Title

Discovering and validating biological hypotheses from coherent patterns in functional genomics data using associative biclustering

Permalink
https://escholarship.org/uc/item/7bn0f82g

Author

Joachimiak, Marcin P.

Publication Date
2009-02-08

Discovering and validating biological hypotheses from coherent patterns in functional genomics data using associative biclustering

Marcin P. Joachimiak[1,2], Cathy Tuglus[3], Mark van der Laan[3], Adam P. Arkin[1,2,4]

[1]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov/; [2]Lawrence Berkeley National Laboratory, Berkeley, CA, 94720; [3]Department of Biostatistics, University of California, Berkeley, CA, 94720; and [4]Department of Bioengineering, University of California, Berkeley, CA, 94720

Acknowledgements
This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy

Functional genomics confronts researchers with a deluge of new functional genomic experiments and technologies aimed at understanding biological function on the genome scale. For example, the Genomes to Life (GTL) Environmental Stress Pathway Project generates gene expression, gene knockout, proteomic, metabolomic, and protein-protein interaction data. How to rationally construct biological interpretations and determine their significance based on many instances of multiple data types?

Two dimensional clustering, i.e., biclustering, of gene expression data can reveal 'modules' of genes and experiments in which genes exhibit a common pattern. Such modules represent associations between genes and can be used to reconstruct regulatory networks. As functional genomic datasets and data types proliferate it has become advantageous to: a) utilize multiple data types simultaneously, b) determine confidence from combined data, and c) systematically form hypothesis from multiple types of evidence. Basic biclustering has limitations, which can be overcome by algorithms instead designed to search a dataset for biclusters. However, while there is a small number of bicluster search methods which allow for multiple data types, non allow for flexible integration of broadly understood gene and protein features.

We have developed a random-walk statistical algorithm to search for biological modules that maximize a summary criterion. The main novelty of the algorithm lies in modeling three different common data types: gene-by-experiment, gene-by-gene, and gene-by-feature (where 'protein' can be substituted for 'gene'). An overall criterion is computed from a weighted linear combination of summary statistics and correlation measures. Significant features for sets of genes are discovered via a cross-validated R2 of association sub-criterion from data-adaptive fitting. An empirical null distribution provides significance scores for the sub-criteria and overall criterion. The algorithm identifies multiple potentially overlapping biclusters , each with distinct contributions from specific sub-criteria and datasets.

To benchmark module discovery we evaluate this and related methods using a highly annotated functional genomic compendium as well as simulated datasets with synthetic modules. We also present preliminary findings for Saccharomyces cerevisiae and select prokaryotes.