Using pre-existing microarray datasets to increase experimental power: Application to insulin resistance

Bernie J. Daigle, Jr.[1], Philip S. Tsao[2], and Russ B. Altman[3]

***Short Abstract* — A central task in systems biology is the accurate analysis of high-throughput datasets. We have developed a method that increases the power of a microarray experiment by leveraging the vast number of publicly available microarray datasets in concert with the underlying modularity of transcriptional responses. We demonstrate its effectiveness on simulated data, three highly replicated human datasets, and data from a novel human insulin resistance study.**

***Keywords* — Differentially expressed genes, expression modules, microarray analysis, singular value decomposition, systems biology.**

I. Background

EXTRACTING accurate and meaningful conclusions from high-throughput datasets is an important task in systems biology. DNA microarrays are a widely used high-throughput technique, but they are notorious for generating noisy data. This is especially apparent in discrepancies found between studies that identify differentially expressed (DE) genes [1-3]. A common strategy for mitigating the effects of noise is to perform many experimental replicates, but this approach is often costly and sometimes impossible given limited resources. Thus, analytical methods are needed which increase accuracy at no additional cost.

One inexpensive source of potentially relevant knowledge is the collection of published, freely available microarray datasets. To date, data from hundreds of thousands of microarray experiments are in the public domain, yet few existing methods use this information to help identify DE genes. This work is an attempt to use such information in a mathematically principled way to better identify DE genes in a dataset of interest.

II. Results

We present the Singular value decomposition (SVD) Augmented Gene expression Analysis Tool (SAGAT), a data-driven statistical approach for identifying DE genes. SAGAT increases the power of a microarray experiment by leveraging the vast number of publicly available microarray datasets in concert with the underlying modularity of transcriptional responses. SAGAT works by (1) applying SVD to a compendium of pre-existing datasets to identify expression modules (eigengenes) and (2) using this information to create a more robust statistic for detecting differential expression.

We tested the method on simulated data and three well-replicated human microarray datasets [4-6], and we demonstrate that use of SAGAT increased effective sample size by as many as 2.72 arrays. Upon applying SAGAT to a novel microarray study searching for human insulin resistance-related genes, we discovered and experimentally validated a number of novel candidates that would have been missed using existing methods.

III. Conclusions

Transcriptional responses are often modular, with groups of genes undergoing coordinated expression changes. To leverage knowledge of this phenomenon to more accurately identify DE genes, we have developed SAGAT, a method that integrates pre-existing expression data with a dataset of interest. SAGAT improves the accuracy of DE gene identification without increasing experimental cost. We provide SAGAT as a freely available software package that is immediately applicable to any human microarray study.

Acknowledgements: This work was funded by an HHMI predoctoral fellowship, an NLM training grant, and funding from Microsoft Research.

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA. E-mail: bdaigle@stanford.edu

[2]Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA.

[3]Departments of Genetics and Bioengineering, Stanford University School of Medicine, Stanford, CA.

References

[1] Tan PK, et al. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**, 5676-5684.

[2] Miklos GL, Maleszka R (2004) Microarray reality checks in the context of a complex disease. *Nat Biotechnol* **22**, 615-621.

[3] Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *PNAS* **103**, 5923-5928.

[4] Singh D, et al. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-209.

[5] Miller WR, et al. (2007) Changes in breast cancer transcriptional profiles after treatment with the aromatase inhibitor, letrozole. *Pharmacogenet Genomics* **17**, 813-826.

[6] Sabates-Bellver J, et al. (2007) Transcriptome profile of human colorectal adenomas. *Mol Cancer Res* **5**, 1263-1275.