

Discovering Statistically Significant Clusters by Using Genetic Algorithms in Gene Expression Data

Hwa-Sheng Chiu¹, Han-Yu Chuang¹, Huai-Kuang Tsai¹, Tao-Wei Huang¹,
Cheng-Yan Kao¹

Keywords: genetic algorithms, clustering statistically significant, gene expression data, microarray

1 Introduction.

This study presents an iterative GA (genetic algorithm) approach to find significant clusters in gene expression data. Clustering genes of similar expression patterns is useful for predicting gene functions and pathways. Many heuristic algorithms have developed to cluster genes on microarray data [1]. However, the results of most heuristics may be dominated by some predefined criteria, such as the number of clusters and initial mean points of clusters. Moreover, the present methods assign each one of genes in the dataset into some cluster while some of them are irrelevant to the interested conditions. To automatically cluster informative genes together, we regard clustering as an optimization problem of maximizing correlativity among genes in a set. Thus, an iterative GA approach is proposed to find the tightest clusters subject to our statistical definition of a significant cluster. The proposed method is applied on a yeast cell cycle dataset of 614 genes and 77 conditions [2]. It efficiently finds 21 statistically significant clusters and discards 105 genes. The experiment results show that genes of the same cluster are highly related to their mean pattern, and genes not belonging to any clusters are far away from each mean pattern. We proceed to apply the proposed method to biclustering.

2 System and method.

The proposed method works as follows. Our GA finds the cluster of the maximal fitness in the interested dataset, and removes the corresponding genes in the cluster from the dataset. Then the GA starts the whole searching process again in the remaining dataset. Until there are no any significant clusters at one run, the iterative process is terminated. In our GA, each chromosome is represented as a bit string with g bits, where g is the number of genes in current target dataset. A chromosome corresponds to a cluster of genes. If a bit of a chromosome has value of 1, its corresponding gene is included in the gene set corresponding to the chromosome. N sets of genes, subject to the constraints of a cluster, are randomly generated as the initial population. After evaluating the fitness, the sequential steps of family competition [3] and stochastic universal sampling (SUS) are run with a k -point probabilistic-elitism crossover and a bitwise uniform mutation. The GA is terminated when the following criterion is satisfied: the improvement ratio between all of the children generated and their respective family parents is less than 0.001 in five continuous generations.

■ Significant Clusters

A subset of genes can be regarded as a cluster in our method, if the average pairwise Pearson correlation (aPPC) of genes in the subset is more than the one of genes in the whole set excluding the subset. Moreover, the cluster is statistically significant if its aPPC is more than the sum of the expected pairwise Pearson correlation in the dataset and τ , where τ is the expected S.T.D. and τ is a threshold to decide the significance.

¹ Bioinfo Lab., Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan.
E-mail: {r91031, r90002, d7526010, d90016, cykao}@csie.ntu.edu.tw

■ Fitness Function

Each chromosome S_i in the population is evaluated by using the sum of all pairwise Pearson correlation of its corresponding genes, instead of using the average one. If only using the average Pearson correlation to choose a cluster, the trivial cluster of two genes will be derived. Our GA prefers chromosomes of larger sum subject to the constraints of the significant cluster. Besides, we introduce the aPPC as the weight to the fitness function, in order to get the set of quite correlative genes.

■ k-point Probabilistic-elitism Crossover

The crossover is modified from the multipoint crossover. The crossover works by randomly selecting k points in each parent and splitting them into k segments. Then it glues segments of parents to obtain offspring by the probability proportional to the fitness of the segment. If the fitness of the segment of one parent is better than the corresponding fitness of the other, the offspring would inherit the better segment more likely.

3 Result.

The proposed method is applied on a yeast cell cycle dataset with 614 genes and 77 conditions. 21 statically significant clusters are derived and 105 genes are discarded. The top-6 statically significant clusters are list in Fig. 1. The thick line is the mean pattern of this cluster and the thin line above or below to the mean pattern is 1 standard deviation from the mean pattern. As shown, the different clusters have the different expression patterns and the standard deviation in all clusters is small. We also found that the pairwise Pearson correlations between 105 discarded genes and 21 cluster mean patterns are small. The experimental results indicate that the proposed approaches can find tight and significant clusters efficiently and filter out irrelative genes effectively.

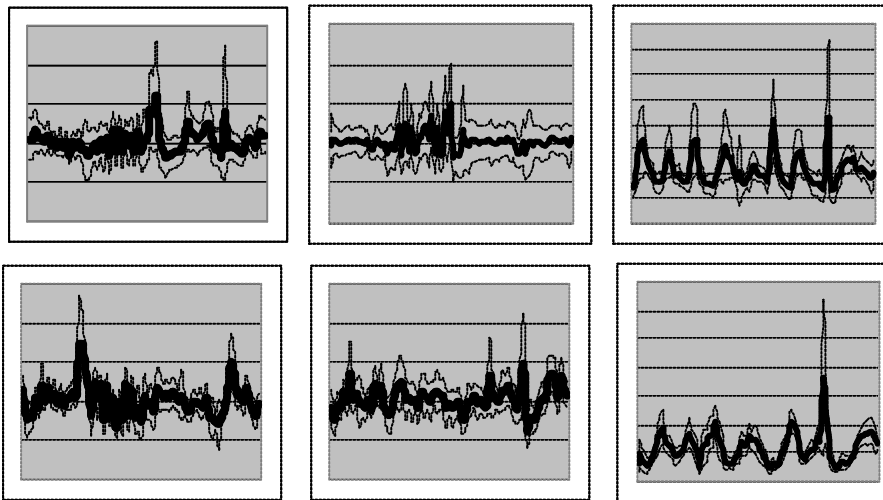


Fig. 1: The top-6 statically significant clusters. The x-axis represents experiments and the y-axis represents the expression levels. Each cluster has different expression patterns from others.

References

- [1] Tou, J. T. and Gonzalez, R.C.. 1974. *Pattern Recognition Principles*, Addison-Wesley, Reading.
- [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. *Molecular Biology of Cell*, vol. 9, no. 12, pp. 3273–3297
- [3] Yang, J. M. 2001. A Family Competition Evolutionary Approach of Global Optimization in Neural Networks, Optical Thin-film Design, and Structure-based Drug Design. *Ph. D. thesis, National Taiwan University, Taiwan*.