

Biclustering Gene Expression Data by Using Iterative Genetic Algorithm

Hua-Sheng Chiu¹, Tao-Wei Huang² and Cheng-Yen Kao³

Keywords: gene expression data, transcriptional module, bicluster, genetic algorithm, BIGA

1 Introduction.

The general starting point for uncovering the regulatory network in the gene expression data is to identify the *transcriptional modules* — a set of genes that are co-regulated under certain of experimental conditions. Genes in the same transcriptional module are supposed to have the same function and are regulated by the common transcriptional factors that bind to their short DNA sequence motifs in the up-stream promoter region. Several methods have been devised to address these issues for identifying biologically meaningful condition-specific transcriptional modules, so called *biclusters*. The concept of biclusters is first introduced [1] to capture the coherence of a subset of genes and a subset of conditions in the analysis of large-scale gene expression data. FLOC (FLexible Overlapped biClustering) [3] is proposed to address serious drawbacks observed in [1]. In this study, a novel model and Biclustering by Iterative Genetic Algorithm (BIGA) are proposed for identifying significant transcriptional modules.

2 Method.

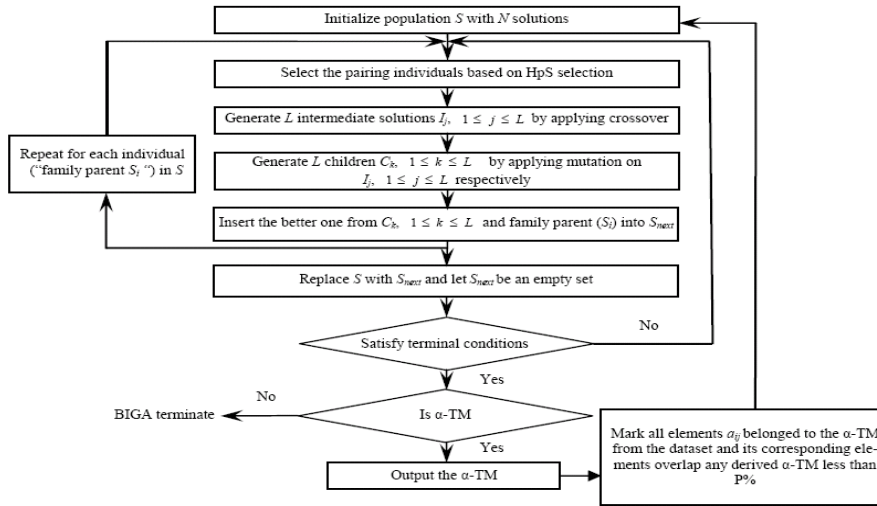


Figure 1: Overview of the proposed biclustering by iterative genetic algorithm (BIGA) approach.

¹ Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, E-mail: r91031@csie.ntu.edu.tw

² E-mail: d90016@csie.ntu.edu.tw

³ E-mail: cykao@csie.ntu.edu.tw

The Biclustering by Iterative Genetic Algorithm (BIGA) approach is proposed to identify transcriptional module (TM) in gene expression data, avoiding the intrinsic limitations of the heuristic biclustering algorithms. Every TM is composed of the gene subset and the condition subset from the original gene expression data and also possesses α significant level of correlativity requested by user. Besides, the novel fitness for a statistically significant and condition-specific clusters, i.e, the α -TM, is defined concisely for GA fitness function. In a word, the proposed approach is regarded biclustering as optimization problem in iterative manner based on the new α -TM model and novel fitness function. Figure 1 shows the main steps of the proposed BIGA.

3 Results.

Synthetic data as well as yeast cell cycle dataset [2] are used to test the proposed BIGA approach. The results are characterized by three properties — widespread, dense, and condition-specific. Finally, comparing BIGA with a number of present heuristic strategies emphasizes its superiority of capturing sharp coherent tendency among gene expression data. The expression patterns of one of the TM with 18 genes and 10 conditions generated from yeast cell cycle dataset by performing BIGA is displayed in Figure 2. It's interesting to point out that, the noisy conditions underlined in Figure 2(b) against to these genes are all filtered out while the contributive ones are left.

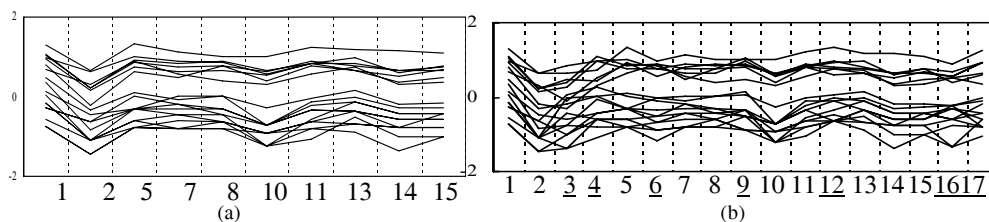


Figure 2: The expression patterns of 18 genes contained in one of the TM derived from yeast cell cycle dataset. (a) Under the specified conditions. (b) Under all conditions.

4 Conclusion.

In this study, the new bicluster model called α -TM is proposed to formulate the problem of identifying significant transcriptional modules among gene expression data based on statistical concepts. Finally, BIGA is applied to evaluate its performance in identifying the significant transcriptional modules on both the synthetic data-set and real gene expression data. Moreover, the BIGA is outperforming in detecting biclusters with higher biological significances than other existing approaches.

References

- [1] Cheng, Y., and Church, G. M. 2000. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* 8, pp. 93-103.
- [2] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9, pp. 3273-3297.
- [3] Yang, J., Wang, W., Wang, H., and Yu, P. 2002. δ -clusters: capturing subspace correlation in a large data set. *18th International Conference on Data Engineering, 2002. Proceedings.*, pp. 517-528.