A Framework for Determining Outlying Microarray Experiments

Raymond Wan[1] rwan@kuicr.kyoto-u.ac.jp
Åsa M. Wheelock[2] asa@para-docs.org
Hiroshi Mamitsuka[1] mami@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan
[2] Lung Research Lab L4:01, Respiratory Medicine Unit, Department of Medicine, Karolinska Institutet, 171 76 Stockholm, Sweden

Keywords: microarrays, distance-based outliers, data cleaning, error function

1 Introduction

While microarrays have been shown to be useful in assessing the expression levels of thousands of genes, one problem is the amount of noise. This abstract summarizes some earlier work on a graph-based framework for identifying noisy microarray experiments and an error function for cleaning individual expression levels [3].

2 Method

The basis of our framework is to use a microarray data set R to evaluate the level of noise in a new microarray slide t. The evaluation result is a score which indicates the percentage of probes which were deemed outliers.

We represent R as an undirected graph $G(V, E)$, as shown in Figure 1, using a dissimilarity function and a suitable threshold d_t. As we are interested in numerical dissimilarity and not profile similarity, we employ the Euclidean distance instead of a correlation function. We assume that R is of sufficient quality to compare against.

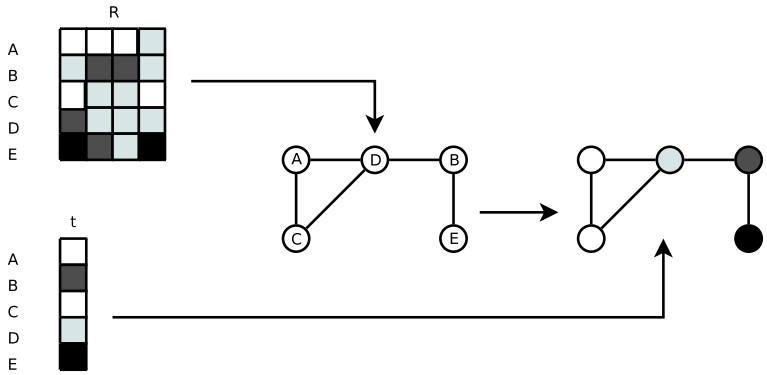

Figure 1: Illustration of our graph-based framework.

In the figure, R has five probes and four microarray slides. Its purpose is to create the *structure* of the graph. The structure dictates which probes a given probe should be compared against. Afterwards, the values of the microarray test slide t are inserted into the graph, which we show as shades of gray at the far right of the figure.

The slide t is scored by using a distance-based outlier approach [2]. Distance-based outlier detection compares every record with every other record using a suitable distance function (such as the Euclidean distance in our case). Here, we use the graph's structure to limit the number of comparisons done for both efficiency and practical reasons – two probes that are not similar in expression levels in R should not be expected to be similar in t. We employ a second threshold, e_t, to judge whether or not a neighboring expression level is an outlier.

An obvious extension is whether expression levels can be cleaned (corrected) using the same graph structure. The previously described method *marks* expression levels as outliers to obtain the score. Here, we focus on these marked expression levels for cleaning. Again, we turn to the Euclidean distance since we are interested in similarity in numerical values. We define the global energy E for microarray slide t as:

$E_t = \frac{1}{2}\sum_i^m \sum_j^m (\tilde{v}_i - \tilde{v}_j)^2$, where m is the number of probes and $\tilde{v}_i$ and $\tilde{v}_j$ are expression levels of two nodes such that an edge exists between them in G.

If we take the partial derivative of E with respect to some probe v_k, equate it to 0, and solve for v_k, we are left with m linear equations: $\mathbf{v} = \mathbf{A} \cdot \mathbf{v} + \mathbf{c}$, where $\mathbf{v}$ is the solution vector, $\mathbf{A}$ is an $m \times m$ matrix, and $\mathbf{c}$ is a constant vector whose value is derived from the the expression levels that are not marked as outliers. We then clean t by solving for the m linear equations using LU-decomposition and back substitution.

3 Results

In order to evaluate our system, we created synthetic, dye-swapped microarray data sets for R and t using the SIMAGE system [1]. Each microarray slide has $11{,}664$ probes. The size of R was 100 slides and 10 "noisy" slides were used one at a time for t. Gaussian noise is introduced into R using the default distribution of $\mathcal{N}(0, 0.219)$. The test slides have a larger standard deviation of noise of 0.500.

Our experiment pipeline is as follows. The 100 experiments in R is used to form G. We score each of the 10 noisy test slides and report the average score as the "Initial test set" score. The expression levels that were deemed outliers are cleaned using our error function. Then, the outlier detection phase is applied again, yielding our "Final test set". As a comparison, we extract the first 10 experiments from R and apply a single outlier detection step and call this our "Baseline".

We vary both d_t and e_t to illustrate their effect and present our results in Figure 2. We investigated two values for d_t: 3% and 10%, and produced a different set of lines for each. We vary e_t from 1% to 10% for the horizontal axis. The vertical axis is the average score across 10 experiments, out of 100%.

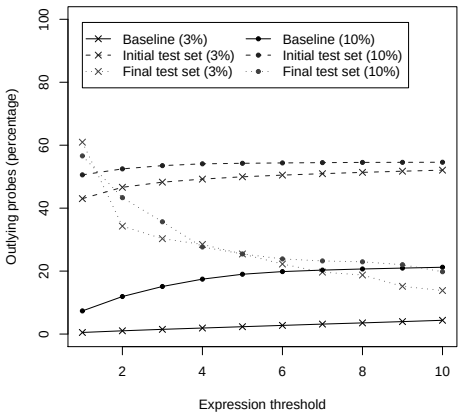

Figure 2: Results using simulated data.

Generally, the number of outlying probes is low for the baselines, with a noticeable difference between the two values of d_t. The initial percentages for the test set are larger: from 40% to 60%, indicating many of the expression levels differ from R. Cleaning the expression levels with the error function gradually lowers the number of outlying probes as e_t increases, for both values of d_t.

Additional results where we compare with statistical methods that rely on replicate microarrays (distance from the interquartile range, *Z*-test, and *Q*-test) are given elsewhere [3]. In the future, we would like to apply our work to actual microarrays. Also, our method relies on the Euclidean distance in several instances; we are considering whether other metrics would be more appropriate.

Acknowledgements: RW was supported by a postdoctoral fellowship from the Japan Society for the Promotion of Science (JSPS). ÅMW was supported by an EU Fp6 Marie Curie Fellowship. This work has been supported in part by BIRD of the Japan Science and Technology Agency (JST).

References

[1] C. J. Albers, R. C. Jansen, J. Kok, O. P.Kuipers, and S. A. van Hijum. SIMAGE: Simulation of DNA-microarray gene expression data. *BMC Bioinformatics*, 7(205), 2006. URL: `http://bioinformatics.biol.rug.nl/websoftware/simage/`.

[2] E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *Special Issue on the Best Papers of VLDB '98, VLDB Journal*, 8(3–4):237–253, February 2000.

[3] R. Wan, Å. M. Wheelock, and H. Mamitsuka. A framework for determining outlying microarray experiments. In *Proc. 8th International Workshop on Bioinformatics and Systems Biology (IBSB)*, volume 20 of *Genome Informatics*, (To appear).