above, biclusters are seeking genes whose expression level either *stay constant* or increase across an ordered subgroup of conditions. These 3 genes are actually known to be unclassified.

The partial analysis results that we present here for the yeast data were obtained by first quantizing the expression data using a dictionary with 50 codewords determined by the k-means algorithm. We sought all biclusters with 3 or more genes and 11 to 17 conditions. Because of the large number of biclusters found, we will present here a few illustrative results that will help the reader grasp the magnitude of the problem and the nature of the results produced by the algorithm.

A preliminary assessment of the biological significance of the biclusters was performed by using functional categories from the Comprehensive Yeast Genome Database [19]-[20]. The database categorizes yeast genes into fine groupings. The results that we present here however utilize a level of classification that divides the yeast genome into 19 groups based on the function of the protein the specified gene codes. The annotation system the CYGD utilizes is called FunCat, for functional classification catalog. More information on this can be found in [21]. Because of the large number of biclusters that we obtained, we discuss here the biclusters that we identified with 12 to 16 conditions. The analysis of these biclusters is representative of what we have seen so far. It also illustrates the complexity of the additional investigations that must be performed on the biclusters once they have been identified. Table 1 provides a preliminary biological significance

analysis of the biclusters with 12 or more conditions. The first row of Table 1 lists how many biclusters were found. Rows two through five show how many biclusters belong to one of 4 mutually exclusive categories. The second row shows how many of those biclusters contained genes that were all annotated under the same function. An example of a bicluster in this grouping would be three genes that all produce proteins whose main purpose is metabolism. The third row displays how many of the biclusters picked up only genes that were unclassified. The fourth row lists the number of biclusters that contained genes annotated to the same function as well as unclassified genes. The final row just shows how many biclusters detected genes that are functionally unclassified.

Note that the biclusters picked up a low number of completely functionally annotated sets. This was to be expected since the initial data contained many unclassified genes. Interestingly, the algorithm picks up many biclusters that are completely comprised of functionally unclassified genes. Analysis of these unknown genes that are co-regulating, and possibly analysis of the conditions under which they do so, could lead to further classification of the *S. cervisiae* genome and are the object of current biological investigations.

Another unexpected result was the number of biclusters that contained "mixed" data. The appearance of such biclusters led us to pose several questions that we are attempting to answer in collaboration with researchers in the biological sciences. The genes in these mixed biclusters showed patterns of coherent evolution but did not fall necessarily in the same functional category.

The presence of these biclusters may be indicative of the fact that co-regulated genes do not necessarily belong to the same functional category. On the other hand, it may indicate that these genes have other unknown functions or functions that were not captured in the annotation we used. It is also possible that the expression levels of certain genes that belong to a given functional category affect those of some other genes that belong to a different functional category.

Many of the mixed biclusters are of biological interest because they contain genes that either belong to a single functional category or are unclassified. Current investigations are attempting to determine whether the unclassified genes in these biclusters do actually belong to the same functional category as the others. With colleagues, we are examining the literature to identify the theorized functions of many of the unclassified genes that appear in mixed biclusters or biclusters with unclassified genes. We are also studying alternative gene annotations sources, such as GO-Slim [22], to answer some of the questions that we posed here.

5. REFERENCES

[1] Y. Cheng, G.M. Church, "Biclustering of Expression Data", In *Proc. ISMB'00*, pages 93-103. AAAI Press, 2000.

[2] J. Yang, H. Wang, W. Wang, and P.S. Yu, "Enhanced Biclustering on Expression Data," *Proc. Third IEEE Conf. Bioinformatics and Bioeng,* pp. 321-327, 2003.

[3] A. Tanay, R. Sharan, and R. Shamir, "Discovering Statistically Significant Biclusters in Gene Expression Data," *Bioinformatics*,vol. 18, pp. S136-S144, 2002.

[4] G. Getz, E. Levine, E. Domany, "Coupled Two-way Clustering Analysis of Microarray Data, *Proc. Natl. Acad. Sci.* USA, 97(22): 12079-84, 2000.

[5] L. Lazzeroni , A. Owen, "Plaid Models for Gene Expression Data", *Statistica Sinica*, 12: 61-86, 2002.A.

[6] Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, "Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem," *Proc. Sixth Int'l Conf. Computational Biology (RECOMB '02)*, pp. 49-57, 2002.

[7] A. H Tewfik and A. B. Tchagang, "Biclustering of DNA Microarray Data with Early Pruning" *Proc. 2005 IEEE International Conference on Acoustics, Speech and Signal Processing,* Philadelphia, PA, March 2005.

[8] A. B. Tchagang and A. H. Tewfik, "Robust Biclustering Algorithm (Roba) For DNA Microarray Data Analysis," *XIII European Signal Proc. Conf. (EUSIPCO2005),* Antalya, Turkey, September 2005.

[9] S. C. Madeira, A. L. Oliveira, "Biclustering Algorithms for Biological Data Analysis: A Survey", *IEEE Transactions on computational Biology and Bioinformatics*, Vol. 1, No. 1, Jan-March 2004.

[10] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. Altman, "Missing Value Estimation for DNA Microarrays", *Bioinformatics* 17(2001), 1-6.

[11] H. Vikalo, A. Hassibi and B. Hassibi, "Optimal Estimation Of Gene Expression Levels In Microarrays," in *Proc. IEEE International Workshop On Genomic Signal Processing and Statistics*, 2005.

[12] R. Albert and H. Othmer, "The topology of the regulatory interactions predict the expression pattern of the segment polarity genes in *Drosophila melanogaster*, *J. Theor. Biol.*, vol. 223, pp. 1-18, 2003.

[13] T. Chung and S. Kim, "Quantization of Genome-wide Expression Data," in *Proc. IEEE International Workshop On Genomic Signal Processing and Statistics*, 2005.

[14] L.L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis,* New York: Addison-Wesley Publishing Co., 1990.

[15] A. Hassibi and H. Vikalo, "A Probabilistic Model For Inherent Noise And Systematic Errors Of Microarrays" in *Proc. IEEE International Workshop* On Genomic Signal Processing and Statistics, 2005.

[16] P. J. Chase, "Combinations of m out of n objects (Algorithm 382)," *Commun. ACM*, vol. 13, p. 368, 1970.

[17] C. M. Liu and D. T. Tang, "Enumerating combinations of rn out of n objects (Algorithm 452)," *Commun. ACM*, vol. 16, p. 485, 1973.

[18] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, and G. Church. Yeast micro data set. [Online]. Available: http://arep.med.harvard.edu/biclustering.

[19] U. Güldener, et al, "CYGD: the Comprehensive Yeast Genome Database," *Nucleic Acids Research*, 33 Database issue, pp. D364 - 8, January 1, 2005.

[20] Munich Information Center for Protein Sequences (MIPS) and GSF-National Research Center for Environment and Health, "Comprehensive Yeast Genome Database," 2002. Available online: http://mips.gsf.de/genre/proj/yeast/ (visited July 21, 2005).

[21] A. Ruepp, et al, "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucleic Acids*, Res 32, pp. 5539-5545, 2004.

[22] R. Balakrishnan et al, "Saccharomyces Genome Database," Available Online: http://www.yeastgenome.org/

Number of Conditions	12	13	14	15	16
Number of Biclusters	32,589	5,377	597	39	1
Number of Functionally Defined Biclusters	2,386 (7.3%)	360 (6.7%)	29 (4.9%)	1 (2.6%)	0
Biclusters Composed Entirely of Unclassified Genes	2,766 (8.5%)	592 (11.0%)	66 (11.1%)	2 (51.3%)	0
Biclusters with Unclassified Genes and Genes of One Function	14,581 (44.7%)	2,683 (49.9%)	345 (57.8%)	28 (71.8%)	1
Biclusters with Genes of Mixed Annotation	12,856 (39.4%)	1,742 (32.4%)	157 (26.3%)	8 (20.5%)	0
Biclusters containing Unclassified Genes	27,620	4,593	512	34	1

Table1