Biological Network Inference Using Redundancy Analysis

Patrick E. Meyer, Kevin Kontos and Gianluca Bontempi

{pmeyer,kkontos,gbonte}@ulb.ac.be
ULB Machine Learning Group
Computer Science Department
Université Libre de Bruxelles
1050 Brussels – Belgium
http://www.ulb.ac.be/di/mlg/

Abstract. The data flood phenomenon that biology is experiencing has propelled scientists toward the view that biological systems are fundamentally composed of two types of information: genes, encoding the molecular machines that execute the functions of life, and transcriptional regulatory networks (TRNs), specifying how and when genes are expressed.

Two of the most important challenges in computational biology are the extent to which it is possible to model these transcriptional interactions by large networks of interacting elements and the way that these interactions can be effectively learned from measured expression data.

The reverse engineering of TRNs from expression data alone is far from trivial because of the combinatorial nature of the problem and the poor information content of the data. However, progress has been made over the last few years and effective methods have been developed. Well-known state-of-the-art methods are Boolean Network Models, Bayesian network models and Association Network Models.

We focus on information theoretic approaches which typically rely on the estimation of mutual information from expression data in order to measure the statistical dependence between genes.

Information-theoretic network inference methods have recently held the attention of the bioinformatics community also for very large networks. We introduces an original information-theoretic method, MRNet, inspired by a known feature selection algorithm, the maximum relevance-minimum redundancy (MRMR) algorithm. This algorithm has been used with success in supervised classification problems to select a set of non redundant genes which are explicative of the targeted phenotype. The MRMR selection strategy consists in selecting a set of variables that both have high mutual information with the target variable (maximum relevance) and are mutually maximally dissimilar (minimum redundancy). The advantage of this approach is that the trade-off between relevance and redundancy is properly taken into account.

The proposed MRnet strategy consists in (i) formulating the network inference problem as a series of input/output supervised gene selection tasks where each gene at the time plays the role of the target output, (ii)

adopting the MRMR principle to perform the gene selection for each of supervised tasks.
The rationale is that the MRMR selection relies on a square matrix of bivariate mutual information that can be computed once for all, making then the network inference computationally affordable also for a large number of genes.
We benchmark MRNet against two state-of-the-art information-theoretic network inference methods, namely relevance networks and ARACNe. The comparison relies on six different synthetic microarray datasets obtained with two different generators.