Improved Modeling and Analysis of Gene Expression

Annamarie Bair*, Matthew B.A. McDermott*, Jennifer Wang†, Wen-Ning Zhao‡, Steven D. Sheridan†, Peter Szolovits*, Isaac Kohane⋆, Stephen J. Haggarty‡, Roy H Perlis†

*CSAIL, MIT; †CEDD & ‡CGM, MGH; ⋆DBMI, Harvard

annabair@mit.edu

Abstract

Gene expression is the process by which information encoded in DNA is transcribed and translated into proteins. Analyzing this data is extremely valuable for understanding disease and the effects of various treatments. However, analysis can be challenging because gene expression datasets are high-dimensional, noisy, and sometimes not normally distributed. This work demonstrates two main methods to better analyze gene expression data: a method of improved hypothesis testing and a method of correlational analysis. We show that our improved hypothesis testing is able to perform a more meaningful measure of statistical significance on non-normally distributed data. We also demonstrate the power of analyzing correlative relationships between genes and identify significantly differentially expressed genes between a disease and control state.

Data

1. Synthetic data
2. Patient data from Massachusetts General Hospital (MGH)
   - 2 patients with Schizophrenia (SCZ) and 2 Healthy Control Subjects (HCS).
   - L1000 technique resulted in 758 HCS and 756 SCZ measurements of 978 genes.

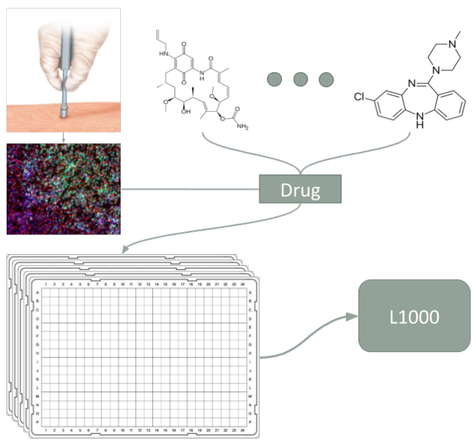


Figure 1: L1000 technique was used to analyze amount of mRNA produced by transcription of each gene in patients' Neural Progenitor Cells.

Methods

Zeroth order Method

Perform a more meaningful test of statistical significance on non-normally distributed data.

- Some genes (Figure 2) display evidence of non-normality.
- Use Gaussian Mixture Model to fit an appropriate distribution.
- Use probabilistic measure of significance.

Normal p- value:

$$p = \mathbb{P}_{\mathsf{n}}(n > X)$$

Probability-based measure of significance:

$$p = \mathbb{P}_{\mathsf{n}}(p_{\mathsf{n}}(n) > p_{\mathsf{n}}(X))$$

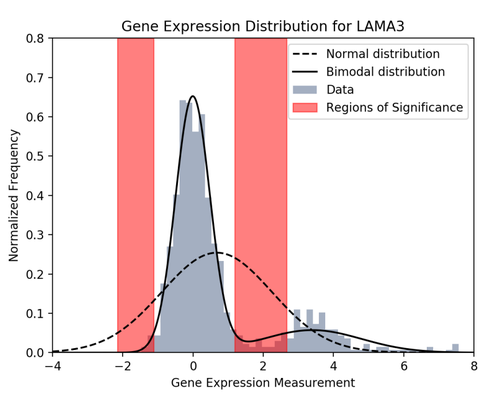


Figure 2: LAMA3 expression levels are non-normally distributed.

First order Method

Detect disruption of co-regulatory relationships between genes.

- Compare the correlative relationships between genes in the control population and those in the disease population.
- Genes whose correlative relationships differ the most are flagged as anomalous.
- Can detect information not evident in direct comparison of expression levels.

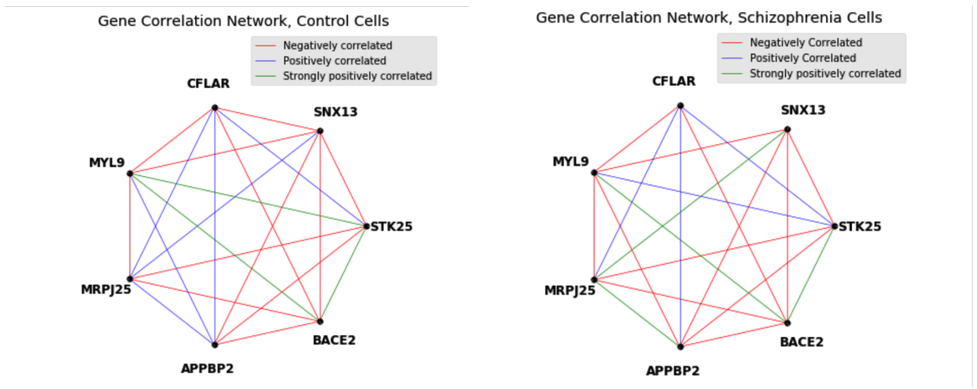
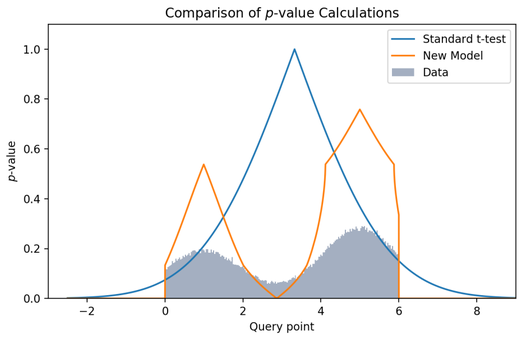


Figure 3: Differences in correlative relationships between HCS cells and SCZ cells. Note the differences between pairs of genes (MYL9, APPBP2), (CFLAR, SNX13), (MRPJ25, APPBP2).

Results

Synthetic Data

- p-value as calculated by a standard t-test misrepresents the distribution on the right.
- Our probabilistic measure of significance accounts for both bimodality and truncation.



Patient Data

- This table shows genes with significantly different correlative relationships between SCZ and HCS cells.
- MPZL1, ACLY, STXBP1, and FOXJ3 are all related to neural processes.

Gene	Significance
MPZL1	≤ 0.01
ACLY	0.02
STXBP1	0.02
FOXJ3	0.05

Conclusion

1. We demonstrate a *zeroth order* method to model non-normal distributions, including bimodal mixtures with truncated components.
2. We use a *probabilistic measure of significance* to better perform hypothesis testing for non-normal distributions.
3. We use a *first order* method of correlative analysis to identify genes with significantly different patterns of expression in patients with Schizophrenia as opposed to healthy control subjects.

This research was funded in part by grants from the National Institutes of Health (NIH): National Institute of Mental Health (NIMH) grant P50-MH106933, National Human Genome Research Institute (NHGRI) grant U54-HG007963, and the MIT SuperUROP Program.