TOWARDS SYSTEM LEVEL MODELING OF FUNCTIONAL MODULES AND REGULATORY PATHWAYS USING GENOME-SCALE DATA

Tom Michoel[1,2,*], *Anagha Joshi*[1,2], *Eric Bonnet*[1,2], *Vanessa Vermeirssen*[1,2] *and Yves Van de Peer*[1,2]

[1]Department of Plant Systems Biology, VIB
[2]Department of Plant Biotechnology and Genetics, Ghent University,
Technologiepark 927, B-9052 Gent, Belgium
[*] Corresponding author, E-mail: tom.michoel@psb.vib-ugent.be

ABSTRACT

Understanding complex biological processes such as development and pathology of multicellular organisms at a system level requires the study of dynamic networks of interacting molecules. High-throughput methods have generated large-scale, static networks of physical interactions which may occur or not depending on spatial, temporal or context-specific variation, while genome-wide microarray datasets measure the functional state of a cell precisely under such varying conditions. Here we present recent results towards integrative, system level modeling of functional modules and regulatory pathways using diverse genome-scale data.

1. INTRODUCTION

Networks of physical interactions between DNA, RNAs, proteins and metabolites represent the functional organization of multiple levels of regulation in a cell [1]. Regulatory networks can be probed experimentally by measuring genome-wide expression levels under various conditions. The correlations between genes in large-scale expression compendia are indirect manifestations of the underlying network architecture and can be analyzed for the purpose of network reconstruction [2]. In addition, more and more high-throughput experimental methods are becoming available to map the various networks of physical interactions in a cell directly [3].

Biological networks are significantly different from random networks and already purely on the basis of their topology, we can understand general principles about their function [4] or evolution [5]. However, physical interaction networks are inherently static, with edges which may occur or not in specific cellular locations, developmental timepoints or environmental conditions. Furthermore, even the knowledge of *where* and *when* each interaction occurs is not sufficient to know *how* a particular regulatory influence is exerted. An important challenge for systems biology is therefore to develop computational methods for predictive modeling of dynamic and context-specific networks by integrating diverse sources of genome-scale data [6].

In this paper we present recent results towards attaining this goal. We have developed a set of algorithms for inferring transcriptional modules and their condition dependent regulatory programs from gene expression data [7, 8, 9, 10, 11, 12]. Regulatory interactions inferred from expression data are often indirect, involving several hidden, intermediate physical interactions. In order to characterize these indirect regulation mechanisms, we have introduced the notion of regulatory path motifs, short significantly enriched paths in physical networks which connect cause-effect protein pairs in perturbational expression data [13].

2. MODEL-BASED INFERENCE OF TRANSCRIPTIONAL REGULATORY MODULES

Reconstructing transcriptional control networks from gene expression data has been an important subject for many years and a lot of methods have been developed (reviewed in [2, 14, 15, 16]). An important class of methods are those which also infer a model of the biological system which explains the observed expression patterns and generates testable hypotheses. Such models can take the form of probabilistic graphical models [2, 17, 18] or simplified kinetic equation models [19]. It is known that the correlation in expression between coregulated gene pairs is much stronger than between regulator-target pairs, especially in eukaryotes [20]. Network inference methods based on clustering genes into modules with a putative common regulation program therefore come as a natural approach and have the additional advantage of a significant dimensionality reduction [17, 18].

We have extended the approach of [17] to infer regulatory modules and their condition-specific regulators from gene expression data by using a more representative solution extracted from an ensemble of possible statistical models to explain the data. We use a Gibbs sampling approach for two-way clustering of genes and conditions to generate an ensemble of partially overlapping partitions of genes into modules and produce an averaged solution. This centroid solution consists of so-called *tight clusters*, subsets of genes which consistently cluster together in almost all local optima [8]. Furthermore we use a probabilistic method for learning regulatory programs. These regulatory programs model the expression level x of the genes in a module as a mixture of normal distributions, conditional on the expression levels of a limited set of reg-

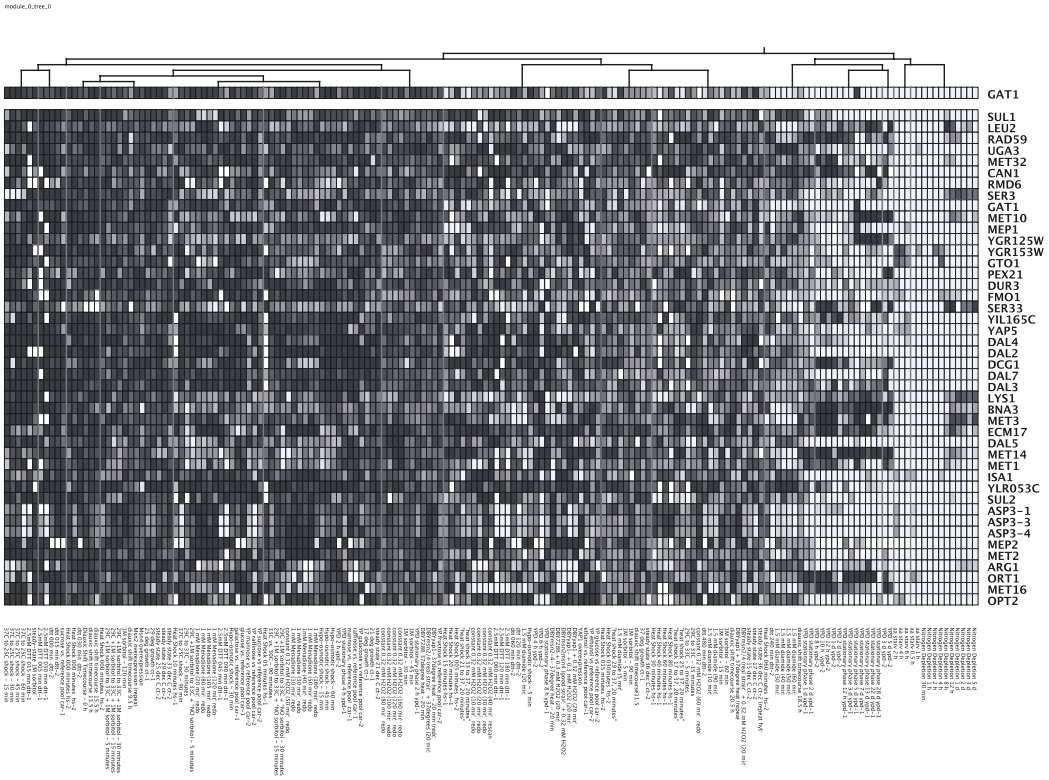

Figure 1. Example of a transcriptional regulatory module in yeast inferred by LeMoNe. Rows represent genes, columns conditions and the grayscale is proportional to the expression level (black = low, white = high expression).

ulators $\mathcal{R}$:

$$p\big(x \mid \{x_r, r \in \mathcal{R}\}\big) = \sum_n \alpha_n\big(\{x_r, r \in \mathcal{R}\}\big)\, p(x \mid \mu_n, \tau_n),$$

where the value of the mixture component weights α_n is determined by the regulator expression levels x_r through a fuzzy decision tree [9]. Together, the Gibbs sampling cluster algorithm and probabilistic regulatory program learning provide a computationally efficient method to identify statistically reliable modules and their condition-specific regulatory programs which can be used for generating experimentally verifiable hypotheses. Our algorithm is freely available for academic use as a software package called LeMoNe. Compared to the original, direct-ascent optimization strategy of [17], ensemble-averaged clusters are functionally more coherent and regulators assigned consistently in multiple solutions are more often supported by literature [9]. Compared to the best information-theoretic method, CLR [21], it is found that in the prokaryote *E. coli* CLR has a higher AUC-score with respect to the network of known direct transcriptional interactions, while LeMoNe has a higher AUC-score in the eukaryote *S. cerevisiae* [10]. More importantly, both methods recover largely distinct parts of the underlying regulatory networks and optimal results are obtained by integrating the predictions of both methods [10].

Figure 1 shows a typical example of a module and regulatory program in yeast, keeping only the most significant regulator as determined by the frequency of occurrence in an ensemble of optimal solutions, inferred from an expression compendium of environmental stress conditions [22]. Each normally distributed mixture component is represented by a cluster of conditions (leaf node in the tree). The predicted regulator is GAT1, a transcriptional activator of genes involved in nitrogen catabolite repression. 28 of the 44 genes in the module are involved in nitrogen compound metabolic process (hypergeometric p-value 10^{-23}, corrected for multiple testing). The regulatory program predicts that the module is upregulated if GAT1 is upregulated (rightmost condition clusters), and most of these conditions belong to a nitrogen depletion time course experiment. Only one gene in the module is known to be bound in its promoter by GAT1 in the ChIP-chip dataset of [23], but eight genes (p-value 10^{-6}) are confirmed in the YEASTRACT database [24]. Most likely this module indeed consists of *bona fide* direct or indirect targets of GAT1.

In multicellular organisms, differential gene expression drives development, function and pathology. Reconstructing transcriptional modules and their regulatory programs should lead to a better understanding of development in health and disease. We applied LeMoNe to a compendium of expression profiles for the worm *C. elegans* and obtained functionally coherent coexpression modules and predicted regulators related to similar biological processes as their target modules [11]. In addition to checking for known protein-DNA interactions, we also searched

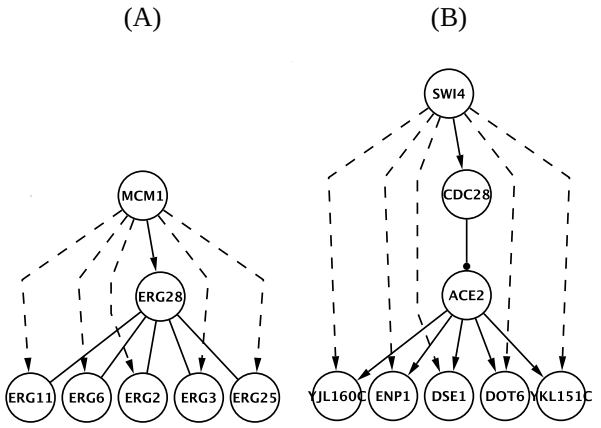

Figure 2. Examples of regulatory path modules identified by Pathicular. (A) A TRI-PPI module in knockout data. (B) A TRI-PHI-TRI module in overexpression data. Legend: — (PPI), → (TRI), —● (PHI), ⇢ (perturbational interactions). See main text for details.

for indirect regulatory paths between predicted regulators and target genes. For several modules, complete regulatory paths could be identified. Transcription is influenced by protein-protein interactions and (de)phosphorylation events, explaining why the components in a hidden regulation path could not be detected from gene expression data alone. Regulation by microRNAs is another form of posttranscriptional regulation with critical functions in development and disease. We used a public dataset of expression profiles for 89 human tumor and normal tissue samples measuring both mRNA and microRNA levels, and inferred mixed regulatory programs of transcription factors and microRNAs to predict the differential expression of transcriptional modules [12]. A detailed analysis showed that the assignment of microRNAs to several modules is functionally coherent and supported by literature. We further validated one of those modules experimentally by overexpression and inhibition of the assigned microRNA and showed that the expression of genes in the module indeed changes significantly as predicted by LeMoNe [12]. Like for transcription factors, the actual connection of physical interactions between predicted microRNAs and their target modules is usually through a hidden regulatory path not observable in the expression data.

3. IDENTIFICATION OF REGULATORY PATHWAYS

Transcriptional regulators inferred from correlations in expression data can often be validated by a knockout or overexpression experiment, but as discussed in the previous section, the regulatory influence is often exerted through a hidden regulatory path. In order to characterize these regulatory pathways more systematically, we introduced the notion of *regulatory path motifs*, short paths in integrated networks of physical interactions which occur significantly more often than expected by chance, as measured by comparison to degree-preserving randomized data, between transcription factors and their targets in perturbational expression data [13]. We combined perturbational data for 157 knockout and 55 overexpression experiments with large-scale physical networks of transcriptional (TRI) [23], protein-protein (PPI) [25] and phosphorylation (PHI) [26] interactions. Among all 39 paths of length at most three, eight regulatory path motifs were significantly enriched across multiple transcription factors, and thus represent general indirect regulation mechanisms (TRI, TRI–TRI, TRI–PPI, PPI–TRI, PPI–PHI–TRI, PPI–TRI–TRI, TRI–PHI–TRI, TRI–PPI–TRI) [13]. Our method for analyzing regulatory path motifs is available as a Cytoscape plugin Pathicular [13].

Like topological network motifs [4], regulatory path motifs cluster into larger modules. Figure 2(A) shows an example of a TRI-PPI regulatory path module. If the transcription factor MCM1 is knocked out, five enzymes in the ergosterol biosynthesis pathway are differentially expressed, although none of them is directly regulated by MCM1. However MCM1 does regulate ERG28, a protein which tethers ergosterol biosynthesis enzymes to the endoplasmic reticulum and physically interacts with all of them. A likely explanation is thus that the disruption of ERG28 by knocking out MCM1 prevents proper functioning of the ergosterol biosynthesis pathway, which in turn affects the expression levels of the corresponding enzymes. Figure 2(B) shows an example of a TRI-PHI-TRI regulatory path module. Five genes are differentially expressed if the cell-cycle transcription factor SWI4 is overexpressed but none of them are known to be direct targets. The indirect regulatory path inferred by Pathicular is that they are transcriptionally regulated by ACE2, another cell-cycle transcription factor, whose activity is controlled by phosphorylation by CDC28, the central coordinator of the yeast cell cycle, which is a direct transcriptional target of SWI4.

4. CONCLUSIONS

The automatic identification and modeling of context-specific functional modules and regulatory pathways is one of the great challenges of computational systems biology. We have shown that transcriptional coexpression modules and their condition-specific regulators can be reliably inferred from gene expression data using probabilistic methods, but that the predicted regulators often act indirectly via intermediate transcriptional, protein-protein or phosphorylation interactions. By combining perturbational expression data with integrated networks of physical interactions, we have shown that some of these indirect regulatory mechanisms are generic and manifest themselves as so-called regulatory path motifs. Further development of these integrative methods will undoubtedly lead to a better system-level understanding of complex processes in development and disease of multicellular organisms. The software presented in this paper is freely available for academic use and can be downloaded from our homepage http://bioinformatics.psb.ugent.be/.

5. ACKNOWLEDGEMENTS

This work was supported by IWT (SBO-BioFrame) and IUAP P6/25 (BioMaGNet).

6. REFERENCES

[1] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet*, vol. 5, pp. 101–113, 2004.

[2] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 308, pp. 799–805, 2004.

[3] X. Zhu, M. Gerstein, and M. Snyder, "Getting connected: analysis and principles of biological networks," *Genes & Dev*, vol. 21, pp. 1010–1024, 2007.

[4] U. Alon, "Network motifs: theory and experimental approaches," *Nat Rev Genet*, vol. 8, pp. 450–461, 2007.

[5] C. G. Knight and J. W. Pinney, "Making the right connections: biological networks in the light of evolution," *BioEssays*, vol. 31, pp. 1080–1090, 2009.

[6] T. M. Przytycka, M. Singh, and D. K. Slonim, "Toward the dynamic interactome: it's about time," *Brief. Bioinformatics*, vol. 11, pp. 15–29, 2010.

[7] T. Michoel, S. Maere, E. Bonnet, A. Joshi, Y. Saeys, T. Van den Bulcke, K. Van Leemput, P. van Remortel, M. Kuiper, K. Marchal, and Y. Van de Peer, "Validating module networks learning algorithms using simulated data," *BMC Bioinformatics*, vol. 8, pp. S5, 2007.

[8] A. Joshi, Y. Van de Peer, and T. Michoel, "Analysis of a Gibbs sampler for model based clustering of gene expression data," *Bioinformatics*, vol. 24, no. 2, pp. 176–183, 2008.

[9] A. Joshi, R. De Smet, K. Marchal, Y. Van de Peer, and T. Michoel, "Module networks revisited: computational assessment and prioritization of model predictions," *Bioinformatics*, vol. 25, no. 4, pp. 490–496, 2009.

[10] T. Michoel, R. De Smet, A. Joshi, Y. Van de Peer, and K. Marchal, "Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks," *BMC Systems Biology*, vol. 3, pp. 49, 2009.

[11] V. Vermeirssen, A. Joshi, T. Michoel, E. Bonnet, T. Casneuf, and Y. Van de Peer, "Transcription regulatory networks in *Caenorhabditis elegans* inferred through reverse-engineering of gene expression profiles constitute biological hypotheses for metazoan development," *Mol. BioSyst.*, vol. 5, pp. 1817–1830, 2009.

[12] E. Bonnet, M. Tatari, A. Joshi, T. Michoel, K. Marchal, G. Berx, and Y. Van de Peer, "Module network inference from a cancer gene expression data set identifies microRNA regulated modules," *PLoS One*, vol. 5, pp. e10162, 2010.

[13] A. Joshi, T. Van Parys, Y. Van de Peer, and T. Michoel, "Characterizing regulatory path motifs in integrated networks using perturbational data," *Genome Biology*, vol. 11, pp. R32, 2010.

[14] T. S. Gardner and J. J. Faith, "Reverse-engineering transcription control networks," *Phys Life Rev*, vol. 2, pp. 65–88, 2005.

[15] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo, "How to infer gene networks from expression profiles," *Mol Syst Biol*, vol. 3, pp. 78, 2007.

[16] H. J. Bussemaker, B. C. Foat, and L. D. Ward, "Predictive modeling of genome-wide mRNA expression: from modules to molecules," *Annu Rev Biophys Biomol Struct*, vol. 36, pp. 329–347, 2007.

[17] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nat Genet*, vol. 34, pp. 166–167, 2003.

[18] M. A. Beer and S. Tavazoie, "Predicting gene expression from sequence," *Cell*, vol. 117, pp. 185–198, 2004.

[19] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson, "The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*," *Genome Biol*, vol. 7, pp. R36, 2006.

[20] M. J. Herrgård, M. W. Covert, and B. o. Palsson, "Reconciling gene expression data with known genome-scale regulatory network structures," *Genome Res*, vol. 13, pp. 2423–2434, 2003.

[21] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles," *PLoS Biol*, vol. 5, pp. e8, 2007.

[22] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic expression programs in the response of yeast cells to environmental changes," *Mol Biol Cell*, vol. 11, pp. 4241–4257, 2000.

[23] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young, "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, pp. 99–104, 2004.

[24] M. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A. Fernandes, N. Mira, M. Alenquer, A. Freitas, A. Oliveira, and I. Sà-Correia, "The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae," *Nucleic Acids Res.*, vol. 34, pp. D446–451, Jan 2006.

[25] T. Reguly, A. Breitkreutz, L. Boucher, B.-J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O. G. Troyanskya, T. Ideker, K. Dolinski, N. N. Batada, and M. Tyers, "Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*," *J Biol*, vol. 5, pp. 11, 2006.

[26] J. Ptacek, G. Devgan, G. Michaud, H. Zhu, X. Zhu, J. Fasolo, H. Guo, G. Jona, A. Breitkreutz, R. Sopko, R. McCartney, M. Schmidt, N. Rachidi, S. Lee, A. Mah, L. Meng, M. Stark, D. Stern, C. De Virgilio, M. Tyers, B. Andrews, M. Gerstein, B. Schweitzer, P. Predki, and M. Snyder, "Global analysis of protein phosphorylation in yeast," *Nature*, vol. 438, pp. 679–684, Dec 2005.