

Genetic Network construction in CML gene expression profile data analysis ¹

Luigi Augugliaro², Angelo M. Mineo
Dipartimento di Scienze Statistiche e Matematiche S. Vianelli,
e-mail: {augugliaro, elio.mineo}@dssm.unipa.it

Giuseppe Cammarata
I.B.I.M., C.N.R. di Palermo, e-mail: cammaratacammarata@hotmail.com

Alessandra Santoro
Dipartimento di Ematologia 1, Azienda Ospedaliera V. Cervello,
e-mail: santoro.al@libero.it

Abstract: Aim of this paper is to define a new statistical framework to identify central modules in Gaussian Graphical Models (GGMs) estimated by gene expression data measured on a sample of patients with negative molecular response to imatinib. A central module is defined as a module of a GGM which contains genes that are defined differentially expressed.

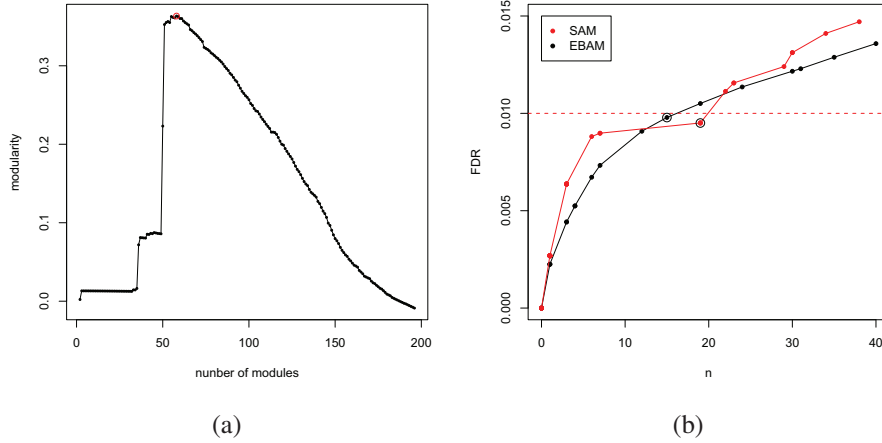
Keywords: Gaussian graphical models, modularity, differentially expressed genes.

1. Introduction

Imatinib, a specific ABL tyrosine kinase inhibitor, has been reported to have a significant clinic effect on chronic myeloid leukemia (CML) in chronic phase as well as in blast crisis. However, many patients in blast crisis who are being treated with imatinib relapse at a relatively early time, suggesting that leukemia cells tend to acquire resistance to imatinib. Such innate mechanism of resistance is poorly understood, but some evidences suggest that activation of alternative oncogenic pathway may confer BCR-ABL independent survival to CML cells. To understand the complexity of the regulation of gene expression, genetic networks have been used (Schlitt and Brazma, 2007, for more details). Genetic network are usually constructed using probabilistic models such as Gaussian Graphical Models (Edwards, 2000). Unfortunately their practical application is strongly limited by the amount of available experimental data. In a typical microarray dataset the number of genes exceed the sample number. To overcome this problem, Schäfer and Strimmer (2005a) proposed an empirical Bayesian framework to improve the inference on a GGM. However, this method does not consider an important property of a genetic network: the property of community structure (Segal *et al.*, 2003). Aim of this paper is to define a statistical framework to identify central modules in a Gaussian Graphical Model estimated by gene expression data measured on a sample of patients with negative molecular response to imatinib. We define *central* a module that includes differentially expressed genes.

¹Work partially supported by grants of the University of Palermo.

Figure 1: (a) modularity index w.r.t. number of modules; circled points correspond to the optimal number of modules. (b) estimated false discovery rate w.r.t. number of differentially expressed genes.



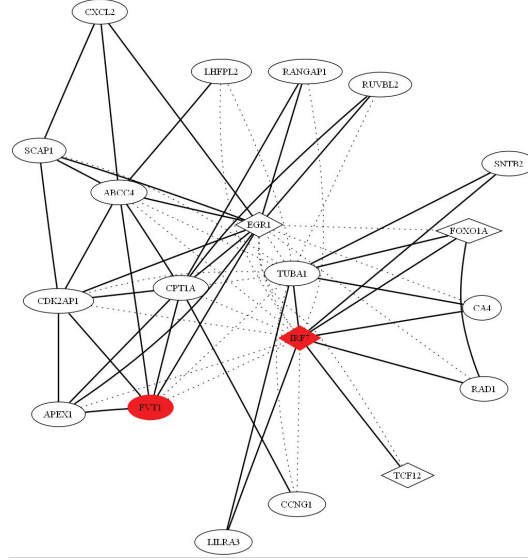
2. Finding modules in a Gaussian Graphical Model

In recent years, the analysis of genetic regulatory networks has received a major impetus from the huge amount of data made available by high-throughput technologies such as DNA microarrays. To better understand how genes are related to each other within the class of patients with negative molecular response to imatinib, we performed a genetic network analysis by means of GGMs. When we work with a microarray experiment the number of gene (variables) exceed the sample number, for this reason the sample covariance matrix is not positive definite and cannot be inverted. To overcome this problem, we used the method proposed by Schäfer and Strimmer (2005a), namely, in the first step the estimated partial correlation coefficients are obtained using the shrinkage approach (Schäfer and Strimmer, 2005b). In the second step a mixture distribution is used to address the statistical testing problem of non-zero partial correlation. Formally, we shall assume that the partial correlation coefficients $\hat{\rho}_{ij}$ across all edges in the network follow a mixture distribution $f(\hat{\rho}_{ij}) = \pi_0 f_0(\hat{\rho}_{ij}, \kappa) + (1 - \pi_0) f_1(\hat{\rho}_{ij})$ where π_0 is the prior for the null distribution

$$f_0(\hat{\rho}_{ij}, \kappa) = (1 - \hat{\rho}_{ij}^2)^{(\kappa-3)/2} \frac{\Gamma(\kappa/2)}{\pi^{1/2} \Gamma((\kappa-1)/2)} \quad (1)$$

where κ are the degrees of freedom. The (1) is the distribution of the sample normal correlation coefficient under $H_0 : \rho_{ij} = 0$. $f_1(\rho)$ is computed by means of a non-parametric method. Using a sample of 9 patients with negative molecular response to imatinib and a value 0.9 as threshold for the empirical posterior probability of a edge being present, a Gaussian graphical model consisting of 197 genes and 895 edges was obtained. As observed by Barabasi and Oltvair (2004), cellular functions are likely to be carried out in a highly modular manner, namely, division of network nodes into groups where the network connections are dense within the group, but are sparser between the groups. In order

Figure 2: Graph of the fourth subnetwork. Transcript factors are labeled as diamonds. Differentially expressed genes are in red color. Positive estimated partial correlations are labeled by continuous lines; dotted lines identify negative estimated partial correlations.



to identify central modules related with the negative molecular response to imatinib, we applied the method proposed by Newman and Girvan (2004) to identify a set of modules in the estimated GGM. This method is based on the idea that edges connecting separate modules have high edge betweenness, since all the shortest paths from one module to another have to cross them. Then, modules were identified gradually removing the edge with the highest edge betweenness score. Panel (a) in figure 1 shows the modularity index (Newman and Girvan, 2004) as function of the number of subnetworks identified. The peak corresponds to 58 subnetworks; however, as we shall see in the following section, only four modules are related with a set of genes called differentially expressed.

3. Finding central modules and conclusions

In order to identify modules related with the negative molecular response to Imatinib, a sample of 8 patients with positive molecular response was used to identify a set of differentially expressed genes. A list of differentially expressed genes was defined combining the results obtained using the significant analysis of microarrays (Tusher *et al.*, 2001) and the empirical Bayes analysis of microarrays (Efron *et al.*, 2001). In order to make comparable the results given by the two methods, we chose 0.01 as optimal false discovery rate. Figure 1 (b) shows the estimated false discovery rate as function of the number of differentially expressed genes. The highest n corresponding to an estimated false discovery rate lower than 0.01 was chosen as number of differentially expressed genes. Using this criterion, SAM led to the identification of 19 differentially expressed genes while EBAM led to the identification of 16 genes, with thirteen of these that were found with the SAM. As we have seen in section 2., we can identify 58 subnetworks into the estimated Gaus-

sian graphical model. However, only four of them contain the differential expressed genes identified using SAM and EBAM methods. In our approach, these submodules are defined central. Annotated gene list is not reported for the sake of brevity. Figure 2 shows one of the central modules identify with the proposed method. Validating our results by PReMod (a data base of genome-wide mammalian cis-regulatory module predictions, Ferretti *et al.* (2007)), several interactions have been confirmed. Some of these interactions, like Egr-1– Foxo1A (see figure 2), are also confirmed in a recent functional study indicating that Foxo1A behaves as a negative regulator of Egr-1 expression (Cabodi *et al.*, 2009). Moreover, the transcription factor Egr-1, central in the module reported in figure 2, has been described in a recent paper as a potent tumor suppressor gene whose deregulation is involved in hematological malignancies (Gibbs *et al.*, 2008).

References

- Barabasi A.L. and Oltvair Z.N. (2004) Network biology: understanding the cell's functional organization, *Nat. Rev. Genet.*, 5, 2, 101–113.
- Cabodi S., Morello V., Masi A., Cicchi R., Broggio C., DiStefano P., Brunelli E., Silengo L., Pavone F., Arcangeli A., Turco E., Tarone G., Moro L. and Defilippi P. (2009) Convergence of integrins and EGF receptor signaling via PI3K/Akt/FoxO pathway in early gene Egr-1 expression., *J. Cell Physiol.*, 218, 2, 294–303.
- Edwards D. (2000) *Introduction to Graphical Modelling*, Springer Verlag, New York.
- Efron B., Tibshirani R., Storey J. and Tusher V. (2001) Empirical Bayes Analysis of a Microarray Experiment, *J. Amer. Statist. Assoc.*, 96, 456, 1151–1160.
- Ferretti V., Poitras C., Bergeron D., Coulombe B., Robert F. and Blanchette M. (2007) PReMod: a database of genome-wide mammalian cis-regulatory module predictions., *Nucleic Acids Res.*, 35, D122–D126.
- Freeman L. (1978) Centrality in social networks: Conceptual clarification, *Social Networks*, 1, 215–239.
- Gibbs J., Liebermann D. and Hoffman B. (2008) Egr-1 abrogates the E2F-1 block in terminal myeloid differentiation and suppresses leukemia., *Oncogene.*, 27, 1, 98–106.
- Horvath S. and Dong J. (2008) Geometric interpretation of gene coexpression network analysis, *PLoS Comput. Biol.*, 4, 8, e1000117.
- Newman M. and Girvan M. (2004) Finding and evaluating community structure in networks, *PHYSICAL REVIEW*, E 69, 026113.
- Schäfer J. and Strimmer K. (2005a) An empirical bayes approach to inferring large-scale gene association networks, *Bioinformatics*, 21, 6, 754–764.
- Schäfer J. and Strimmer K. (2005b) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statist. Appl. Genet. Mol. Biol.*, 4, 32.
- Schlitt T. and Brazma A. (2007) Current approaches to gene regulatory network modelling, *Bioinformatics*, 8(Suppl 6).
- Segal E., Shapira M., Regev A., Pe'er D., Botstein D., Koller D. and Friedman N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data., *Nat. Genet.*, 2, 34, 166–76.
- Tusher V.G., Tibshirani R. and Chu G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences*, 98, 9, 5116–5121.