Deep Surveys of Biological Modules: K-biclustering Gene Expression and Phenotype Data

Marcin P. Joachimiak[1*], Cathy Tuglus[2], Mark van der Laan[2], Adam P. Arkin[1,3]
[1] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
[2] Dept. of Biostatistics, University of California, Berkeley, 94720, USA
[3] Dept. of Bioengineering, University of California, Berkeley, 94720, USA
*To whom correspondence should be addressed: MJoachimiak@lbl.gov

Abstract:

Experimental efforts are targeting genomes, cells, and populations of organisms with a widening array of high throughput experimental techniques. In light of this bountiful data it is advantageous to: a) simultaneously query multiple data types to uncover new biological associations, b) jointly determine confidence across data types, and c) systematically form hypothesis from multiple lines of evidence. We have developed an accurate and sensitive biclustering algorithm, Massive Associative K-biclustering (MAK), for the discovery of biological data associations across multiple data types. The algorithm framework models data archetypes, such as object-by-value (e.g. gene expression), object-by-feature (e.g. phylogenetic profiles), and object-by-object (e.g. protein interactions). The objects can be for example genes, proteins, regulators, orthologous sequence families, or experiments. For each data archetype we designed statistical criteria to detect the expected association patterns. True associations in biological data are mostly unknown, therefore to evaluate biclustering methods we design a simulated dataset modeled on yeast gene expression data, with implanted associations forming known patterns. Using this and other evaluations we find that MAK compares favorably to other biclustering methods.

We applied the MAK algorithm to reconstruction of a condition-specific transcriptional co-expression network for Saccharomyces cerevisiae using combinations of gene expression, experimentally determined transcription factor binding, protein interaction, and phylogenetic profile data. Using gene expression data alone we find more biclusters with higher enrichment for known transcription factor regulation and functional terms than other biclustering methods. Pooling biclusters independently discovered using different data type combinations leads to an improvement in most evaluation measures. We also used the discovered biclusters to generate an annotated co-expressed module network, integrating the MAK biclusters, with human-curated groups of experimental conditions, human-curated functional terms, categories, and cellular localization, known regulation, and known sequence motifs. We were able to assign putative functions and regulation to a number of novel biclusters.