Independent component analysis: identifying data-driven human gene modules

Jesse Engreitz\textsuperscript{1}, Bernie Daigle\textsuperscript{2}, Yoni Marshall\textsuperscript{1}, and Russ B. Altman\textsuperscript{1,2}

\textsuperscript{1}Stanford University, \textsuperscript{2}Department of Bioengineering, \textsuperscript{3}Department of Genetics

\textbf{motivation}

Cellular physiology, including disease states and drug responses, results from the combined influences of many genes. Experimentalists have now sampled many conditions and cell types, contributing vast amounts of gene expression data that represent many biological processes. The expansion of public microarray databases such as the Gene Expression Omnibus (GEO) allows the use of intelligent data mining approaches to extract information about these biological processes in a data-driven manner. Using 9,395 human microarrays measuring over 20,000 genes, we use independent component analysis to identify functional gene modules, or sets of genes, that describe a wide range of biological conditions.

\textbf{methods}

9,395 human microarrays
Affymetrix HG U133 Plus 2.0
Diverse experimental conditions

Preprocessing
We used hierarchical clustering to reduce the contributions from highly replicated experimental systems. We consolidated clusters of over-represented conditions to create a meta-compendium of 423 meta-arrays.

Independent component analysis
Independent component analysis (ICA) models gene expression data as a linear combination of transcriptional patterns, termed independent components. Given a set of microarrays, ICA identifies components so that statistical independence is maximized. Since ICA is a stochastic method, we run the algorithm 20 times and cluster the component estimates.

Gene modules
Each independent component has a weight for each gene that quantifies relative expression level. For each component, we use a weight cut-off to generate gene modules of over-expressed and under-expressed genes.\textsuperscript{2}

Gene module analysis
We can calculate the expression of each independent component in a new microarray experiment. We predict that gene modules associated with highly-expressed components play an important role in the experiment.

\textbf{results}

We identified 423 independent components and defined 846 gene modules. Annotation using the Gene Ontology (GO) suggests that while some of the gene modules represent known biological processes, some may describe transcriptional programs not well characterized by GO. ICA gives reproducible component estimates when applied to a large compendium of gene expression data, and performs better than PCA in describing the data.

\textbf{application: parthenolide}

9,395 human microarrays
Affymetrix HG U133 Plus 2.0

We ranked these genes based on their variance explained and calculated the number of enriched GO terms. We used hierarchical clustering to reduce the contributions from highly replicated experimental systems. We consolidated clusters of over-represented conditions to create a meta-compendium of 423 meta-arrays.

Independent component analysis
Independent component analysis (ICA) models gene expression data as a linear combination of transcriptional patterns, termed independent components. Given a set of microarrays, ICA identifies components so that statistical independence is maximized. Since ICA is a stochastic method, we run the algorithm 20 times and cluster the component estimates.

Gene modules
Each independent component has a weight for each gene that quantifies relative expression level. For each component, we use a weight cut-off to generate gene modules of over-expressed and under-expressed genes.\textsuperscript{2}

Gene module analysis
We can calculate the expression of each independent component in a new microarray experiment. We predict that gene modules associated with highly-expressed components play an important role in the experiment.

\textbf{conclusion}

We extracted gene modules from a large corpus of expression data using data-driven means, providing a new method for predicting functional relationships between genes. These modules are useful for differential expression analysis and may be applied in a number of other settings, including Gene Set Enrichment Analysis, phenotype classification, drug discovery, and content-based microarray search.

Support
Bio-X Undergraduate Research Award
Bio-X2 Supercomputing Cluster

References