Identifying Local Gene Expression Patterns in Biomolecular Networks

A. Y. Sivachenko, A. Yuryev, N. Daraselia, I. Mazo
Ariadne Genomics, Inc., 9700 Great Seneca Hwy., Rockville, MD 20850
sivachenko@ariadnegenomics.com

Abstract

This study is aimed at elucidating putative transcription regulators (TRs) responsible for the observed differential expression pattern. Combined direct promoter binding and indirect transcriptional regulation networks were used, and the expression levels of each TR's downstream targets were collectively analyzed, as a sample, for significance. Statistical procedure was also developed that takes into account the sign of the expression change, thus requiring the downstream targets to exhibit expression pattern consistent with the regulatory relationship effect signs.

1. Introduction

Analysis of large scale expression datasets poses a significant challenge because of natural variability of gene expression, high levels of noise, and multiple testings performed in one run. Conventional methods for microarray analysis (e.g. [1]) are aimed at finding significant features in the data and, for instance, selecting differentially expressed genes based on the data statistics alone. However, using the expression data in combination with additional information, such as relationships among genes, improves the power of statistical tests and leads to better and more extended predictions. Examples of promising existing integrative approaches include, *e.g.* co-clustering of gene expression and interaction data [2], and discovering putative regulatory pathways [3].

2. Statistical model

Our goal is to find a putative set of TRs driving the differential expression of other genes and thus suggesting an explanation for the observed data.

We refer to a TR as "significant" if its downstream targets in the regulatory network exhibit, as a set, a pattern of differential expression significantly deviating from the distribution expected by random chance. The definition of such "significance" is, however, model-specific. The simplest model would deal with pre-selected set of differentially expressed (DE) genes determined from the microarray data and evaluate the overrepresentation of DE genes among the targets of each TR. We found, however, that this model lacks sensitivity and that the results strongly depend on the (arbitrary) *p*-value cutoff used to pre-select DE genes.

Hence, we suggest a more robust test, in which the expression values (log-ratios) measured for each target of particular TR are considered as a *sample* to be tested against the sampling distribution.

The simplest choice for the sampling distribution is the collection of all log-ratio absolute values measured on the array. We argue however that "network rewiring" randomization procedure [4] should be generally preferred. In resampling terms, this procedure requires breaking all the network edges and then randomly reconnecting the dangling edge halves. This model has certain biologically sound benefits. In particular, consider a gene that can be regulated by a large number of different TRs in the network. Even if such a gene exhibits very high or very low log-ratio, it still makes little contribution to, or is not too prohibitive for the significance of each particular sample (and upstream TR) it belongs to. While such behavior is achieved automatically through network resampling, the brute force approach is too expensive computationally and we suggest here a simple and efficient alternative. To approximate the resampling it is sufficient to replicate each measured log-ratio absolute value on the array by the number of TRs regulating this gene (*i.e.* by the in-degree of the gene in regulatory network). Each replicated log-ratio is now effectively associated one-to-one with the in-going *edge* adjacent to the gene, thus we refer to this distribution as "edge distribution".

Importantly, the edge distribution also provides us with the means to take into account not only the fold change absolute value, but also its direction. Indeed, if

we take the sign of gene's log-ratio (up-regulated, +1, and down-regulated, -1) and define edge (regulatory relationship) effect signs (positive, +1, or negative, -1, regulation), then for any given TR truly exerting its regulatory function, the product of effect sign by the target expression change sign is expected to be the same for all targets (*e.g.*, if a TR is activated, then the targets it positively and negatively regulates are expected to be activated and suppressed, respectively, $1 \cdot 1 = (-1) \cdot (-1) = 1$). We can thus build a "signed edge distribution", in which each measured log-ratio is replicated by the gene's in-degree and each replica is multiplied by the corresponding edge's effect sign. Drawing from such a distribution still approximates network resampling, when a TR is allowed to randomly pick dangling edges bearing effect signs. The sample for such test should be also modified: all the log-ratios of the downstream targets must be taken with signs and multiplied by the corresponding edge signs. If some edge signs are unknown, we run the test on the subsets of targets (edges) for which effect signs are available.

4. Data

The whole-genome expression dataset (Affymetrix U133 array) comparing primary tumor and isogenic metastatic colon cancer cell lines [5] was used; the transcripts were mapped to 12,902 loci.

The regulatory network was extracted from the ResNet database [6], which contains a collection of ~500,000 relationships automatically mined from the biomedical literature with the natural language processing full-sentence parsing algorithm [7]. For this work we used a set of ~12,000 direct and indirect transcription regulation relationships among 3,845 genes and gene products.

5. Results

First, we pre-selected DE genes with p-value cutoff 0.001 (291 genes). Testing for overrepresentation of DE genes among downstream targets we found the following significant TRs: E2F1, LEF1, FLI1, RB1, TP53 (p-values 0.008, 0.022, 0.026, 0.026, 0.050, respectively). At the cutoff $p<0.005$ (860 DE genes), we still observed LEF1, TP53, and FLI1 (p=0.017, 0.018, 0.021) as significant, however other TRs lost significance, while a few new TRs (MAP2K3, E2F4, TCF4) became significant. Some of the observed TRs are indeed major elements of colon cancer pathway (LEF1, TCF4) and others might be implicated, but it is seen that the procedure has low sensitivity and demonstrates strong dependence on the DE cutoff.

Next, we built the "edge distribution" and compared to it samples of absolute log-ratio values downstream of each TR. Mann-Whitney ranked test was used to calculate the p-value. All significant TRs found with the discrete test were recovered, except for LEF1 (p=0.18). However, many other potentially interesting regulators were found (discussed below). We suggest using unsigned test when most of the effect signs are unknown or when higher sensitivity is desired.

Finally, when we apply the signed test, a number of TRs from the unsigned test results lose their significance, indicating that their downstream targets exhibit somewhat elevated log-ratios, but the change directions do not correlate well with effect signs. Checking for consistent expression change pattern, the signed test is expected to perform the most stringent selection. In the absence of the golden standard, we refer to the list of 10 most significant TRs obtained with the signed test (p-values 0.004–0.017): JUND, PPARA, PPARG, AKT1, CTCF, NOG, LEF1, TP73, ETV4, PKCZ, and CITED2. Indeed, all these genes except for CITED2 are either known to play a role or implicated in colon cancer, or are oncoproteins implicated in many cancer types. This suggests the meaningfulness and relevance of the model used.

In conclusion, we have demonstrated simple yet effective procedures that allow pinpointing putative TRs governing the observed differential expression pattern. When applied to the whole-genome dataset, our tests successfully retrieved highly relevant TRs.

[1] N. J. Armstrong and M. A. van de Wiel, "Microarray data analysis: from hypotheses to conclusions using gene expression data," Cell Oncol, vol. 26, pp. 279-90, 2004.

[2] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer, "Co-clustering of biological networks and gene expression data," *Bioinformatics*, vol. 18 Suppl 1, pp. S145-54, 2002.

[3] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18 Suppl 1, pp. S233-40, 2002.

[4] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, pp. 910-3, 2002.

[5] www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1323.

[6] ResNet database, http://www.ariadnegenomics.com

[7] S. Novichkova, S. Egorov, and N. Daraselia, "MedScan, a natural language processing engine for MEDLINE abstracts," *Bioinformatics*, vol. 19, pp. 1699-706, 2003.