Supervised inference of biological networks with trees: Application to genetic interactions in yeast

Marie Schrynemackers
2.2.2015

Networks or graphs provide a natural representation of molecular biology knowledge, in particular to model relationships between biological entities such as genes, proteins, drugs, or diseases. Because of the effort, the cost, or the lack of the experiments necessary to the elucidation of these networks, computational approaches for network inference have been frequently investigated in the literature. In this thesis, we focus on supervised network inference methods. These methods exploit supervised machine learning algorithms to train a model for identifying new interacting pairs of nodes from a training sample of known interacting and possibly non-interacting pairs and additional measurement data about the network nodes.

Our contributions in this area are divided into three parts. First, the thesis examines the problem of the assessment of supervised network inference methods. Indeed, their reliable validation (in silico) poses a number of new challenges with respect to standard classification problems, related to the fact that pairs of objects are to be classified and to the specificities of biological networks. We perform a critical review and assessment of protocols and measures proposed in the literature. Through theoretical considerations and in silico experiments, we analyze in depth how important factors influence the outcome of performance estimation. These factors include the amount of information available for the interacting entities, the sparsity and topology of biological networks, and the lack of experimentally verified non-interacting pairs. From this analysis, we derived specific guidelines so as to how best exploit and evaluate machine learning techniques for network inference.

Second, we systematically investigate, theoretically and empirically, the exploitation of tree-based methods for network inference. We consider these methods in the context of the two main generic classification-based approaches for network inference: the local approach, which trains a separate model for each network node, and the global approach, which trains a single model over pairs of nodes. We present and formalize these two approaches, extending the former for the prediction of interactions between two unseen network nodes, and discuss their specializations to tree-based methods, highlighting their interpretability and drawing links with clustering techniques. Extensive experiments are carried out with these methods on various biological networks that clearly highlight that these methods are competitive with existing methods. The interpretability of the resulting method family is illustrated on a drug-protein interaction network.

In the last part of the thesis, we built on the experience gained in the two
previous parts to try to predict at best the genetic interaction network in yeast 
S.cerevisiae. For that purpose, we collected a large dataset, assembling 4 mil- 
lions gene pairs that were experimentally tested in the context of 11 different 
studies and 23 sets of measurements to use as gene input features for the infer- 
ence. Through several cross-validation experiments on the resulting dataset, we 
showed that predicting genetic interactions is indeed possible to some useful ex- 
tent and that actually in some settings, the accuracy of computational methods 
is not very far from that of experimental techniques