

Combinatorial Analysis of Perturbational Gene Expression Compendia

Steven Maere^{1,2,*}, Patrick Van Dijck^{3,4}, Martin Kuiper^{1,2}

1. Department of Plant Systems Biology, VIB, B-9052 Ghent, Belgium
2. Department of Molecular Genetics, Ghent University, B-9052 Ghent, Belgium
3. Department of Molecular Microbiology, VIB, B-3001 Leuven, Belgium
4. Laboratory of Molecular Cell Biology, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium

*E-mail: steven.maere@psb.ugent.be

Background

Large-scale compendia of gene expression profiles under chemical or genetic perturbations constitute an invaluable resource from a systems biology perspective. However, the perturbational nature of such data imposes specific requirements on the methods used to analyze them. In particular, the distance measures used in traditional clustering algorithms have difficulties in detecting one of the most prominent features of perturbational data, namely partial correlations between expression profiles. Biclustering methods on the other hand are specifically designed to capture such partial correlations. However, most biclustering algorithms do not provide measures for pair-wise expression correlation between genes, but rely on emergent properties of groups of genes and conditions (modules) in order to identify statistically significant subpatterns in the data. This reliance complicates the elucidation of less modular regions in the underlying transcriptional network.

Results

We introduce a novel method to extract (partial) expression correlations and transcriptional modules from perturbational gene expression data, based on the use of combinatorial statistics and graph-based clustering. Briefly, gene expression profiles are discretized into three categories (upregulated, downregulated, unchanged) based on p -values for differential expression. For each pair of profiles, we then assess the probability that the observed overlap of upregulated and downregulated fields is generated by chance. The resulting correlation p -values are corrected for multiple testing and translated to edges in a coexpression network, which is then clustered into (overlapping) expression modules using a graph clustering procedure that identifies densely connected components in the network. Relevant condition sets are then determined for all modules and the modules are screened for enrichment of Gene Ontology categories and transcription factor binding sites. Finally, a regulation program is learned for each module in an attempt to explain the expression behavior of the module's genes as a function of the expression of a limited set of regulators (transcription factors and signal transducers). We have incorporated this methodology in a software tool called ENIGMA [1]. We show that ENIGMA outperforms other methods on both modular and non-modular artificial data (see Figure 1). We also applied ENIGMA to the Rosetta compendium of expression profiles for *Saccharomyces cerevisiae* [2]. In particular, we were able to discriminate a subset of candidate mating-related genes whose expression appears to

be regulated by the transcription factor Tec1, although their promoters lack Tec1 binding sites (see Figure 2). We propose that Tec1 in fact mediates antisense expression of these genes through interaction with nearby Ty1 long terminal repeats (LTRs), and that this effect could be functionally relevant for the mating process. We present evidence that at least one Ty1 LTR-associated gene, namely *YLR343W*, causes a mating-related phenotype upon deletion.

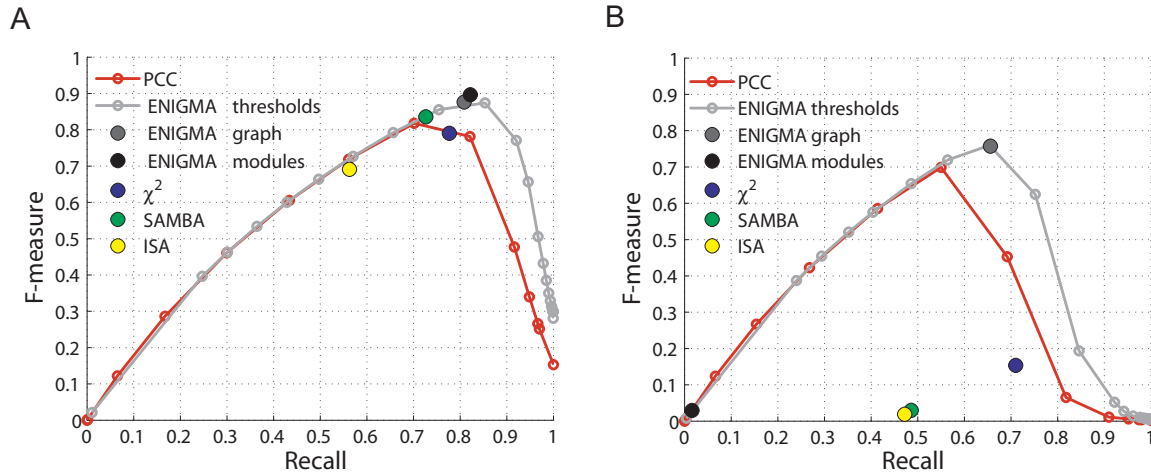


Figure 1: Performance of ENIGMA versus other coexpression measures and biclustering methods on (A) modular and (B) non-modular artificial gene expression data (1000 genes on 100 conditions, (A) 20 overlapping biclusters of varying size and (B) 500 partial expression correlations between individual genes). We tested the performance of ENIGMA on two levels by assessing the overlap between the artificial input correlation network and (i) the network of significant correlations obtained in the first step of the ENIGMA algorithm (The 'ENIGMA thresholds' series shows the performance for different uncorrected p -value thresholds. 'ENIGMA graph' denotes the network obtained when correcting the p -values for multiple testing at FDR=0.05. The performance of 'ENIGMA graph' is close to the 'ENIGMA thresholds' optimum); (ii) the module network inferred by ENIGMA ('ENIGMA modules'). The performance of the combinatorial statistic built into ENIGMA was compared with that of two other similarity measures, namely Pearson's correlation coefficient (PCC, different thresholds) and the χ^2 -statistic (FDR=0.05). We also compared ENIGMA with two established biclustering methods, namely SAMBA [3] and ISA [4]. ENIGMA finds very few modules in the non-modular data, hence the low recall and F-measure of 'ENIGMA modules' in panel B.

References

- [1] ENIGMA [<http://bioinformatics.psb.ugent.be/ENIGMA/main.htm>]
- [2] Hughes, TR, et al (2000) *Cell* **102**:109-126.
- [3] Tanay, A, et al (2002) *Bioinformatics* **18 Suppl 1**:136-144.
- [4] Bergmann, S, et al (2003) *Phys Rev E Stat Nonlin Soft Matter Phys* **67**:031902.

