# Classifying Microarray Data using Pairwise Similarity Between Gene Profiles

**Raymond Wan**[1] **Åsa M. Wheelock**[1] **Matthew J. Bartosiewicz**[2] **Hiroshi Mamitsuka**[1]

rwan@kuicr.kyoto-u.ac.jp asa@para-docs.org matt.bartosiewicz@sandiego.ppdi.com 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

## Abstract

*A method for classifying microarray data is introduced which is based on similarities between gene profiles. We found that around 4.3% of the gene-pairs were unique to a particular condition. Our aim is to use this work as the basis for two other projects.*

## 1. Introduction

RNA microarrays allow the expression levels of thousands of genes to be measured simultaneously. Common methods for analyzing this data include hierarchical clustering and k-nearest-neighbor [Wit and McClure, 2004]. The result of either is a set of clusters of co-expressed genes.

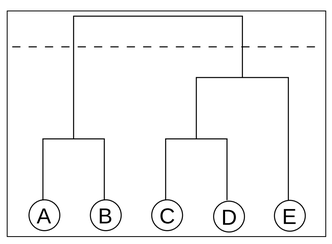


**Figure 1:** *Dendrogram with a threshold applied (---) to create two clusters.*

Similarity (or dissimilarity) forms the basis of any gene clustering algorithm. Hierarchical clustering pairs genes in order of decreasing similarity. Each gene cluster is assigned a score based on the relationships within the group [de Hoon et al., 2004]. This process resembles bottom-up creation of a binary tree, as shown in Figure 1.

Cluster formation assumes that each gene is related to every other gene within its cluster. Also, each gene is unrelated to every other gene in all of the other clusters. While this categorization of each gene is ideal for visualizing gene relationships, an alternative method that retains the similarity scores is possible if the information is not immediately used by the user.

In this poster, we describe experiments which show how similarity scores between gene profiles can be used to differentiate between microarray conditions. This information is used as the foundation for two other projects (see Section 5).

## 2. Data Set

Our data set is obtained from experiments with thirty-two male Swiss Webster mice (28-33 g, 5-8 wk) which 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 [Bartosiewicz et al., 2001]. MWG mouse microarrays containing 10,030 50mer oligonucleotides were hybridized with the RNA, and the resulting data was normalized using print-tip loess. Of the 10,030 gene profiles, 28 were duplicate in name and removed prior to the experiments described below.

## 3. Method

Hierarchical clustering pairs every gene with another gene or an existing cluster. Instead of this recursive process, we create a more general graph structure such that each node in this graph is a gene. Undirected edges are added between genes which have a similarity less than some threshold $\tau$. Each gene is permitted to be connected to zero or more genes (except itself). Thus, if $\tau = \infty$, then the graph is complete.

Two separate graphs are created for each of the two classes from our data set (control and $CdCl_2$ treated). Since the nodes in these graphs are labelled and identical, comparing the edges is trivial. If edges exist between gene A and gene B in both graphs, then these two genes are co-expressed in both the control and treated animals. When an edge exists in one graph, but not the other, then the pair of genes were expressed under only one of the conditions.

## 4. Experimental Results

The aim of these experiments is to quantify the number of edges that exist in one microarray condition, but not the other. Similarity between genes was calculated across all 16 experiments for both the control and treatment groups using Euclidean distance. Table 1 and Figure 2 presents our results. Regardless of the threshold chosen ($\tau$), at most 4.3% of the total number of edges appear in only one of the microarray data sets. For each threshold, the set of edges will vary.

| $\tau$ | 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.*

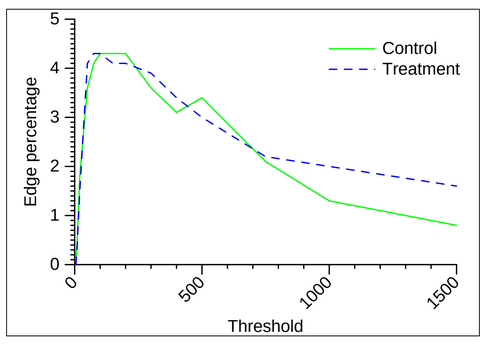


**Figure 2:** *Distribution of edges in one graph, but not the other, for varying values of $\tau$.*

## 5. Our Related Work

The work described in this poster has formed the basis for two current projects. The first is on outlier detection of microarray expression levels; the second is a comparison of microarray data with transcription factor binding sites for genes.

### 5.1 Outlier Detection

In earlier work, we showed how microarray data could be cleaned using the paradigm suggested in this poster [Wan et al., 2005]. That method had the danger of over-cleaning useful values. Since then, we have shifted our focus to outlier detection (Figure 3). Assuming a set of four replicates, a single experiment is tested while the remaining three are used to generate the graph structure described above ("training"). A neighboring gene is a gene which is similar to the current gene in the training data. The outlier (shown in red) is located by comparing it with its neighbors.

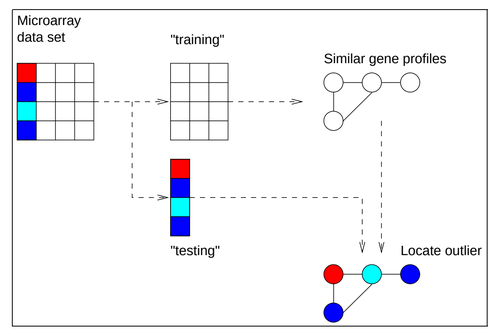


**Figure 3:** *Proposed method for outlier detection for microarray data.*

### 5.2 Transcription Factor Binding Sites

The two graphs outlined before represent pairs of genes which are co-expressed under specific conditions. One way in which genes are regulated when exposed to a given condition is through the transcription factor binding sites (TFBS) within their respective regulatory regions, as illustrated in Figure 4. As with the gene similarity described in this poster, a scoring mechanism can be employed to score these two genes based on their similar transcription factors (colored cyan in this figure). Then, a graph produced from this set of similarities can be compared with the two graphs described in Section 4.

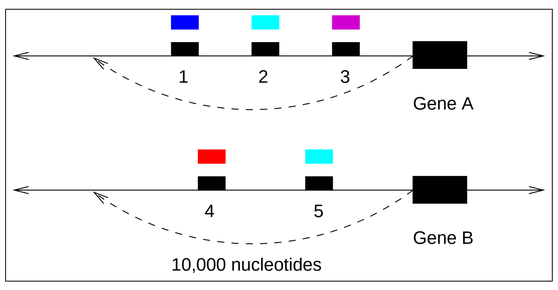


**Figure 4:** *Transcription factor binding sites in a regulatory domain of 10,000 nucleotides upstream.*

## 6. Discussion

This poster has proposed a method of classifying microarray experiments using pairs of gene profiles. While the output from such a method is unwieldy for a user, we intend to use it for the basis of two other projects which are currently on-going: outlier detection in microarray data and comparison with TFBS. One possible extension for this work is to look at sets of 3 or more genes, instead of confining ourselves to only pairs.

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

Any suggestions on our on-going work would be greatly appreciated.

## References

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.

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

R. Wan, H. Mamitsuka, and K. F. Aoki. Cleaning microarray expression data using Markov random fi elds based on profi le similarity. In *Proc. 20th ACM Symposium on Applied Computing*, pages 206–207, 2005.

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