A computationally efficient method for gene network identification

ID - 132

Lauria Mario[1], Di Bernardo Diego[1]
[1]TIGEM, Napoli

Motivation

An increasing amount of data from gene expression array experiments is becoming available due to increasing interest toward gene regulatory networks and the decreasing costs of the DNA chip technology. A number of approaches have been proposed to 'reverse engineer' such data. All known inference methods are based on statistical techniques and are characterized by a trade-off between selectivity (ability to discriminate between true inferred gene interactions from false ones) and sensitivity (percentage of true interactions recovered). Currently, no methods exist that can recover more than a small fraction of the regulatory interactions with high confidence; high confidence is desirable because the in vitro verification of putative interactions is expensive. A new promising approach is based on combining a statistical inference method and an iterative algorithm in which a solution is obtained through successive approximations. In order to make this approach feasible, a computationally fast inference method needs to be developed.

Methods

In this work we focus on a two-stage approach in which we first select a subset of genes whose expression profiles appear to be correlated, and then apply a method to estimate the coupling coefficients for the selected genes based on their profiles. For the second step we chose a fast and theoretically elegant statistical inference method based on entropy maximization proposed by Lezon et al.~\cite{Lezon2006}. The first step of gene selection is the most critical and less theoretically characterized of the two; we focused most of our efforts on this step. We empirically tested several different approaches and compared their performance using a synthetic data set of 100 genes and 100 experiments. The synthetic data set was the same used in the review paper by Bansal et al.~\cite{Bansal2007}, therefore we were able to directly compare our results with those obtained with mainstream network inference tools.

The first selection gene rule we tested was a straightforward implementation of the selection rule presented in the Lezon et al. paper~\cite{Lezon2006}, based on the selection of genes whose expression profiles have values well above the average value. We then improved on the original approach by using selection criteria based on the value of correlation coefficient between pairs of gene expressions; more specifically, we tried using the values of the coefficients themselves, and then the associated p-values. The rule based on p-values turned out to be the one giving the best results.

We tried other approaches to selection of genes, inspired to methods found in the literature on gene regulatory network inference. Briefly, for one of these rules we computed the entropy of each gene expression profile; the computation was performed using an histogram of the expression values. In another rule we tried to apply the CLR algorithm to the pre-filtering of genes, instead of post filtering of coupling coefficients as originally proposed~\cite{Faith2007}. In the last two rules we used the value of Mutual Information (MI) as the selection criteria; MI is computed using the histogram method and the kernel density estimation method respectively, as described in \cite{Steuer2002}.

Results

In this paper we propose a fast method for inferring genetic interaction networks from gene expression data. By decomposing the problem of inferring a network into two sub-steps, gene selection and coupling coefficient determination, we were able to use simpler computational methods than those currently proposed in the literature.

Based on an evaluation with synthetic data, our two-stage approach is between 2 and 3 orders of magnitude faster than the best published method and only a factor of 2.5 to 3 worse in terms of selectivity for a comparable level of sensitivity.

Email: lauria@tigem.it