Title

Statistical search for coherence in functional genomics data

Authors

Joachimiak, Marcin P.
Tuglus, Cathy
Laan, Mark van der
et al.

Publication Date

2008-06-02

Statistical Search for Coherence in Functional Genomics Data

293/R-131 (1024)

Marcin P. Joachimiak[1,2], Cathy Tuglus[3], Mark van der Laan[3], Adam P. Arkin[41,2,4]

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

http://vimss.lbl.gov/

Introduction

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? Biclustering of gene expression data is an analysis that can reveal modules of genes and experiments.

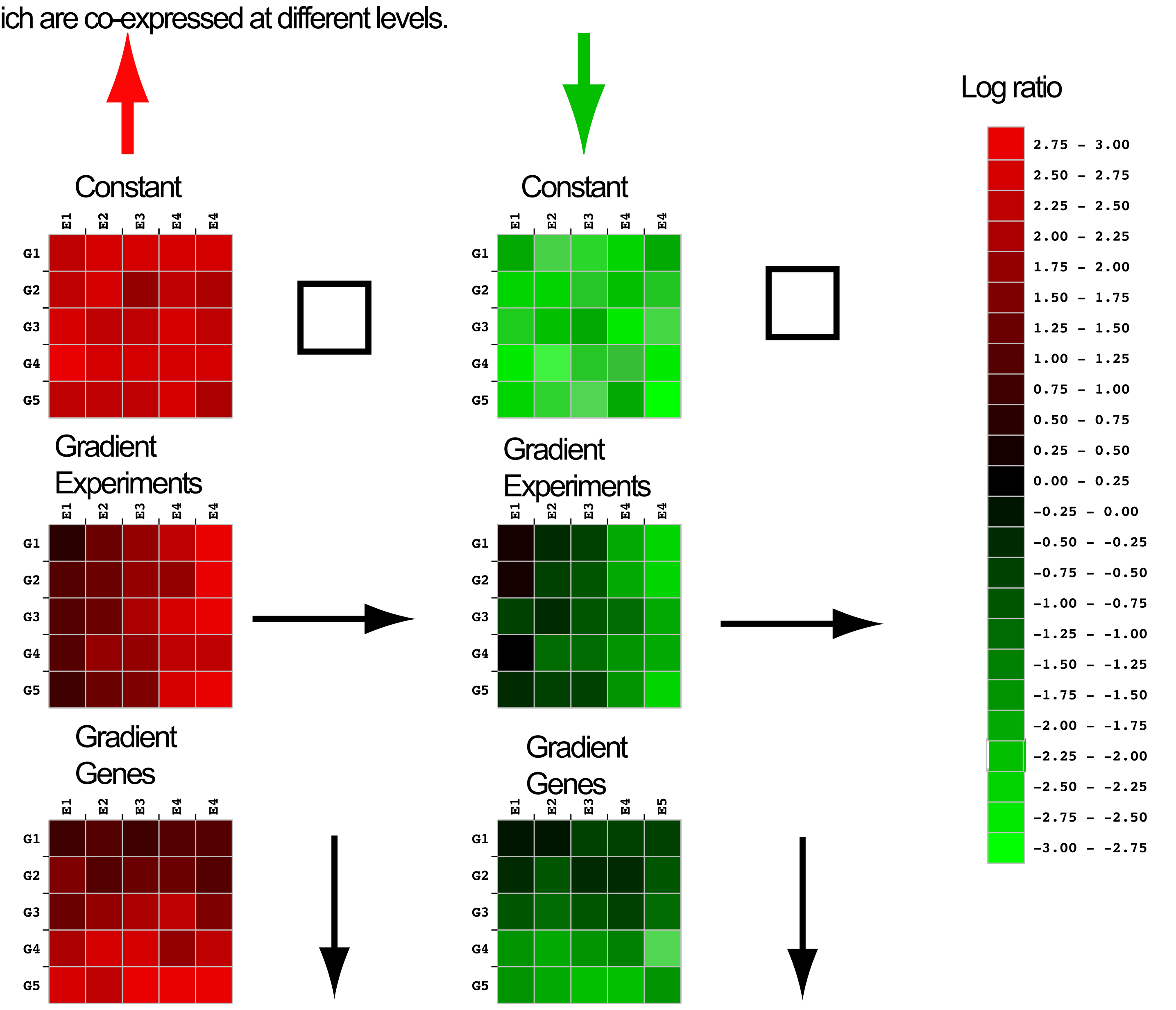
These modules serve to analyze gene-gene associations and reconstruct regulatory networks. As datasets and data types proliferate it has become advantageous to: a) utilize multiple data types simultaneously, b) determine confidence from combined data, and c) systematically form hypothesis from multiple types of evidence. A recent method, CMONKEY (Reiss et al. BMC Bioinformatics 2006), searches for co-regulated genes using simulated annealing and a Markov chain bicluster model with multi-parametric logistic regression for module membership based on gene expression, association networks, -and sequence motifs.

We have developed a random-walk algorithm to search for biological modules that maximize a summary criterion. The main novelty of the algorithm lies in modeling three common data types: gene-experiment, gene-gene, and gene-feature. The summary criterion is computed from a weighted linear combination of correlation measures.

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

Coherent patterns in noisy data: gene expression

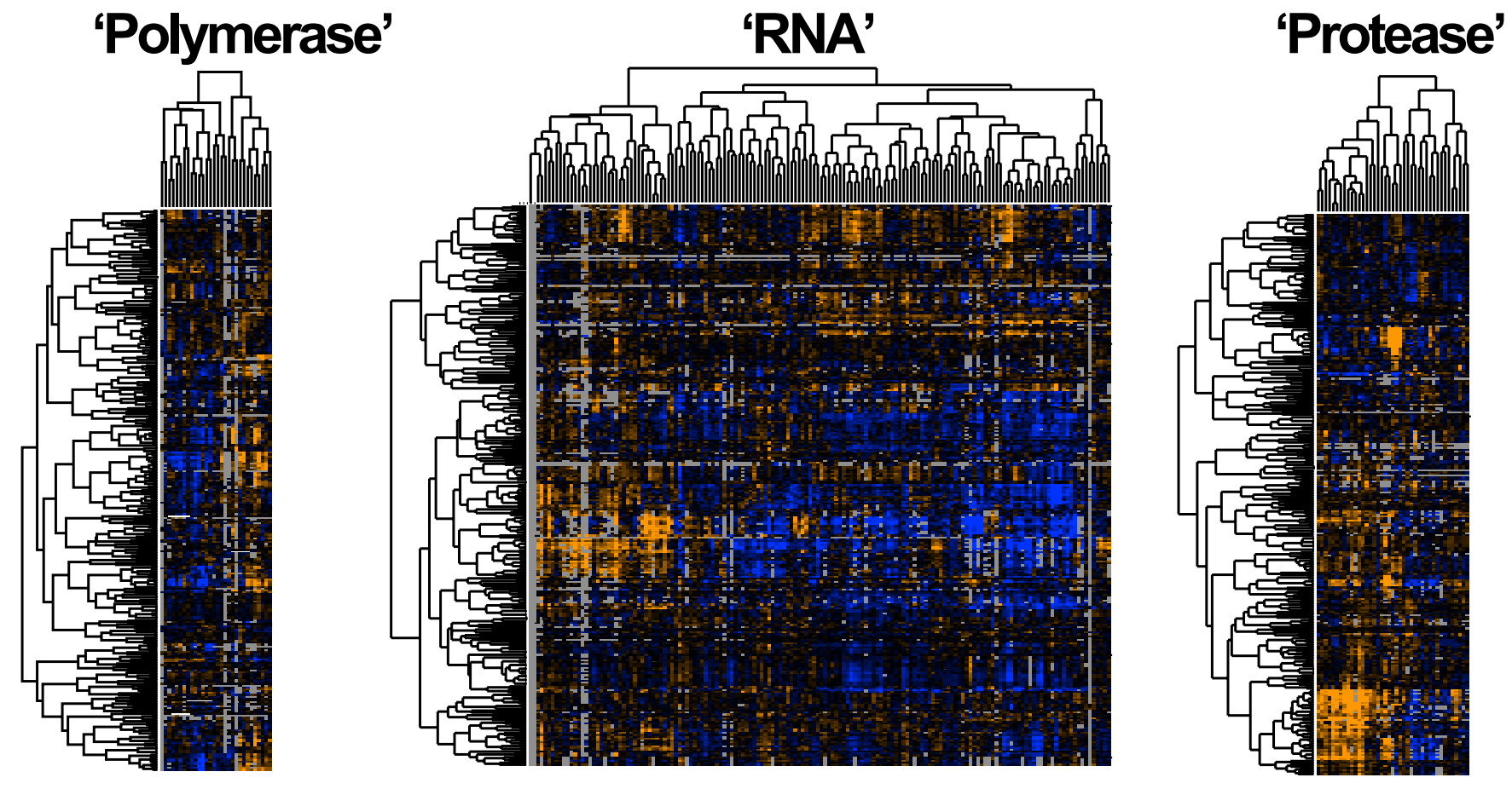
Shown are examples of coherent gene expression data patterns. The first pattern is the case of constant values across genes and experiments, with ribosomal genes being an example. The second pattern is characterized by a gradient in the experiment dimension, indicating a concerted change in expression over a series of experimental conditions such as time points or stressor concentrations. The third pattern is characterized by a gradient in the experiment dimension and this is expected for sets of genes which are co-expressed at different levels.



Standard clustering to detect transcriptonal modules

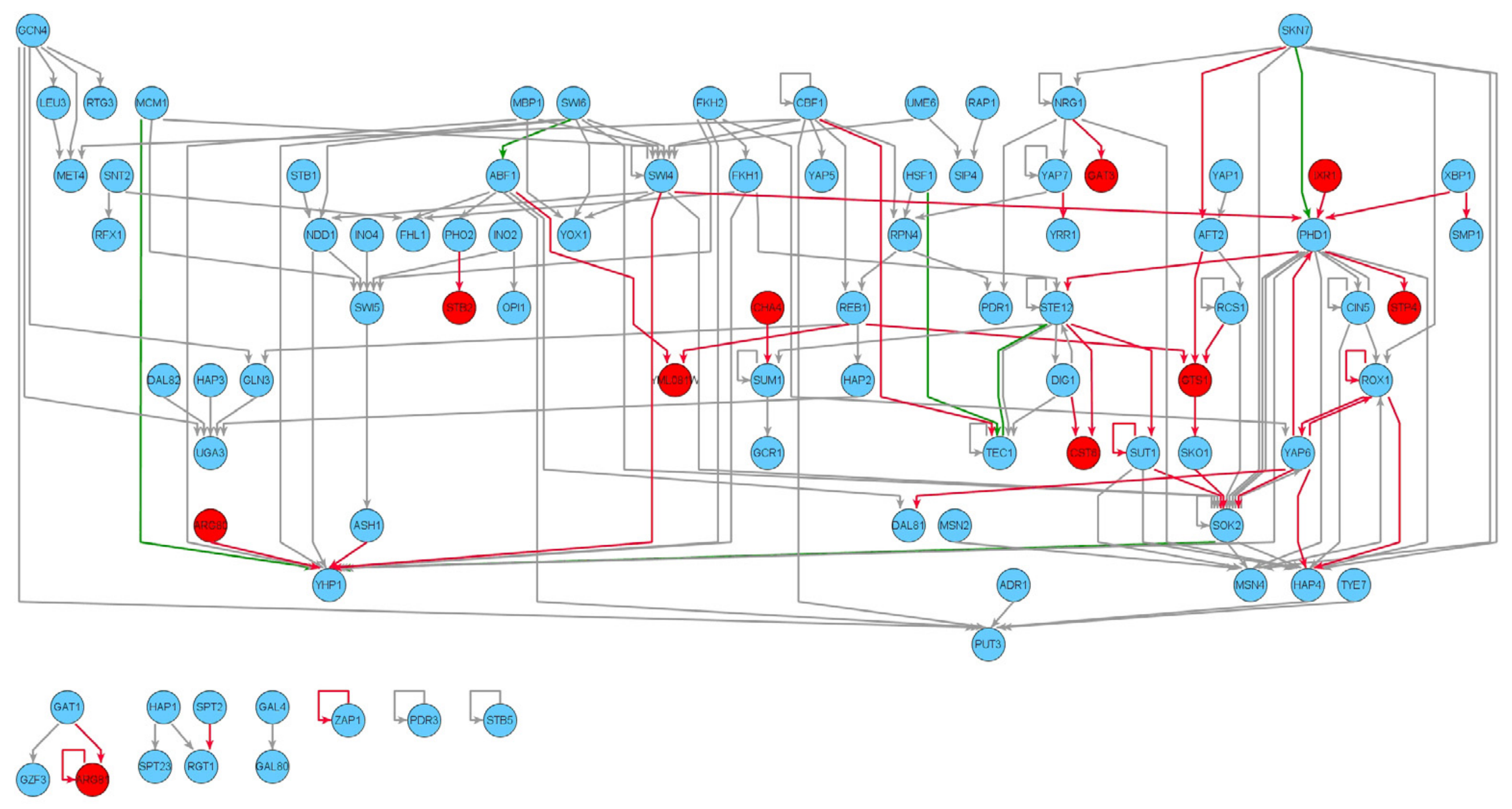
Desulfovibrio vulgaris Hildenborough biclustered sets of genes whose annotations match keywords (www.microbesonline.org).

Limitations:
- Uses information from all experiments.
- Each gene/experiment pair can only be assigned to one module.
- Liimted to a single dataset of a single type.



Modules allow to reconstruct transcriptional networks

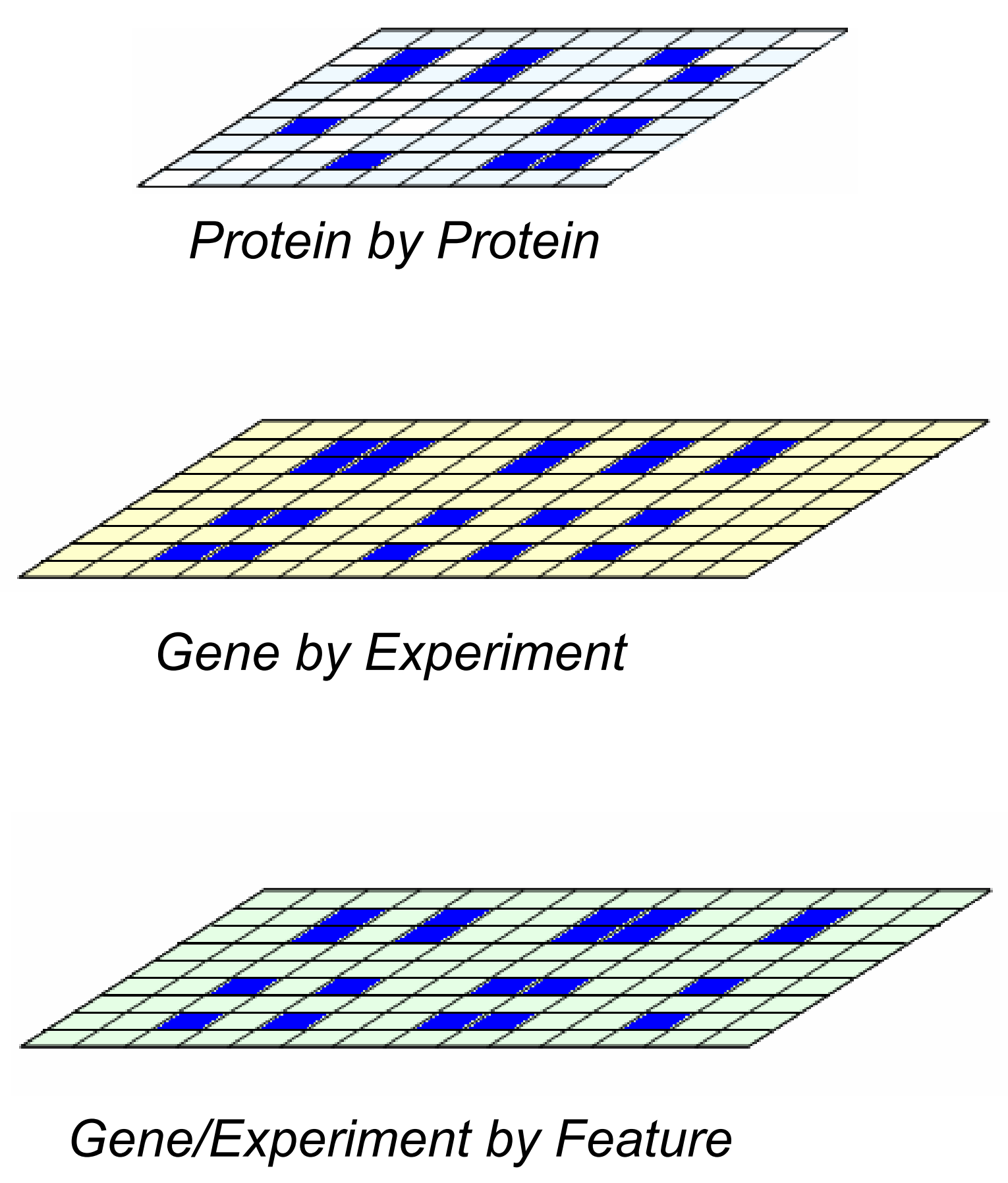
The *Saccharomyces cerevisiae* regulatory network (MacIsaac et al. BMC Bioinformatics 2006).



Functional genomics data fusion

The results of functional genomics research are represented by a variety of data types including continuous measurements, binary ones such as protein-protein interactions, and gene/protein as well as experiment features usually occurring in the form of text labels (e.g., functional annotations) or string representations (e.g., phylogenetic profiles). Each data type requires specific infrastructure and statistical frameworks. We have developed a statistical data mining random walk algorithm that allows to incorporate multiple datasets and types.

Each data type is accommodated using different criteria, which we refer to as subcriteria, geared specifically for its data structure. The algorithm searches for coherent patterns within a base data matrix of gene by experiment. Our search criterion is a weighted linear combination of the subcriteria. The weights are user specified and determine the overall importance of each subcriteria, providing the user some control over the search process such as down-weighting less relevant or less confident datasets.



Criterion = ∑Sub-Criteria

Proportion score

$$\mathrm{MSE}(\overline{X}) = \mathrm{E}((\overline{X}-\mu)^2) = \left(\frac{\sigma}{\sqrt{n}}\right)^2$$

Mean Squared Error (MSE), Row MSE, **Correlation**
- Significance score calculated from an empirical null distribution created from random draws of all allowed bicluster sizes.
- Probability that a value as extreme or more extreme would occur if bicluster was randomly sampled.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

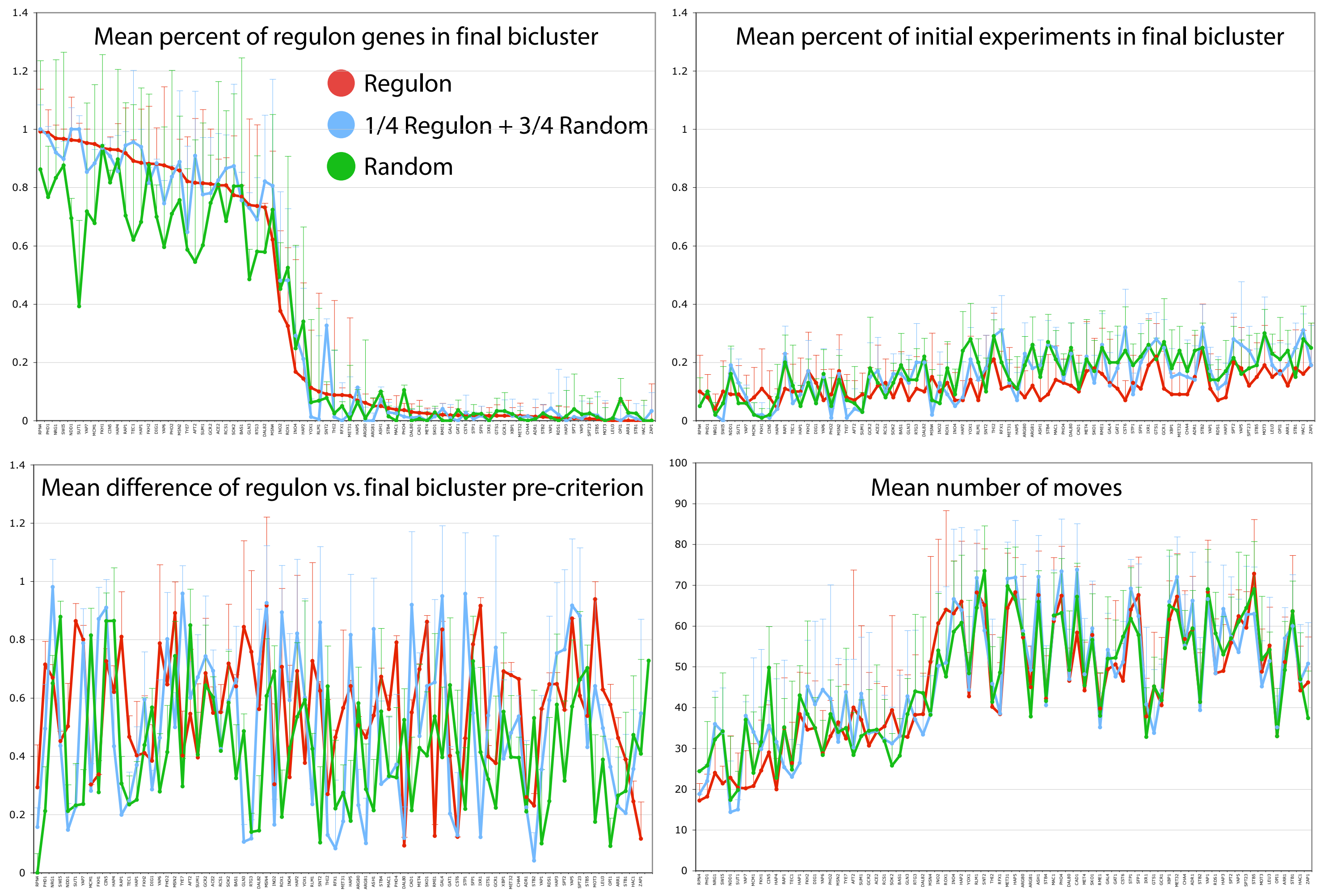
Cross-validated R^2
- Calculated using data-adaptive software with polynomial spline fitting.
- Selects subset of features using cross-validation.

Current subcriteria are restricted to significance scores and cross-validated R-squared measures of association. Significance scores are calculated with respect to a null distribution from randomly sampling the space of biclusters of all allowed sizes. Currently, significance scores are used for the following subcriteria: bicluster expression mean squared error, bicluster expression column/row mean square error, and proportion of interacting genes within the set of genes. To incorporate feature data from the gene or experiment set, we use the cross-validated R-square of association of bicluster membership with a set of gene and/or experiment feature data using data-adaptive software. The data-adaptive algorithm, currently a polynomial spline based algorithm is used (polymars), allow to use all features or merely a subset of features.

With multiple subcriteria the algorithm can identify multiple potentially overlapping global maxima, each with distinct contributions from specific sub-criteria and datasets. To date our algorithm supports compendia of continuous measurements of the form gene-by-experiment, gene-by-gene binary data, and gene-by-experiment features.

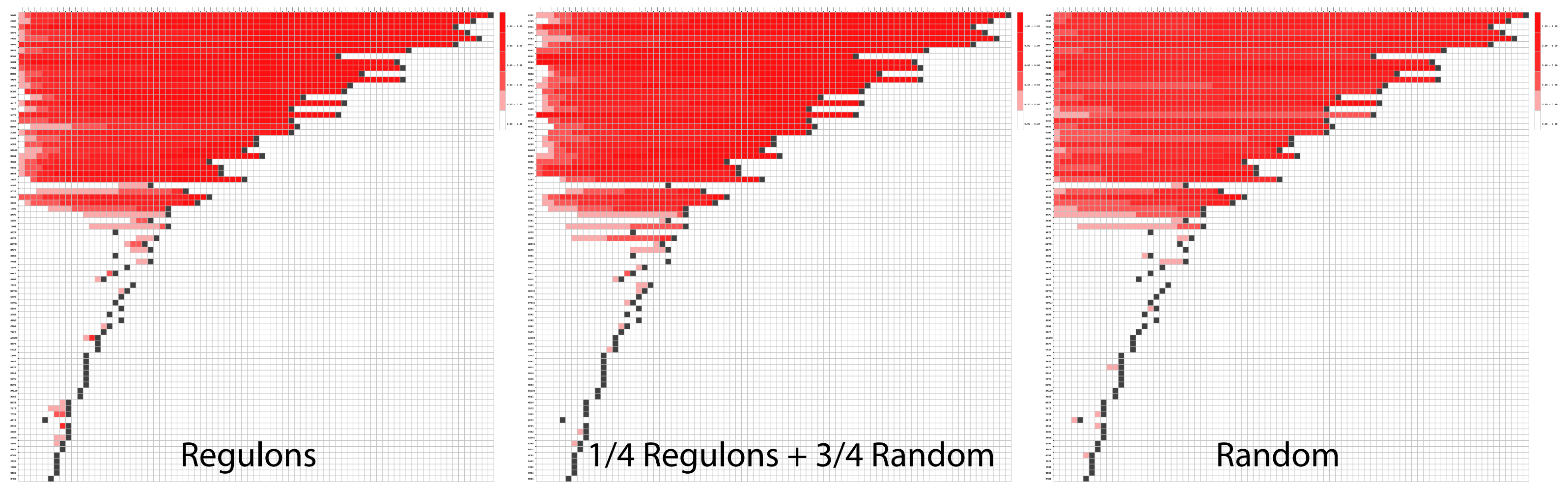
Bicluster search evaluation

To evaluate the performance of our bicluster search method we assessed the results of bicluster searches with known *Saccharomyces cerevisiae* regulons (MacIsaac et al. 2006 BMC Bioinformatics). As the gene expression data for the search we used data on 173 conditions from Gasch et al. Mol. Biol. Cell 2002. Shown are summary statistics based on 10 independent runs per regulon (80 regulons total, the 11 largest and 24 smallest ones were excluded). Each plot lists regulons on the x-axis sorted by highest to lowest percentage of regulon genes in the final bicluster. Values are the mean for the regulon and errors bars indicate the standard deviation. For the '1/4 Regulon + Random' and 'Random' results, searches were performed using the same random experiments as selected for 'Regulons'. The percentage of genes from the regulon present in the final bicluster is largely determined by regulon size, however intermediate size regulons vary widely in their behavior. Larger regulons tend to keep fewer of the initial (randomly selected) experiment and also require fewer algorithm moves.



Regulon stability

The series of regulon random walk searches allows to assign an occupancy for each gene with respect to the final bicluster identified in each search. Most genes in large regulons have high occupancy, meaning that they are nearly always present in the final bicluster. Genes in smaller regulons generally have much lower occupancy. Large regulons are more likely to be completely defined (or over-defined) while smaller ones are more likely to be under-defined. Since the gene expression compendium used did not sample many possible conditions, it is possible that under the tested conditions some of the regulons are not differentially expressed.



Conclusions

The increasing volume and diversity of experimental and sequence-based datasets requires efficient and flexible statistical methods to discern significant patterns and to validate biological hypothesis. We have developed a multi-purpose algorithm which searches multiple datasets and data types simultaneously, allowing to identify inter-dataset relationships and to enrich biological hypothesis with significant data patterns.

Acknowledgment

ESPP2 (MDCASE) is part of the Virtual Institute for Microbial Stress and Survival (VIMSS) 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.