

Small-module-targeting (SMT) coexpression: prioritizing gene and module specific coexpressions by accounting for expression heterogeneity

Nick Furlotte¹, Hyun Min Kang², Chun Ye³ Eleazar Eskin⁴

Understanding the biological relationships among regulatory elements has been one of the primary challenges in the efforts to model biological systems in a mathematically and computationally tractable form. The rapid development of high-throughput gene expression array and RNA sequencing technologies enables us to infer the landscape of regulatory networks.

Two types of analysis have been popular for inferring regulatory networks from expression data. One is the correlation-based approach, which relies on the pairwise correlation between genes. The other is the causality-based approach, which aims to identify causal regulatory relationships between genes with more rigorous statistical criteria. The simple correlation-based approach has been used to identify many large-sized coregulated gene clusters, many of which have been experimentally verified to be biologically meaningful. The causality-based approach, which focuses on the direction of causal relationships between genes has been used to identify many master regulators that have also been experimentally verified.

One of the main problems with correlation-based approaches is the indirect correlation between genes. When a gene co-regulates hundreds or thousands of other genes, the correlations among the co-regulated genes lead to the formation of large-sized clusters. These clusters contain excessive numbers of indirect coexpressions induced through the master regulator rather than direct coexpressions with the regulator itself. This can complicate the analysis of individual gene-specific regulatory relationships. Moreover, it has been demonstrated that technical bias in the expression data sets such as batch effects, plate effects, ozone levels, and sample preparation bias can generate spurious coexpression between genes, further obscuring the identification of true biological coexpression between regulatory modules. On the other hand, causality-based approaches focus on identifying a relatively small number of strong causal relationship through several steps of stringent criteria given a limited size of data, so their outcome may not directly substitute the typical measure of coexpression between every pair of genes.

In this work, we introduce a novel method for calculating gene coexpression that takes into account expression heterogeneity. By correcting for the dependence structure found within the expression data, our method is able to deprioritize many of the indirect and spurious coexpressions. Our measure of coexpression is termed small-module targeting (SMT) coexpression, and can be used in place of the traditional Pearson or Spearman correlations to find small gene and module specific gene sets.

We applied our approach to an expression data set for *Saccharomyces cerevisiae* segregants and the human HapMap samples. Across these data sets, we observed a significant reduction in the enrichment of strong coexpression because the indirect and spurious coexpressions are deprioritized. By leveraging the pairs of probes targeting the same gene, we are able to show that our method prioritizes gene-specific coexpressions more highly when compared with the Pearson's correlation. We also utilize MIPS functional categories to define small sets of coregulated genes and show that the coexpressions among these small sets are more highly prioritized with our method when compared to Pearson's correlation.

¹Department of Computer Science, University of California, Los Angeles, CA 90095. E-mail: nfurlott@cs.ucla.edu

²Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093. E-mail: h3kang@cs.ucsd.edu

³Bioinformatics Program, University of California San Diego, La Jolla, CA 92093. E-mail: yimmieg@gmail.com

⁴Department of Computer Science, University of California, Los Angeles, CA 90095. Department of Human Genetics, University of California, Los Angeles, CA 90095. E-mail: eeskin@cs.ucla.edu