Classifying Microarray Data using Pairwise Similarity Between Gene Profiles

Raymond Wan[1]
`rwan@kuicr.kyoto-u.ac.jp`

Åsa M. Wheelock[1]
`asa@para-docs.org`

Matthew J. Bartosiewicz[2]
`matt.bartosiewicz@sandiego.ppdi.com`

Hiroshi Mamitsuka[1]
`mami@kuicr.kyoto-u.ac.jp`

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan
[2] Microarray Facility, University of California, Davis, California, USA

Keywords: microarray expression data, gene expression profile, graph similarity, gene co-expression

1 Introduction

RNA microarrays permit the expression levels of thousands of genes to be measured simultaneously. However, interpretation of the large data sets is complex. Commonly used methods include hierarchical clustering and k-nearest-neighbor [3]. After applying these methods, clusters of co-expressed genes are produced, which can then be used as a fingerprint to identify a biological response. The basis of any gene clustering algorithm is the similarity (or dissimilarity) measure between gene expression profiles. After every possible gene-pair has been considered, hierarchical clustering recursively pairs the genes in order of increasing dissimilarity. A new score is assigned to a gene group which is obtained by aggregating the scores of the genes within the group [2]. Thus, this process resembles bottom-up creation of a binary tree, resulting in a dendrogram as depicted in Figure 1. By applying a threshold, the tree is cut horizontally to form subtrees, each representing a cluster. Hierarchical clustering has the advantage of allowing a user to visualize how genes merge to form clusters. However, individual relationships between genes within a cluster is lost once the cluster is formed. For example, gene E in the example may be more similar to gene C, but not gene D. In this poster, we focus on the pairwise nature of the hierarchical clustering process and propose a method of microarray classification using gene pairs instead of single genes or gene clusters.

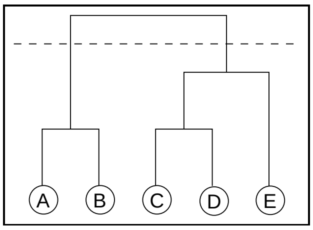

Figure 1: Dendrogram with a threshold applied (---) to create clusters. The first cluster has genes A and B while the second one has genes C, D, and E.

2 Method

Thirty-two male Swiss Webster mice (28-33 g, 5-8 wk) received an interperatoneal injection of $CdCl_2$ in saline (group 1; $n = 16$), or saline vehicle (group 2; $n = 16$). mRNA was isolated from liver tissue as previously described [1]. MWG mouse microarrays containing 10,000 50mer oligonucleotides were hybridized with the RNA, and the resulting data was normalized using print-tip loess and analyzed as described below. In contrast to the commonly used clustering methods, we stop short of cluster formation and retain relationships between genes, as shown in Figure 2. Instead of constructing a tree, we built a more general undirected graph which permits cycles. Each gene is a graph node, and an edge is added between all gene-pairs whose distance is less than the threshold.

Our method is then applied as follows for the purpose of classifying a test set using an initial training set. Two separate graphs are constructed for each of the two conditions (control and $CdCl_2$ treated) of the training set using the Euclidean distance with a fixed threshold τ. If $\tau = \infty$, then every node is connected to every other node in both graphs. The similarity between these two graphs is calculated by examining the overlap in edges since the two sets of vertices are equivalent. Of particular interest is when an edge appears in one graph but not in the other. These pairs of genes together represent a means for identifying the condition of a microarray data set within the test set.

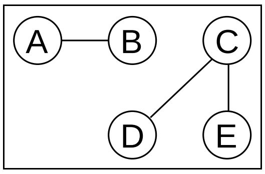

Figure 2: Graph of pair-wise similarities.

3 Results

Preliminary experiments were conducted on a 2.4 GHz AMD Opteron (dual processor) with 16 GB RAM and 1,024 KB cache. For both the control and the treatment groups, the dimensions of the data used for graph construction was $10,003$ unique gene profiles by 16 arrays. There were 28 duplicate gene profiles that were removed. Construction of both graphs and reporting on the level of overlap of the two sets of edges required less than 5 minutes. These results are shown in Table 1 and Figure 3. The number of gene pairs that exist solely in the control and treatment reaches a maximum of ~4.3% at a threshold of ~100 (Figure 3).

τ	Neither	Control	Treatment	Both
10	98.7	0.5	0.4	0.4
25	90.4	2.3	2.1	5.2
50	75.3	3.6	4.1	17.0
100	56.0	4.3	4.3	35.4
200	35.9	4.3	4.1	55.6

Table 1: Overlap of edges, represented as percentages of total possible edges.

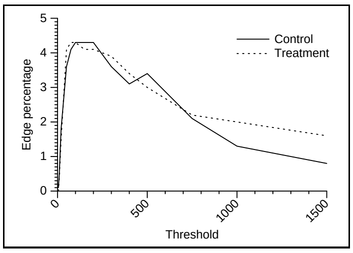

Figure 3: Distribution of edges in one graph, but not the other, for varying values of τ.

4 Discussions

This poster has proposed a method of classification which constructs graphs of relationships of mRNA expression between gene-pairs. As future work, the effectiveness of our technique is being validated with a test set and a list of co-expressed genes which have been determined by the transcription factor binding sites located in their respective regulatory domains. Also, we would like to extend our method to hypergraphs, where an edge can connect more than two nodes. This would allow classification using groups of 3 or more genes.

Acknowledgements Å. M. W. was supported by a Japan Society for the Promotion of Science (JSPS) fellowship.

References

[1] M. J. Bartosiewicz, D. Jenkins, S. Penn, J. Emery, and A. Buckpitt. Unique gene expression patterns in liver and kidney associated with exposure to chemical toxicants. *J. Pharmacol. Exp. Ther.*, 297(3):895–905, June 2001.

[2] M. J. L. de Hoon, S. Imoto, J. Nolan, and S. Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

[3] E. Wit and J. McClure. *Statistics for Microarrays.* John Wiley & Sons Ltd., 2004.