MicroarrayCAKE: a simulation and analysis framework to guide experimental design and gene expression data analysis

Rajanikanth Vadigepalli[1,3], Rishi Khan[2],Guang Gao[2] and James Schwaber[1]

*[1]Daniel Baugh Institute for Functional Genomics and Computational Biology,
Department of Pathology, Thomas Jefferson University, Philadelphia, PA 19107
[2]Department of Electrical Engineering, University of Delaware, Newark, DE 19711
[3]Author to whom correspondence should be addressed: raj@mail.dbi.tju.edu*

Key words: bioinformatics tool, microarray data, gene expression analysis, experimental design

Prepared for Presentation at the 2004 Annual Meeting, Austin, TX, Nov 7-12,
Session #49 'Simulation of Biomolecules I: Computational Representation of Genomics and Proteomics'

November 2004

Unpublished

AIChE shall not be responsible for statements or opinions contained in papers or printed in its publications

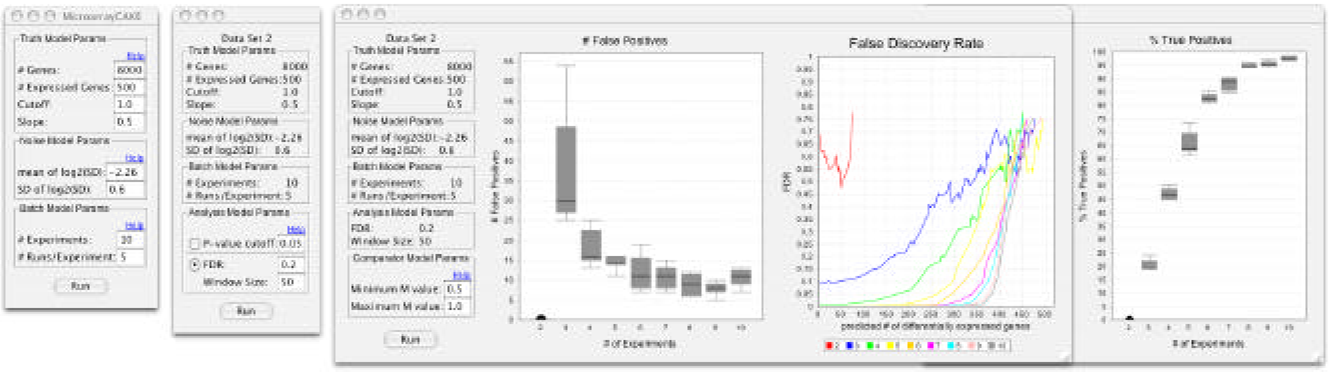
ABSTRACT

We describe a framework, MicroarrayCAKE (Correct Answer Known Evaluator), in order to aid researchers design microarray experiments. The framework allows a user to build a dataset that reflects the "known correct answer" and simulates gene expression measurements based on the expected statistical characteristics of the biological and experimental processes. The simulated gene expression datasets are analyzed using multiple statistical methods to predict differentially expressed genes. These predictions are compared to the simulated "Known Correct Answer" to compute the true/false positive rates of the analysis methods. MicroarrayCAKE allows researchers to apply multiple analysis methods on the same data set in order to arrive at an optimal number of experimental repeats necessary to achieve the user-specified true/false positive rates. Conversely, the MicroarrayCAKE results provide an estimate of the true/false positive rates of a given number of experimental repeats, thus providing researchers the ability to evaluate previously collected gene expression data sets. The software, associated documentation and a tutorial are available at http://www.dbi.tju.edu/MicroarrayCAKE.

SUMMARY and RESULTS

We describe a framework, MicroarrayCAKE (Correct Answer Known Evaluator), in order to aid researchers design microarray experiments. The framework allows a user to build a dataset that reflects the "known correct answer" and simulates gene expression measurements based on the expected statistical characteristics of the biological and experimental processes. The simulated gene expression datasets are analyzed using multiple statistical methods to predict differentially expressed genes. These predictions are compared to the simulated "Known Correct Answer" to compute the true/false positive rates of the analysis methods. MicroarrayCAKE allows researchers to apply multiple analysis methods on the same data set in order to arrive at an optimal number of experimental repeats necessary to achieve the user-specified true/false positive rates. Conversely, the MicroarrayCAKE results provide an estimate of the true/false positive rates of a given number of experimental repeats, thus providing researchers the ability to evaluate previously collected gene expression data sets.

MicroarrrayCAKE can compensate for the effects of multiple comparisons using methods such as Bonferroni correction or False Discovery Rate (FDR) correction. MicroarrayCAKE implements 3 analysis methods based on: (1) p-value threshold, (2) FDR threshold, and (3) FDR over a moving window of a p-value interval containing specified number of genes. We have developed a novel method (Adaptive Pairing) to improve discrimination between true positives and false positives. For each gene, Adaptive Pairing considers all permissible combinations of perturbation-control pairs and utilizes the pairing that yields maximum variance. This method has been implemented in MicroarrayCAKE. Through simulations we demonstrate that the Adaptive Pairing method yields better Sensitivity and Specificity compared to the standard one sample t-test independent of p-value threshold.



MicroarrayCAKE

Truth Model Params
Genes: 8000
Expressed Genes: 500
Cutoff: 1.0
Slope: 0.5

Noise Model Params
mean of log2(SD): -2.26
SD of log2(SD): 0.6

Batch Model Params
Experiments: 10
Runs/Experiment: 5

Data Set 2

Truth Model Params
Genes: 8000
Expressed Genes: 500
Cutoff: 1.0
Slope: 0.5

Noise Model Params
mean of log2(SD): -2.26
SD of log2(SD): 0.6

Batch Model Params
Experiments: 10
Runs/Experiment: 5

Analysis Model Params
□ P-value cutoff: 0.05
● FDR: 0.2
Window Size: 50

Data Set 2

Truth Model Params
Genes: 8000
Expressed Genes: 500
Cutoff: 1.0
Slope: 0.5

Noise Model Params
mean of log2(SD): -2.26
SD of log2(SD): 0.6

Batch Model Params
Experiments: 10
Runs/Experiment: 5

Analysis Model Params
FDR: 0.2
Window Size: 50

Comparator Model Params
Minimum M value: 0.5
Maximum M value: 1.0

Figure 1: (a) Data generator plugin builds a “correct answer” data set and a simulated measurement data set (b) Analysis plugin calculates p-values and predicts genes expression based on p-value (c) Comparator plugin compares the analysis predictions with the “correct answer” data set and reports false positive and true positive rates.

A MicroarrayCAKE workflow contains 3 major sections: 1) model based data generation, 2) analysis to predict differentially expressed genes, and 3) comparison of the predictions to the known correct answer, as shown in figure 1. The inputs to the data generation stage are: (1) total number of genes measured (2) expected number of differentially expressed genes, (3) the distribution of expressed genes, (4) the estimates of various sources of variability in microarray experiments, (5) the maximum number of experimental repeats to be considered, and (6) number of times the simulation is repeated. In the "correct answer" model, the log2(perturbation/control) of the differentially expressed genes is uniformly distributed between 0 and a user specified cutoff, and then tapers down to the maximum expression. Corrupting the "correct answer" with a noise model yields the measured expression values. The noise model contains additive noise with mean 0 and a gene-specific standard deviation, thus emulating biological variability of gene regulation. Analysis of gene expression datasets from our group indicates that the distribution of the standard deviation is log-normal (2^N(mean_sd,sd_sd)). Setting sd_sd=0 simulates noise that is not gene-specific. The measured gene expression dataset consists of microarray designs containing 2 to the maximum number of experimental repeats considered.

MicroarrayCAKE allows researchers to examine the relationship between the number of repeated experiments and the number of correct predictions. The simulated data is analyzed using multiple analysis methods and their predictions are evaluated by comparing with the "known correct answer". Differential gene expression is predicted by using a threshold on either p-value or FDR with an optional window size (set window size = 0 for standard FDR analysis). The predictions are evaluated by calculating the percent of false positives overall and the percent true positives detected within a user specified window. If the FDR method was used FDR is plotted against the number of predicted genes.

The framework is implemented in Java as a Java web-start application, but can be installed locally. The software, associated documentation, and a detailed tutorial are available at http://www.dbi.tju.edu/MicroarrayCAKE.