BIOLOGICAL EVALUATION OF BICLUSTERING ALGORITHMS USING GENE ONTOLOGY AND CHIP-CHIP DATA

[1]Alain B. Tchagang, [2]Ahmed H. Tewfik, and [1,3]Panayiotis V. Benos

[1]Dept. of Computational Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA
[2]Dept. of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA
[3]Dept. of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA

ABSTRACT

In this paper, we propose a new framework for assessing the biological significance of the outputs of any biclustering algorithm. The framework relies on the *p*-value computed by a *Fisher's exact test* on a 2x2 contingency table derived from *Gene Ontology* (GO) enrichment level and *chromatin immunoprecipitation* (ChIP) data enrichment level. We illustrate the framework using our recently published *Robust Biclustering Algorithm* (RoBA), the *Cheng and Church (CC) algorithm*, and a well-defined set of yeast cell cycle gene expression data and ChIP-chip data. Our evaluation also shows that the biclusters identified by RoBA are biologically more homogeneous than the ones identified by the Cheng and Church (CC) algorithm.

Index Terms— Biclustering, gene expression, chromatin immunoprecipitation, genetic pathways, transcription factors

1. INTRODUCTION

Biclustering algorithms are widely used to analyze gene expression data. The term refers to a distinct class of algorithms that perform simultaneous row-column clustering and they are able to identify local behaviors of the dataset analyzed. Given a gene expression matrix, biclustering algorithms are capable of identifying subsets of genes that behave coherently across subsets of experimental conditions, time points, or tissue samples. By simultaneously clustering the rows and columns of the gene expression matrix, one can identify candidate subsets of conditions that may be associated with specific cellular processes that exhibit themselves only on 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.

Since the publication of the first biclustering algorithm by Cheng and Church for gene expression data analysis [1], several other biclustering algorithms have been developed [2, 3, 4, 5, 6]. **Fig. 1** for example shows the statistics of *Pubmed*'s articles dealing with the development and usage of bi-clustering algorithms in biological applications.

We also refer the reader to [7] for a review on biclustering algorithms and their biological applications. Although these algorithms have been used to identify some interesting patterns in genomics, their results are difficult to evaluate. This is a common problem of all clustering algorithms, partly because the rules for such evaluation are not well established and evaluation datasets are scarce [3].

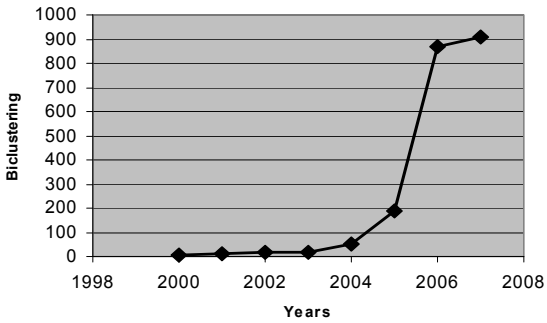

Figure 1: *Pubmed*'s statistics on bi-clustering since 2000

This paper presents a new procedure for assessing the biological significance of the outputs of any biclustering algorithm, based on the *Fisher's exact test*. To illustrate the evaluation procedure, our recently published biclustering algorithm ***RoBA*** (Robust Biclustering Algorithm) [6] and *the Cheng and Church (CC)* algorithm [1] are used to analyze microarray data to identify statistically significant (bi)clusters of co-expressed genes. Subsequently, these biclusters are evaluated using two types of external data: *Gene Ontology* (***GO***) *process* categories and *chromatin immunoprecipitation* (***ChIP***) data. GO classifies gene products according to their associated biological processes, cellular components, and molecular functions in a species-independent manner (www.geneontology.org). ChIP is a well-established methodology used to investigate interactions between transcription factor (***TF***) proteins and their genomic DNA targets *in vivo* [8].

Ideally, a good biclustering algorithm will identify genes with similar expression patterns. These co-expressed genes are expected to be regulated by the same TFs. In addition, genes that are co-expressed frequently participate in the same biological pathways. In both cases, if the

identified biclusters are biologically meaningful, then their corresponding sets of genes should be enriched and annotated under the same GO categories, involved in the same biological pathways, or regulated by the same transcription factors.

2. MATERIALS

2.1. Yeast Expression Dataset

The gene expression data used in this study derive from the yeast *Saccharomyces cerevisiae* cell cycle dataset [9]. It consists of 2,884 genes sampled in 17 conditions (time points) (0 – 160 minutes, 10 minutes interval sampling), covering nearly two full cell cycles. The relative mRNA abundance values (percentage of the mRNA for a gene compared to all mRNAs) were transformed by scaling and taking the logarithm $x \rightarrow 100 \log_{10}(105x)$ and the result was a matrix of integers in the range between 0 and 600. (The transformation does not apply to the values 0 and null element -1) [1].

2.2. Yeast Chromatin Immunoprecipitation Dataset

The ChIP-chip dataset we used here derives from the Lee *et al.* study [10] on the association (or not) of 113 yeast TFs with the promoters of every gene in the yeast genome. It corresponds to a 6270 x 113 dataset, where the rows represent the yeast genes, and the columns the 113 TFs. The entries of the matrix correspond to the *p*-value of the association of a given TF to the promoter of the corresponding gene. In this study, we used a *p*-value threshold of 1.0e-03 to discretize the data as in [10]. In other words, TF-gene association values of 1.0e-03 or less correspond to 1 (*i.e.*, the gene is regulated by the corresponding TF) otherwise correspond to 0.

3. METHODS

Biclusters with *coherent behavior* (see below) were identified from the set of gene expression data described above using RoBA, our recently published algorithm [6] and the Cheng and Church (CC) algorithm [1]. Biological assessment of the biclusters was then performed using GO annotations (www.yeastgenome.org) and a published ChIP-chip dataset [10].

3.1. Robust Biclustering Algorithm (RoBA)

Given an $N \times M$ gene expression matrix $A = [a_{nm}]$ with set of rows or genes $G = \{g_1, \ldots, g_N\}$ and set of experimental conditions or columns $C = \{c_1, \ldots, c_M\}$, a bicluster $B = [b_{ij}]$ is any submatrix of A whose entries follow a specific pattern. Biclusters can be classified into five distinct categories: (a) constant biclusters, (b) biclusters with constant values along rows, (c) biclusters with constant values on columns, (d) biclusters with coherent values, and biclusters with coherent evolutions [7]. Biclusters with coherent evolutions are unique in that they focus on the behavior of the genes across subsets of conditions. Our focus here, and indeed that of most researchers, is on finding subsets of genes that are upregulated or downregulated across subsets of conditions irrespective of their actual expression values. Finding such biclusters provides a starting point for elucidating genetic pathways.

RoBA, the biclustering algorithm we presented in [6], is capable of extracting any type of biclusters mentioned above from a given set of gene expression data in a timely manner. Since the focus of this paper is on the biological evaluation of the biclusters and not the algorithm itself, we refer the reader to [6] for the algorithmic details.

3.2. Biological Assessment of Biclusters

The biological role of the genes in the biclusters was assessed using two statistical criteria: GO process enrichment level and TF-gene association enrichment level.

3.2.1. Gene Ontology (GO) assessment

The significance of a GO process enrichment in a given bicluster can be computed using the *Fisher's exact test* on a 2x2 contingency table. The *p*-value of this test can be calculated using the hypergeometric distribution. Let N denote the total number of unique genes on the microarray, $K(g)=K$ the total number of genes that are in the GO process category (g) of interest, and $I(B)=I$ the number of genes assigned to a bicluster B. Then, the probability of seeing n or more genes in the intersection of the GO category of interest g and the bicluster B is:

$$p = \sum_{i=n}^{I} \frac{\binom{K}{i}\binom{N-K}{I-i}}{\binom{N}{I}} \qquad (1)$$

3.2.2. ChIP-chip assessment

The same idea applies to the statistical analysis of the ChIP-chip data enrichment. Now, K is the total number of genes a certain TF is associated with. Then, the *p*-value of observing n or more genes in bicluster B being associated with the given TF of interest can be computed using (1) above.

4. RESULTS

We ran *RoBA* on the above gene expression dataset. We first filtered out genes with missing values and genes whose expression level did not change significantly during the time course. RoBA yielded four dominant biclusters in the dataset under the whole time course

experiments (**Fig. 2**). Each bicluster had more than 40 genes under all 17 conditions (time points). Note that, since we are dealing with time series gene expression, the sequential order of the time course is very important.

The biological significance of these four dominant biclusters was evaluated using biological (external) data as described in the *Methods* section.

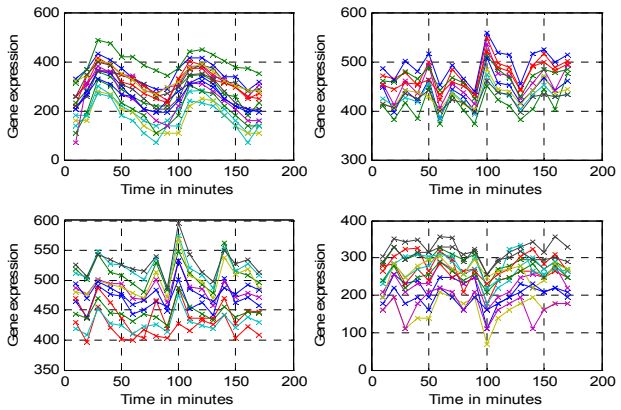

Figure 2: The four dominant bicluster profiles

Biological assessment through statistical analysis of the GO category overrepresentation in the four biclusters was assessed using the online yeast GO Term Finder tool (www.yeastgenome.org). This tool calculates *p*-values of the observed data using the hypergeometric distribution (**Eq. 1**) with Bonferroni correction for multiple testing. All categories depicted as statistically overrepresented in the identified biclusters have some relation with the cell cycle (**Table 1**). In fact, the most significant GO term processes identified for bicluster I is "cell cycle", with a *p*-value = 4.2e-12. Furthermore, 94% of the genes in bicluster II participated in some "cellular process" (*p*-value=1.6e-04). Although more than 20% of the genes in bicluster IV are of unknown function, another 50% of its genes were involved in "regulation of metabolic process" (*p*-value = 7.9e-05.) as calculated by the "*FuncAssociate*" function (http://llama.med.harvard.edu/cgi/func/funcassociate). We note that the genes with unknown functions in bicluster IV may also be involved in "regulation of metabolic process", although further experimentation is required to validate this.

Table 1: Bicluster evaluation using Gene Ontology

Biclusters	No. of genes	Top GO term	percentage of genes in category	P-value
I	48	Cell cycle	45.8%	4.2e-12
		DNA replication	31.2%	5.9e-13
II	51	Translation	58.0%	2.1e-15
		Cellular process	94.0%	1.6e-04
III	43	Translation	62.0%	1.1e-15
		Biosynthetic proc.	67.4%	4.0e-11
IV*, **	47	Reg. of Meta. proc	50.0%	7.9e-05

* More than 20% genes of unknown function, ***FuncAssociate* results

We also evaluated our biclustering algorithm using the published yeast ChIP-chip dataset on 113 TFs. Each of the 113 transcription factors in this dataset was tested for target over-representation in each of the four biclusters using the Fisher's exact test described above. The results are shown on **Table 2**.

Table 2: Bicluster evaluation using ChIP data

	Biclusters			
TFs	**I**	**II**	**III**	**IV**
ASH1	2.0e-02			
CIN5				1.0e-02
FHL1		3.3e-19	1.1e-22	
FKH1				6.0e-02
FKH2	4.0e-02		3.1e-02	3.9e-02
GAT3		9.1e-02	7.0e-02	
HAP4	7.9e-02			
HIR1		2.2e-02		
HIR2		1.1e-02		
MBP1	3.8e-20			
MCM1				2.4e-02
MSN1		9.4e-04		
NDD1				2.9e-02
PDR1		8.9e-02		
RAP1		5.4e-10	4.2e-11	
RLM1			2.7e-02	
SKN7			2.9e-02	
SMP1		9.1e-02		
SRD1		8.9e-02		
SWI4	1.1e-15		1.1e-02	
SWI6	7.3e-25		2.3e-02	
YAP5			6.2e-06	

Transcription factors MBP1, SWI4 and SWI6 have significantly overrepresented number of target genes in bicluster I. Notably, these three TFs participate in the two major transcription complexes regulating G1/S transition: MBF (MBP1/SWI6 heterodimer) and SBF (SWI4/SWI6 heterodimer) [11]. So bicluster I may contain the genes that participate in the G1/S phase transition. Although not so dramatic, the *p*-values for TFs FKH1, FKH2 and MCM1 in bicluster IV are also significant. These TFs are known to be key regulators of the G2/M transition [12]. Interestingly, the FKH2/MCM1 regulated genes need NDD1 to proceed with the G2/M transition, since FKH2 and MCM1 remain bound to their targets throughout the cell cycle [12].

Biclusters II and III are the most similar of the four in terms of gene expression profile (**Fig. 2**), GO category association (both are associated with "translation", **Table 1**) and TF-gene association (both contain overrepresented motifs for FHL1 and RAP1 TFs, **Table 2**). FHL1 TF is known to regulate exclusively the expression of ribosomal protein genes and that its function depends heavily on RAP1 [13]. Their target ribosomal protein genes are important components of the cell cycle and constitute a large component of the GO category "translation", common to both biclusters. Despite their similarities, we note that MSN1 TF (generally involved in response to nutrient

limitation) appears to have overrepresented target genes only in bicluster II, whereas transcription factor YAP5 appears to have overrepresentation target genes only in bicluster III. These are also the only biclusters with overrepresented motifs for these TFs. It could be that these TFs regulate the majority of the genes in the corresponding biclusters that seem to belong to the "cellular process" (94% of bicluster II genes) and "biosynthetic process" (67.4% of bicluster III genes) GO categories (**Table 2**). Since not much is known about the target genes of these two TFs, further experimentation is required to explain their role in yeast cell cycle.

We also compared the evaluation results of the RoBA identified biclusters to that of the biclusters identified by the CC algorithm (**Fig. 3**). For consistency, we only picked the four best (best = low mean squared residue) CC's biclusters under 17 conditions as mentioned on their website (http://arep.med.harvard.edu/biclustering/). We performed analysis of biological significance on the four top biclusters identified by each technique using GO annotations and TF-gene association as described above. A particular bicluster was significant if the *p-value* of the GO category or TF-gene association was smaller than a specified threshold. As shown on **Fig. 3**, the biclusters identified by RoBA are biologically more meaningful than the ones identified by the CC algorithm, using the above two criteria.

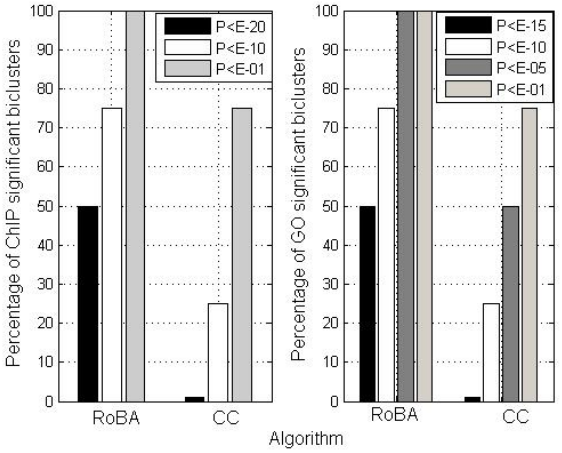

Figure 3: Comparison of RoBA with Cheng and Church's algorithm

5. CONCLUSION

In this paper, we evaluated the biological significance of our recently published algorithm (RoBA) using two main criteria: GO process category enrichment level and ChIP-chip data enrichment level. (Bi)clustering was performed on a well-defined yeast cell cycle gene expression dataset. GO enrichment and TF enrichment analysis showed that our algorithm is able to identify statistically significant and biologically important patterns from a given set of gene expression data. The framework we used in this study to evaluate the biological significance of biclusters can be used as a tool to test and to evaluate any type of clustering algorithms in the future.

Acknowledgements. This work was supported by NIH grants 1R01LM009657-01 and NO1 AI-50018. PVB was also supported by NIH grant 1R01LM007994-01.

6. REFERENCES

[1] Y. Cheng and G. M. Church, Biclustering of expression data, Proc Int Conf Intell Syst Mol Biol, 8 (2000), pp. 93-103.

[2] A. Ben-Dor, B. Chor, R. Karp and Z. Yakhini, Discovering local structure in gene expression data: the order-preserving submatrix problem, J Comput Biol, 10 (2003), pp. 373-84.

[3] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele and E. Zitzler, A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics, 22 (2006), pp. 1122-9.

[4] A. H. Tewfik and A. B. Tchagang, Parallel identification of gene biclusters with coherent evolutions, IEEE Transactions on Signal Processing, 54 (2006), pp. 2408-2417.

[5] W. Ahmad and K. Ashfaq, cHawk: an Efficient Biclustering Algorithm, based on Bipartite Graph Crossing Minimization, Workshop on Data Mining in Bioinformatics, 33rd VLDB 2007, 2007.

[6] A. B. Tchagang and A. H. Tewfik, DNA Microarray Data Analysis: A Novel Biclustering Algorithm Approach, EURASIP Journal on Applied Signal Processing, 2006 (2006).

[7] S. C. Madeira and A. L. Oliveira, Biclustering Algorithms for Biological Data Analysis: A Survey, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1 (2004), pp. 24-45.

[8] M. J. Buck and J. D. Lieb, ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, Genomics, 83 (2004), pp. 349-60.

[9] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart and R. W. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, Mol Cell, 2 (1998), pp. 65-73.

[10] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford and R. A. Young, Transcriptional regulatory networks in Saccharomyces cerevisiae, Science, 298 (2002), pp. 799-804.

[11] V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder and P. O. Brown, Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF, Nature, 409 (2001), pp. 533-8.

[12] J. Bahler, Cell-cycle control of gene expression in budding and fission yeast, Annu Rev Genet, 39 (2005), pp. 69-94.

[13] J. T. Wade, D. B. Hall and K. Struhl, The transcription factor Ifh1 is a key regulator of yeast ribosomal protein genes, Nature, 432 (2004), pp. 1054-8.