Integrating microRNA expression data and other sources in module network reconstruction

Eric Bonnet
Bioinformatics and Systems Biology
Department of Plant Systems Biology
Technologiepark 927, 9052 Gent, Belgium
eric.bonnet@psb.vib-ugent.be

Abstract

Biological functions arise from interaction between a defined set of components, thereby defining modules. Systems biology approaches try to infer module networks from high-throughput 'omics' data. We have created a module network algorithm relying on probabilistic optimization techniques, capable of integrating heterogeneous data types for the candidate regulators. We summarize here the results of the application of this algorithm on two different datasets linked to cancer.

Modular cell biology

Most biological functions arise from interactions among many different components (proteins, DNA, RNA, metabolites). Those discrete sets of components form different *modules*, a critical level of biological organization. Many well-established biological functions can be cited as examples of modular structures, like protein synthesis, DNA replication, glycolysis, signal transduction. Some of those modules have even been reconstructed *in vitro*. The modules can be insulated from or connected to each other, thereby forming a dynamical network structure. Functional modules are not necessarily rigid

structures and their components may belong to various modules at different times. It is also likely that evolution is shaping the composition and the functional relationships between the modules. Those modules can hardly be discovered by 'gene-centric' approaches, that have been heavily used in molecular biology so far. More global and systematic approaches are necessary to identify the core components of the modules, which is one of the main goals of the emerging field of systems biology [3].

Module network algorithms

During the last decade, biological research has switched from a relatively data-poor science, mostly qualitative, to a data-rich, quantitative one. This change is of course exemplified by the rapid adoption of the microarray technology to measure gene expression data on a genomic scale in many organisms and under a wide variety of conditions. There are nowadays several technological platforms available to potentially measure all types of cellular components in a high-throughput, genome-scale manner. Those techniques are collectively referred to as 'omics' data production (genomics, transcriptomics, proteomics, metabolomics, localizomics, etc.) [6]. The application of those techniques on model organisms is already generating hundreds of gigabytes of data, most of them publicly available through portals like GenBank, GEO and others. However, there are some drawbacks that should be taken into account for the analysis of those datasets. Some techniques generate quite a lot of false positives, mainly due to technical artefacts. Standardized representations of the data are not the rule, making cross comparisons between experiments sometimes difficult. The quality of the data is not always assessed properly and the experimental conditions are often poorly annotated. Despite those problems, those datasets constitute a novel and exciting challenge for computational biologists to identify functional modules at a system level. In recent years, many research groups have proposed various algorithms to analyze 'omics' datasets [6]. Segal and co-workers introduced an algorithmic approach to reconstruct module network from a large compendium of microarrays. Their method is using a statistical model to infer modules of co-expressed genes and their corresponding regulators [8].

The LeMoNe algorithm

We have further extended the framework proposed by Segal and colleagues into an algorithm called LeMoNe, standing for *Learning Module Networks*. Our approach is using different probabilistic optimization techniques to produce a more reliable and robust result [4]. The algorithm is divided mainly in two steps. In the first one, we identify cluster of co-expressed genes, where gene expression for a given cluster is modeled by a normal mixture distribution:

$$p(X) = \sum_n C_n p(X \mid \mu_n, \sigma_n) \tag{1}$$

Where C_n denotes the group of conditions having a mean μ_n and a standard deviation σ_n. We are using a Gibbs sampling procedure to generate multiple, equally probable, clustering solutions (usually more than 10). Each single solution represents a different local optima. In order to have a robust result, we construct a centroid representation of all clustering solutions, called the tight clusters, corresponding to genes that are often associated together in all previous clustering solutions. The tight clusters are extracted with a graph spectral method [5].

For the second major step of the algorithm, we are assigning regulators to each tight cluster of co-expressed genes. A hierarchical tree is build by grouping sets of conditions (i.e. experiments) having similar mean and standard deviation. Then a pre-defined list of candidate regulators is build, most of the times by simply using the Gene Ontology (GO) annotation and selecting appropriate categories (for example Transcription Factor Activity or Signal Transduction Activity). Regulators from the list are assigned to each node of the hierarchical tree by logistic regression on the binary splits defined by the set of hierarchically linked condition clusters (Fig. 1). Let C_0 and C_1 be two disjoint sets of conditions. Given a regulator with expression value x in some condition, our model assumes there is a (hidden) binomially distributed random variable Y such that $Y = 0$ if the condition is assigned to C_0 and $Y = 1$ if it is assigned to C_1, with probability

$$p(Y = 1 \mid x) = \frac{1}{1 + e^{-\beta(x-z)}} \tag{2}$$

Given the partition of conditions and their hierarchical tree, we know at each tree node which conditions m belong to C_0 and which to C_1. Using Bayes' rule, we can determine the parameters β and z which maximize the posterior

probability of assigning regulator R. This posterior probability is then used as the score for R at this particular tree node and combined with the scores at other nodes to compute a global assignment score. The parameter z is interpreted as a *split value*, meaning if R is highly expressed ($x_m > z$) the condition is assigned to one side of the split and if R is lowly expressed ($x_m < z$) to the other side. The parameter β is determined by how well a regulator fits the separation of conditions: if $x_m > z$ for all $m \in C_1$ and $x_m < z$ for all $m \in C_0$ (or vice versa), we can take $\beta = +\infty$ and obtain a maximal posterior probability. If there is no split value which achieves a good separation of conditions, β will be close to 0 leading to low values of the posterior probability.

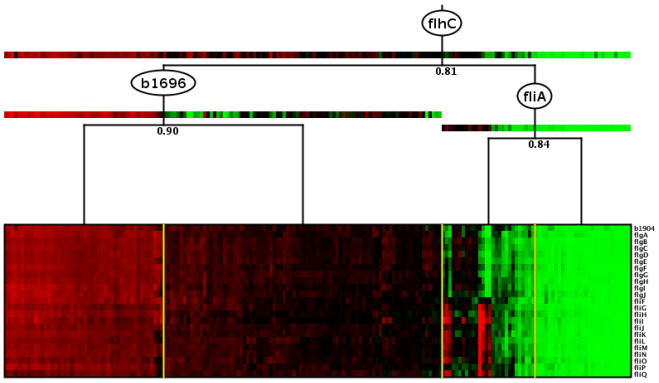

Figure 1: A tight cluster with its regulation tree and three assigned regulators. The bottom panel represents the co-expressed genes. Expression values are color coded from low (red) to high (green). The vertical yellow bars define groups of conditions having similar expression mean and standard deviation. The hierarchical tree on top is build by linking the different groups of conditions. Regulators are assigned to the different nodes of the tree, and their expression profile is shown with the same color scheme as for the module genes.

Here again, in order to build a robust solution, we are generating multiple hierarchical trees for each tight cluster and assigning multiple regulators for each node of each tree. An ensemble score is then calculated, summing the strength with which a regulator participates in each regulatory tree, allowing us to prioritize the list of regulators for each tight cluster.

When building the pre-defined list of candidate regulators, the classical choice is to select genes having *a priori* a regulatory activity like transcription factors, signal transducers or kinases. In that case, the expression values are exactly of the same type for the genes and the regulators, being measured on the same microarray platform. However, looking at equation (2), we can see that there is no need for the values x to be comparable in absolute terms to the expression values determining the co-expression clusters. This means that we can use expression values for different types of regulatory molecules, measured with a different platform, like microRNAs (miRNAs). There is also no need for the values x to be continuous, meaning that we can consider regulators having discrete values. As we are using a probabilistic model and the final regulator score is defined by a posterior probability, the scores of mRNA, miRNA and discrete regulators can all be integrated and compared on the same scale to determine the final module network. We usually determine a cutoff score for assigned regulators by taking the top 1% of the distribution of all assigned regulators. We define a module as a given tight cluster plus its associated list of high-scoring regulators. A given regulator can be assigned to one or multiple tight clusters, therefore the ensemble of modules is forming a module network.

Module network examples

We have applied our module network algorithm to the construction of module networks from cancer related expression datasets. We first used a dataset published by Lu and co-workers [7]. In this study, the authors compared the classifying power of mRNA and miRNA expression data to discriminate between different types of cancer tissues. They measured expression values for $\approx$ 12,000 mRNA and $\approx$ 120 miRNAs using a standard microarray platform and a proprietary solution respectively, on 89 different tissue samples. We applied our algorithm on this dataset, including transcription factors, signal transducers and miRNAs as candidate regulators. We obtained a set of 76 tight clusters for which a total of 294 high-scoring regulators were assigned [2]. Within this set, ten different miRNAs were selected as high-scoring regulators for seven modules. Fig. 2 shows an example of a module regulated by a miRNA. This module is likely involved in epithelial homeostasis and the top regulator is miR-200a. We could demonstrate with quite simple experiments that this miRNA is indeed a key regulator of the module genes, probably

acting on them indirectly through the action of a ZEB transcription factor.

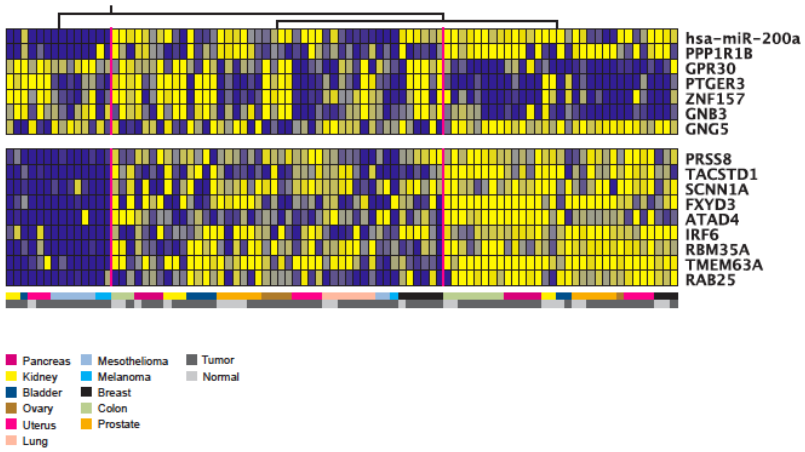

Figure 2: A module inferred from mRNA and miRNA expression data measured on normal/tumor tissues. There are nine module genes (bottom panel) and seven high scoring regulators (top panel). The tissue origin is indicated by a color code at the bottom of the figure, as well as their disease status. The regulators are prioritized by decreasing score.

More recently we applied our algorithm on a dataset of lymphoblastoid cell lines made from blood samples of 90 patient having prostate cancer. Both mRNA and miRNA expression levels were measured on those samples and at the same time the degree of aggressiveness of the tumors were characterized by a clinical parameter, the Gleason score [9]. We used the mRNA expression data to build tight clusters of co-expressed genes, and then used transcription factors, signal transducers, miRNAs and the Gleason score as candidate regulators. In this particular case, the Gleason score was having only two levels, being either 'low' or 'high'. We had expression data for more than 40,000 mRNAs and more than 700 miRNAs. The output was a module network consisting of 43 tight clusters composed of a total of 1,259 genes [1]. From the whole set of regulators assigned, we retained 496 unique candidate regulators. Among them, 30% are miRNAs, a significant increase compared to our previous analysis. Several of those miRNAs have been previously characterized as causal in human diseases, including some forms of cancer. Interestingly, the Gleason score was also retained as a high-scoring regulator, linked to three different modules enriched for cell proliferation, cell

growth and mitosis (Fig. 3). This parameter is of course not a regulator in itself, but this result might be interpreted as a consequence of the degree of aggressiveness of the prostate cancer, triggering subtle changes in expression in the patient blood system. In this study, due to the specificity of our algorithm, we could demonstrate that it is possible to simultaneously evaluate heterogeneous types of regulators in one framework, including discrete ones.

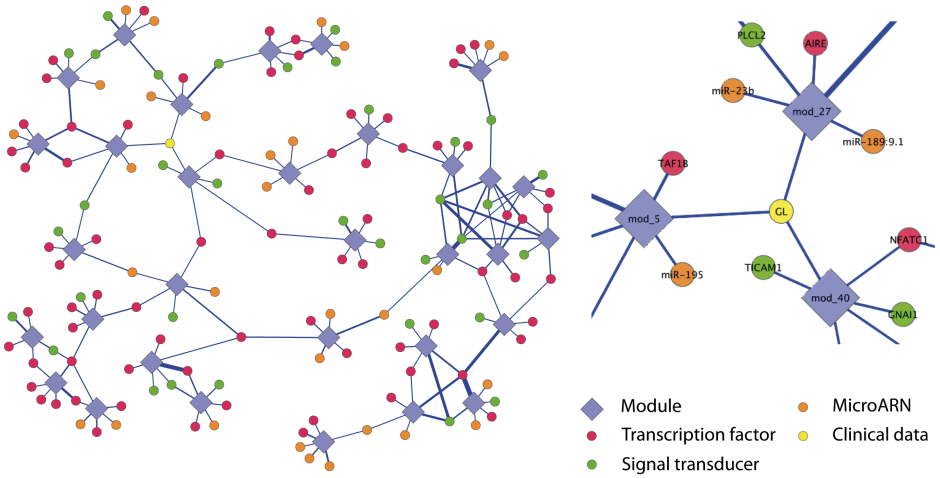

Figure 3: Simplified view of the module network inferred from lymphoblastoid cell lines expression data. On the left panel, modules are pictured as purple diamonds, while regulators are depicted as circles, with a color corresponding to different types of regulators. On the right panel, a zoom on the module network is showing a clinical parameter, The Gleason score, linked to three different modules, as well as other types of regulators.

Acknowledgments

Thanks to Tom Michoel, Anagha Joshi, Vanessa Vermeirssen and Yves Van de Peer for excellent collaborative work, fruitful discussions and stimulation. This work was funded by an IWT grant for the SBO project Bioframe and by an IUAP grant for the BioMaGNet project (ref. p6/25).

References

[1] E. Bonnet, T. Michoel, and Y. Van de Peer. Prediction of a gene regulatory network linked to prostate cancer from gene expression, microRNA and clinical data. *Bioinformatics*, in press, 2010.

[2] 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*, 5:e10162, 2010.

[3] L.H. Hartwell, J.J. Hopfield, S. Leibler, A.W. Murray, et al. From molecular to modular cell biology. *Nature*, 402(6761):47, 1999.

[4] 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*, 25(4):490, 2009.

[5] A. Joshi, Y. Van de Peer, and T. Michoel. Analysis of a Gibbs sampler method for model-based clustering of gene expression data. *Bioinformatics*, 24(2):176, 2008.

[6] A.R. Joyce and B.Ø. Palsson. The model organism as a system: integrating'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210, 2006.

[7] J. Lu, G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando, et al. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–838, 2005.

[8] 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. *Nature genetics*, 34(2):166–176, 2003.

[9] L. Wang, H. Tang, V. Thayanithy, S. Subramanian, A.L. Oberg, J.M. Cunningham, J.R. Cerhan, C.J. Steer, and S.N. Thibodeau. Gene Networks and microRNAs Implicated in Aggressive Prostate Cancer. *Cancer Research*, 69(24):9490, 2009.