Title : Inferring Networks from Multiple Samples with Consensus LASSO

Authors: Nathalie Villa-Vialaneix (1,2), Matthieu Vignes (2), Nathalie Viguerie (3,4) and Magali SanCristobal (5)
(1) SAMM, EA 4543, Université Paris 1 - France
(2) INRA, UR 875 MIA-T, Castanet Tolosan - France
(3) Inserm, Obesity Research Laboratory, Institute of Metabolic and Cardiovascular Diseases, Toulouse - France
(4) University of Toulouse, Paul Sabatier University, Toulouse, France
(5) INRA, Laboratoire de Génétique Cellulaire, Castanet Tolosan - France

Abstract:
The recent development of high-throughput techniques makes available huge datasets where thousand genes are simultaneously measured. However, the number of observations is, comparatively, very small, and those are often measured in a variety of experimental conditions. One of the big challenge of modern systems biology is to understand the influence of controlled experimental conditions on the functioning of living organisms. This question is usually addressed by searching for the difference between gene expressions pertaining to the condition (hence for "differentially expressed genes"). But the differences in the way the genes interact with each others is also a question of interest: finding which regulation pathways are modified by a given experimental condition gives an interesting insight on the influence of the condition on the living system in its whole.

One of the most popular approach to understand the complex relationships existing between the expression of a large set of genes is to infer a co-expression network from a transcriptomic dataset. In such a model, the nodes of the network represent the genes and an edge between two nodes models a strong co-expression between the two genes. A number of different methods have been developed to infer such networks: using correlations (``relevance network'', Butte & Kohane, 2000), Bayesian networks (Pearl, 1998 or Pearl & Russel, 2002), Graphical Gaussian Model (Edwards, 1995)... When the observations have been collected in different conditions, a naive approach would be to infer a network for each experimental condition and to compare them. However, this method will not be able to stress out specifically the differences and the commonalities of regulation phenomenons: since the number of observations is small, inferring the networks independently, forgetting that a common functioning should exist whatever the condition, will lead to emphasize irrelevant differences.

In this proposition, we will present a novel method for inferring co-expression networks from samples obtained in different experimental conditions. This approach is based on a double penalization: a first penalty aims at inferring a sparse solution; then, the second penalty is used to make the networks obtained in different conditions consistent with a consensual network. The "consensual network" is introduced to represent the dependency structure between genes, the common functioning of the living organism under study, whatever the condition. The estimation is made more robust by using a bootstrap approach. Our proposal is tested and compared to existing alternatives, on simulated datasets, investigating the influence of the number of different edges between conditions and of the sample size. It is also applied on a real-world dataset where the transcriptom has been measured for different breeds of a given mammalian species.