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

Abstract

*Microarrays are high-throughput technologies whose data are known to be noisy. In this previously presented work [1], we apply a **graph-based** framework in a way that resembles distance-based outlier detection to:*

- *determine how noisy is a microarray experiment, and*
- *apply an error function to clean expression levels*

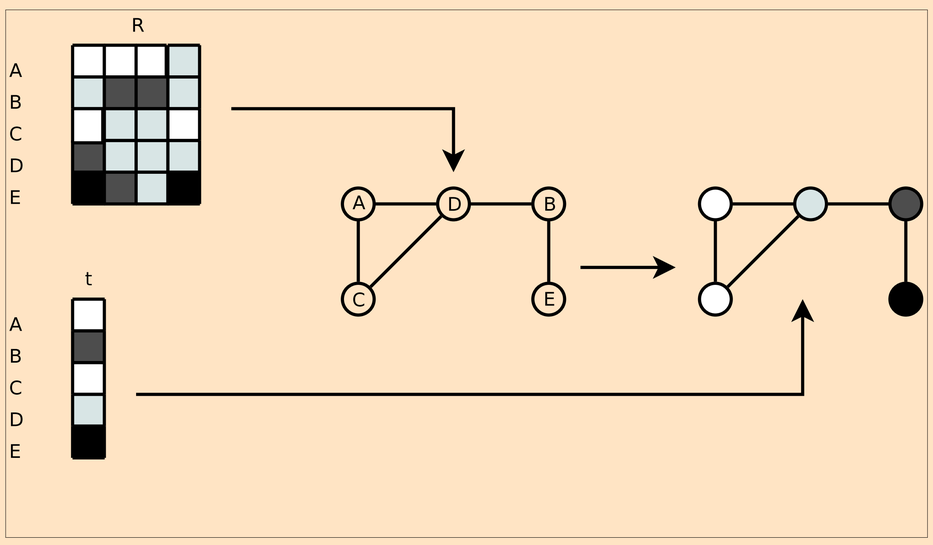
*The graph is built from a separate data set and illustrates another use of **past** microarray data.*

Motivation

- As more and more microarray data is uploaded to repositories, researchers are looking into combining data sets for data mining.
- Assuming that repository data R is of sufficient quality to compare against, we apply it for assessing the reliability of experiment t (where $t \notin R$).
- Distance-based outlier detection finds database records that differ "noticeably" from others (i.e. tax fraud, etc.).
- Instead of comparing every probe with every other probe, we build a **graph** from R to **limit comparisons** on t.
- Can microarrays be "cleaned" with this graph structure using an error function based on the Euclidean distance.

Framework Overview

Overview of how past microarray data from a repository (R) and the experiment (t) to be assessed are employed.



Algorithms

Inputs: R, t, d_t, e_t.
Output: Percentage of outliers in t.

$d \leftarrow$ Calculate all-pairs distance matrix(R)
$E \leftarrow d' \leftarrow$ All distances d_t
$V \leftarrow$ probes and expression levels of t
Build $G(V, E)$
outliers $= 0$
foreach $v_i \in V$
  if (v_i differs more than e_t from **most** of its neighbors)
    outliers++
return (outliers / $|V|$ * 100)

(a) Outlier detection

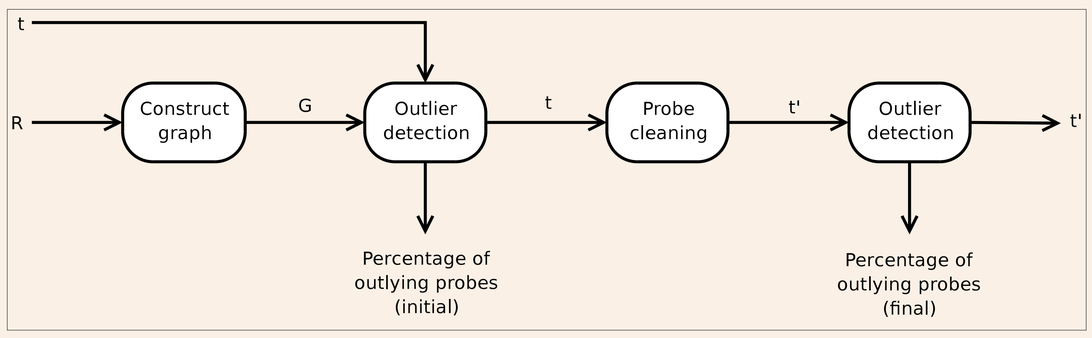
1. Define an **error function** using the Euclidean distance:

$$E = \frac{1}{2}\sum_{i}^{m}\sum_{j}^{m}(\tilde{v}_i - w_{ij}\tilde{v}_j)^2 . \qquad (1)$$

2. To minimize the error, take the partial derivative $\frac{\partial E}{\partial v_k}$, for some vertex v_k.
3. We have one equation for each vertex ($\mathbf{v} = \mathbf{A} \cdot \mathbf{v} + \mathbf{c}$), so solve them **simultaneously** using LU-decomposition and back substitution.
4. **Replace** the expression levels of v_k if they were marked as outliers before.

(b) Probe cleaning

Framework



Simulated dye-swapped data sets created using SIMAGE [2]:

Name	Probes	Experiments	Random noise	Purpose
$\mathcal{D}_1$	11,664	100	$N(0, 0.219)$	R
$\mathcal{D}_2$	11,664	10	$N(0, 0.500)$	t

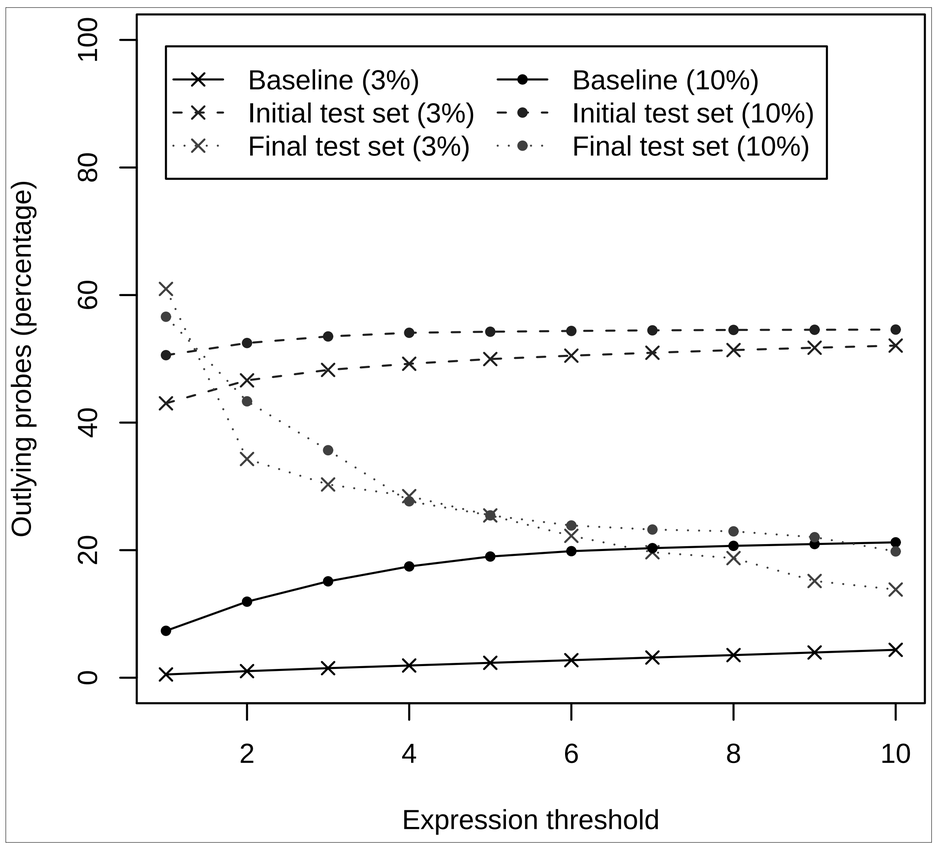
Also, 10 experiments from R used to form our "baseline" t.

About the results...

- Baselines: Number of outlying probes low.
- Increase in outliers from $d_t = 3\%$ to 10% ("denser" graph).
- Test set: Number of outlying probes high (40% to 60%).
- Probe cleaning gradually lowers the number of outlying probes.

Results

Experiments with (a) Two values for d_t (3% and 10%); (b) e_t from 1% to 10%.



Future Work

- We apply some form of the Euclidean distance three times; are there (better) alternatives?
- Evaluate our method on real microarray data.
- Further investigate whether microarray experiment cleaning using the error function can be improved (currently, there is no term to prevent over-cleaning).

[1] 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, October 2008.
[2] C. J. Albers, R. C. Jansen, J. Kok, O. P. Kuipers and S. A. van Hijum, BMC Bioinformatics 7 (2006), URL: http://bioinformatics.biol.rug.nl/websoftware/simage/.

The 2008 Annual Conference of the Japanese Society for Bioinformatics (**JSBi 2008**), December, 15-16, 2008, Osaka, Japan.