A Framework for Determining Outlying Microarray Experiments

Raymond Wan
Bioinformatics Centre
Kyoto University, Japan
rwan@kuicr.kyoto-u.ac.jp

June 9, 2008

Collaborators: Åsa M. Wheelock (Karolinska Institutet, Sweden)
and Hiroshi Mamitsuka (Kyoto University, Japan)

Outline

Framework and Application

Experiments

Conclusion

Overview

- Develop a framework to assess the degree to which an entire microarray experiment T is an outlier using a separate set of n (currently, replicate) experiments.
- Framework is based on an undirected graph indicating similarity between probes across the n replicates.
- Scoring of T is based on a count of the number of probes which differ.

Motivation

1. Microarray repositories
   - Microarray repositories like NCBI GEO and Stanford SMD hold many microarray data sets which are already being used for meta-analysis of microarrays.
   - Despite the variations between laboratories, can they also be used to determine whether or not a newly generated experiment is “suspicious”?
2. Experimenter bias
   - Microarray experiments represent monetary costs to the experimenter.
   - Can an impartial mechanism be developed which makes use of already-made data sets (as a guide)?

Framework and Application

Framework (1)

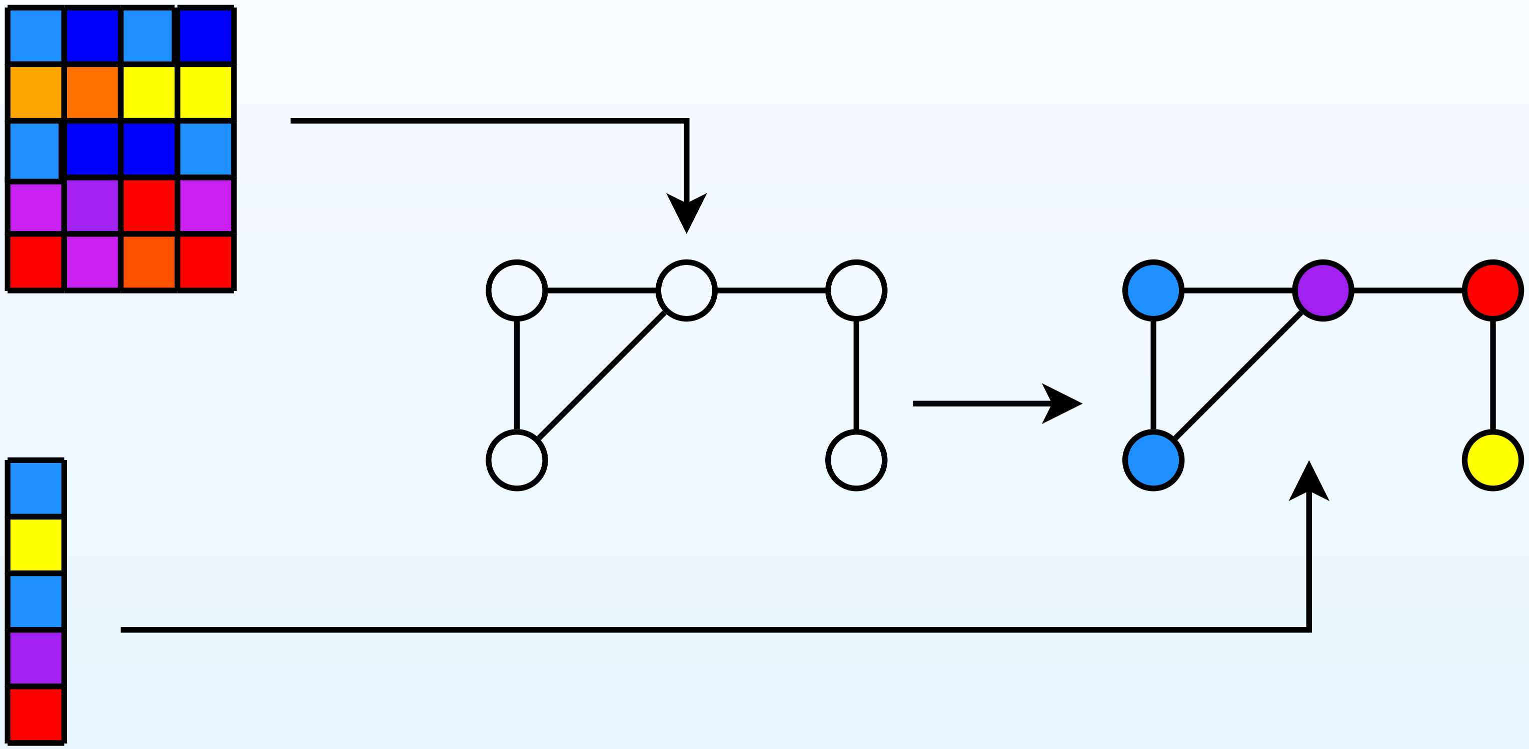
Given: n replicate microarrays and the new experiment T.
Steps:

1. Build an undirected graph $G(V, E)$ of distance similarities using the n replicate microarrays and a distance threshold d_t.
2. Insert the expression levels from the new experiment T.
3. Check how many expression levels differ from their immediate neighbors using an expression threshold e_t; represent this as a percentage on a per-slide basis.

Distance similarities $\Longrightarrow$ Euclidean distance, since we are interested in probes which consistently have the same expression levels.

Framework (2)

Repository / replicates (5 probes, 4 experiments)



New experiment

Application of Framework

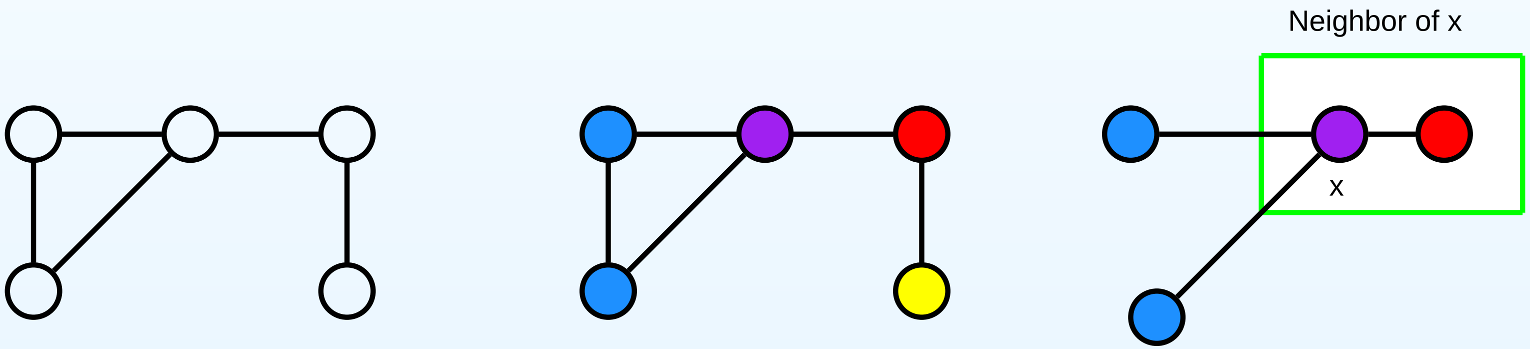
With the undirected graph $G(V, E)$ made, how can we assess the experiments?

We apply “distance-based outlier detection” (from the field of Knowledge Data Discovery [KDD]), which examines how far a database record is from all other records. Some definitions [Bay and Schwabacher, 2003]:

1. Outliers are the examples for which there are fewer than p other examples within a distance d.
2. Outliers are the top n examples whose distance to the kth nearest neighbor is greatest.
3. Outliers are the top n examples whose average distance to the k nearest neighbors is greatest.

Distance-based Outliers for Microarrays

Between every probe p_1 and p_2, there is a distance similarity $d(p_1, p_2)$ and an expression similarity $e(p_1, p_2)$, calculated from the n replicates and T, respectively. These values are regulated by two thresholds: d_t and e_t.



Similar probes in replicates.

Insert the expression values from T.

Focus on a probe x.

Within the probe's neighborhood, if there are more distant-neighbors than close-neighbors, then the probe is counted against T (as an outlying probe).

- Outline
- Overview
- Motivation

Framework and Application

- Framework (1)
- Framework (2)
- Application of Framework
- Distance-based Outliers for Microarrays
- Comparison Method
- Cleaning microarrays...
- Simulating Data

Experiments

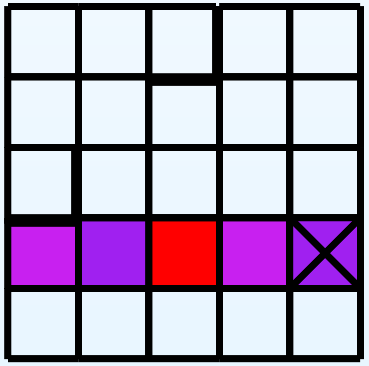
Conclusion

Comparison Method

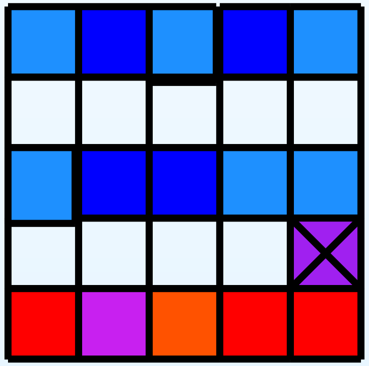
Compare against inter-quartile range (IQR), Z-test, and Q-test, where the Q-test is defined as:

$$Q(x) = \frac{|x - (\text{closest value to } x)|}{\text{range}} \tag{1}$$

So, if we visualize the n replicates with T together:



Applying the statistical methods



Applying the framework

Cleaning microarrays...

Within the same framework, we consider an error function based on an energy function derived from each of the n probes and their neighborhood:

$$E = \frac{1}{2} \sum_{i}^{n} \sum_{j}^{n} (\tilde{p}_i - w_{ij}\tilde{p}_j)^2 \, . \tag{2}$$

Solving for some probe p_k, we obtain n simultaneous equations:

$$\mathbf{p} = \mathbf{A} \cdot \mathbf{p} + \mathbf{c} \, . \tag{3}$$

where **v** is the solution vector and **A** is:

$$a_{ij} = \frac{2w_{ij}}{|\mathcal{N}_i| + \sum_k^N w_{ik}^2} \tag{4}$$

Simulating Data

We evaluate our framework using artificially created microarray data using the SIMAGE web server[1] [Albers et al., 2006].

We created:

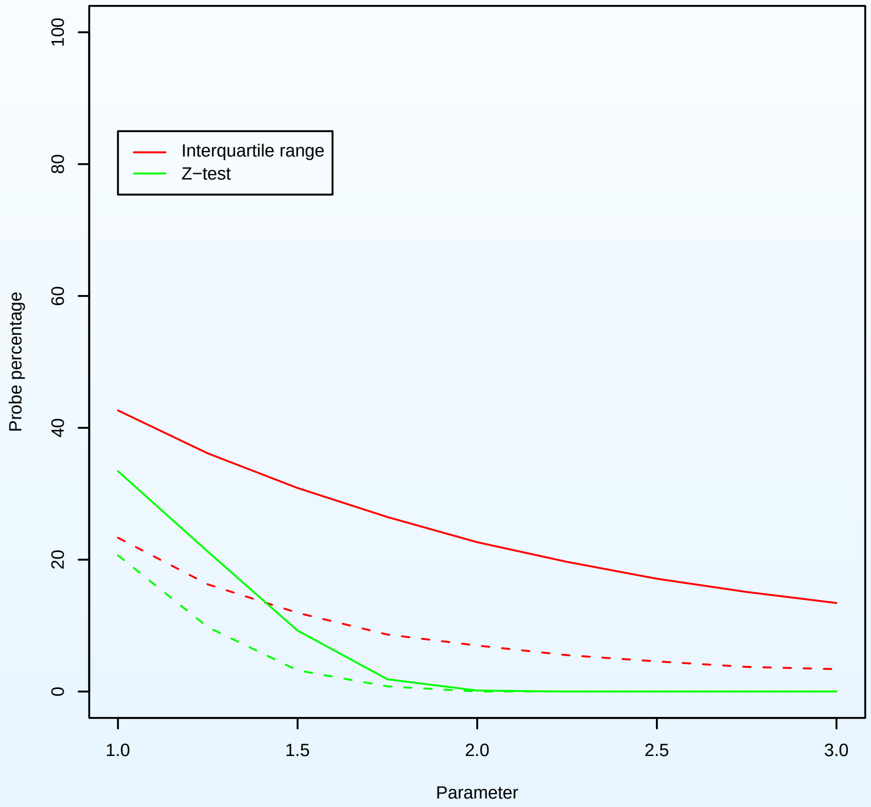
- 6 slides (3 sets of dye-swap)
- $4{,}400$ probes each using default parameters[2]
- 1 slide with the change in the Gaussian noise distribution $N(0, \sigma_\epsilon^2)$ from $\sigma_\epsilon^2 = 0.219$ to 0.438.

[1] URL: `http://bioinformatics.biol.rug.nl/websoftware/simage/`

[2] The SIMAGE maintainers obtained these values by modeling 23 real experiments.

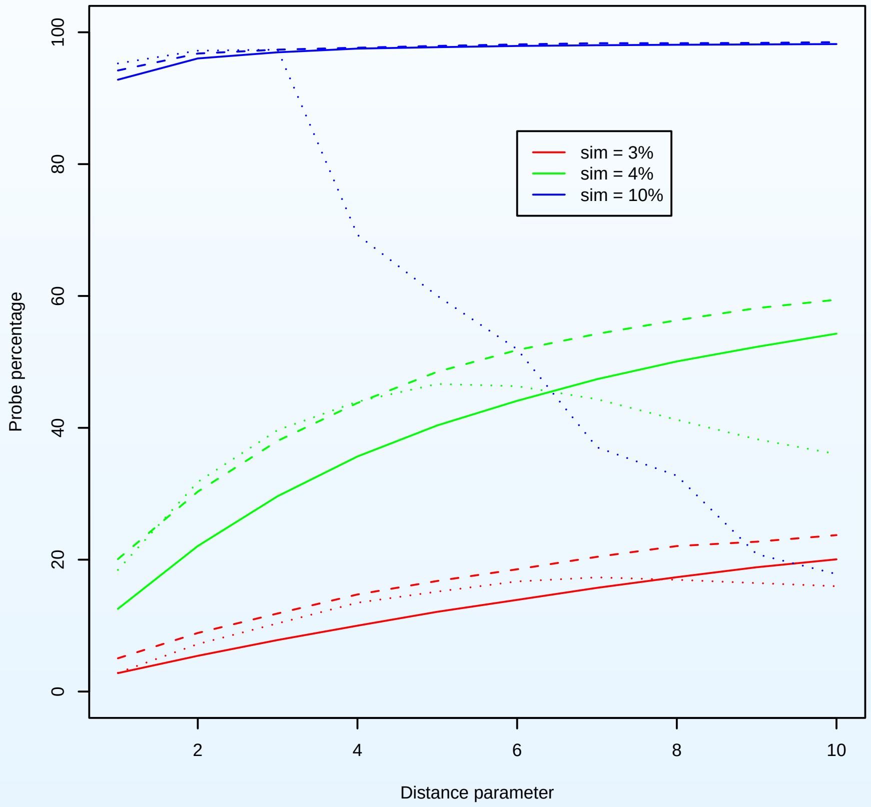
Experiments

Statistical methods



Solid lines: Average IQR or Z-score across the replicates; **Dashed lines**: IQR or Z-score for T. Q-test performed 1.64 % and 1.01 % for replicates and T, respectively.

Distance-based outlier methods



Solid lines: Average across replicates; **Dashed lines**: T. **Dotted lines**: Effect from apply error function to probes marked as outliers with respect to their neighbors.

Summary

Statistical methods:

- X Report higher percentages for the replicates than T.
- ✓ Number of errors reported decreases as we relax the parameter.
- X Q-test appears less strict than the other two tests (low percentage).

Distance-based outlier methods:

- ✓ Results reasonable for small parameter values.
- X The lines for replicates and T are indistinguishable as we increase the parameters (blue lines and moving right in the graph).
- X As we add more edges, the error function over-cleans since the dotted lines are brought closer to the x-axis.

Conclusion

Summary

We have:

- Proposed a framework for assessing the reliability of a single microarray experiment using other [external] experiments and scoring based on the percentage of differing probes.
- Executed preliminary experiments, but more detailed experiments needed to assess parameter choice.

The aim of this work is to give experimenters an unbiased assessment of their microarray experiment prior to data analysis.

Future Work

In the future, we would like to:

- Apply this to actual microarray data. So far, these obstacles:
  - Publicly available data are usually normalized prior to upload to GEO/SMD.
  - “Suspicious” data would not be uploaded to a public repository anyway...

  So, we welcome any ideas on what could serve as the replicates and/or T...
- Consider generalizing graph construction; perhaps using non-replicates or sequence similarity between genes...

Acknowledgements

Collaborators:

- Prof. Hiroshi Mamitsuka (Bioinformatics Centre, Kyoto University, Japan)
- Dr. Åsa M. Wheelock (Department of Medicine, Karolinska Institutet, Sweden)

Acknowledgements:

- Dr. Matthew J. Bartosiewicz (formerly University of California, Davis)
- Dr. Timothy Hancock (Bioinformatics Centre, Kyoto University, Japan)

Data and Software:

- SIMAGE

References

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

S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proc. 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 29–38, 2003

D. P. Shoemaker, C. W. Garland, and J. W. Nibler. *Experiments in physical chemistry*. McGraw-Hill, fifth edition, 1989

Q-test

Critical values for the Q-test according to a 90% confidence interval are[3]:

N	3	4	5	6	7	8	9	10
Q_c	0.94	0.76	0.64	0.56	0.51	0.47	0.44	0.41

[3] Source:Shoemaker et al. [1989, pg. 35]

Error Function Extras

Partial derivative with respect to a probe p_k ($\frac{\partial E}{\partial p_k}$) and solve for p_k:

$$p_k = \frac{2\sum_i^N w_{ki} p_i}{|\mathcal{N}_k| + \sum_i^N w_{ki}^2} \tag{5}$$

An entry in **A** is:

$$a_{ij} = \frac{2w_{ij}}{|\mathcal{N}_i| + \sum_k^N w_{ik}^2} \tag{6}$$

While the solution vector **v** represents “new” values, we are more concerned with how many of the values changed within a small Δ.