

# Gene Specific Co-regulation Discovery from Gene Microarray Data

Ji Zhang, Qing Liu, Kai Xu, Paulo de Souza

{ji.zhang, q.liu, kai.xu, paulo.desouza}@csiro.au

CSIRO Tasmanian ICT Centre, Hobart, TAS, Australia, 7000

## Abstract

The problem of gene specific co-regulation discovery is to identify, for a particular gene of interest, called **target gene**, its strongly co-regulated genes and the condition subsets where such strong gene co-regulations are observed. In this paper, we propose an efficient method for finding gene-specific co-regulations using genetic algorithm (GA). Compared with the previous method, our method is anticipated to achieve a notably improved efficiency by using a number of performance boost strategies.

**Keywords:** Microarray data, gene co-regulations, genetic algorithms

## 1 Introduction

DNA Microarray is an enabling technology for us to have a global view of the expression of a large number of genes. Finding gene co-regulations is one important research focus in Microarray data analysis. One interesting research problem, called *Single Gene Approach* for gene Microarray analysis [2], has been studied recently in [3]. This problem is formulated as: *for a particular gene of interest, called target gene, identify its strongly co-regulated genes and the condition subsets where such strong co-regulations are observed.* The discovered co-regulated genes and the associated condition subsets are gene specific. Gene-specific co-regulation discovery is very helpful for human to better understand and characterize the target gene. Zhang *et al.* proposed an approach for mining local gene-specific co-regulations [3]. This approach is to first find the condition subsets in which the target gene  $g$  is most significantly co-regulated with others using genetic algorithm [1] and the co-regulated genes of  $g$  are then selected from its nearest neighbors in these condition subsets. Given the target gene  $g$ , the fitness function for each condition subset  $s$  in GA is defined as the distance between  $g$  and its  $k^{th}$  nearest neighbor in  $s$ , denoted by  $D^k(g, s)$ . This method is able to find the closely co-regulated genes for the target gene and the associated condition subsets where such co-regulation occur. However, the major performance bottleneck of this approach is the fitness computation, which involves a  $k$ NN search for the target gene in each condition subset that is evaluated in the GA. The typically large number of genes in the Microarray data and the number of condition subsets evaluated in the GA lead to a slow computation. In addition, this method is limited in finding local co-regulations with explicit temporal context. For some Microarray data where no temporal context exists, the search may involve conditions spanning across the whole dimensionality of Microarray data.

## 2 Methodology

To enhance the efficiency of gene-specific co-regulation discovery, three aspects of efforts can be taken in our method, aiming to achieve a noticeable efficiency improvement, which are presented as follows:

- **Using  $k$ NN Search Table ( $k$ NN-ST) for fitness computation.** We propose  $k$ NN-ST to facilitate fitness computation in GA.  $k$ NN-ST for a target gene  $g$ , denoted as  $\mathcal{T}^g$ , is a  $d \times k$  table with

the entry  $x_{ij}$  of the table represents the  $j^{\text{th}}$  nearest neighbor of  $g$  in the  $i^{\text{th}}$  dimension ( $1 \leq i \leq d$ ,  $1 \leq j \leq k$ ).  $\mathcal{T}^g$  can be used to quickly find the lower and upper bounds of  $D^k(g, s)$ , denoted as  $D_{min}^k(g, s)$  and  $D_{max}^k(g, s)$ , respectively. These bounds can be utilized to substantially speed up fitness evaluation in the GA; the fitness of each condition subset can be approximated using the average of  $D_{min}^k(g, s)$  and  $D_{max}^k(g, s)$ . The complexity for constructing  $\mathcal{T}^g$  is only  $\mathcal{O}(dN)$ , where  $N$  and  $d$  are the number and dimensionality of genes in the data set, and the complexity for computing the fitness of a condition subset now becomes only a constant.  $k$ NN-ST of the target gene is pre-constructed before the GA-based condition subset search is performed.

- **Using shorted individual representation in GA.** In the case that the Microarray data does not have explicit temporal meaning, the gene co-regulations may occur in non-consecutive conditions that may span across the whole spectrum of conditions. Given the typical high dimensionality of Microarray data, binary representation of condition subsets, as used in [3], is inefficient in this case for crossover and mutation operations in GA because of its long length. To solve this problem, we propose the use of integer string, a more compact (yet equally effective) representation scheme, for encoding condition subsets in the GA. Due to a significant shorter length, use of integer strings as representation scheme contributes much faster crossover and mutation operations.

- **Using pruning heuristics for searching co-regulated genes.** When the top co-regulated condition subsets of the target gene have been found by GA, we can further find its top co-regulated genes in these condition subsets. To do this efficiently, we propose a heuristic to significantly reduce the number of genes that need to be evaluated. The data space for the Microarray dataset is first partitioned using a multi-dimensional grid structure (the side length of each grid cell is  $l$ ) and each gene is then assigned into one and only one cell in the grid. Let  $g$  be the target gene and  $g'$  be another gene that fall into a different cell in condition subset  $s$ , *i.e.*,  $g \in c$  and  $g' \in c'$  and  $c \neq c'$ . If  $dist(g, centroid(c')) - \frac{1}{2}l > D_{max}^k(g, s)$ , then all the genes in  $c'$  are not closely co-regulated with  $g$  and thus can be pruned away to avoid further consideration, where  $dist()$  is the function calculating distance between two data points in the grid and  $centroid(c')$  is the centroid of cell  $c'$  in the grid.

### 3 Conclusions

In this paper, we propose an enhanced technique for finding gene specific co-regulations using genetic algorithm. A number of boosting strategies are proposed to remarkably improve the efficiency of the existing method. Experimental evaluation using real-life gene Microarray data is underway to validate the efficiency enhancement our technique can achieve.

### 4 Acknowledgment

This research is supported by CSIRO Preventative-Health Flagship program and Tasmanian ICT Centre. Tasmanian ICT Centre is jointly funded by the Australian Government through the Intelligent Island Program and CSIRO. The Intelligent Island Program is administered by the Tasmanian Department of Economic Development and Tourism.

### References

- [1] Holland, J. H., *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, 1975.
- [2] Speed, T., Fridlyand, J., Yang, Y. H., and Dudoit, S., Discrimination and clustering with microarray gene expression data. *2001 Spring Meeting of International Biometric Society Eastern North American Region (ENAR'01)*, Charlotte NC, 2001.
- [3] Zhang, J., Gao, Q., and Wang, H., Discover Gene Specific Local Co-regulations Using Progressive Genetic Algorithm. *IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, 783-790, 2006.