Reconstructing regulatory networks by integrating post-genomic data with biological prior knowledge

Dirk Husmeier

An important and challenging problem in systems biology is the inference of gene regulatory networks from high-throughput microarray data. Various machine learning and statistical methods have been applied to this end. An intrinsic difficulty with these approaches is that complex interactions involving many genes usually have to be inferred from sparse and noisy data. This leads to a poor reconstruction accuracy and suggests that the inclusion of complementary information and biological prior knowledge, related for instance to transcription factor binding locations in promoter regions or partially known signalling pathways from the literature, is indispensable.

We have developed a Bayesian approach to systematically integrate postgenomic data with independent sources of prior information. A hyperparameter that is automatically inferred when training the model determines the weight of the prior knowledge and trades its influence against the data. The approach is based on the methodology of Bayesian networks, reviewed e.g. in Husmeier et al. (2005). Details about the proposed scheme can be found in Werhli and Husmeier (2007).

We have evaluated the method on the Raf–Mek–Erk signal transduction pathway. Raf is a critical signalling protein involved in regulating cellular proliferation in human immune system cells. The deregulation of the Raf pathway can lead to carcinogenesis, and this pathway has therefore been extensively studied in the literature (e.g. Sachs et al., 2005). Figure 1 shows the representation of the currently accepted regulatory network of 11 phosphorylated phospholipids (Pip2 and Pip3) and proteins (all other nodes), taken from Sachs et al. (2005). Several of the connections are direct enzyme–substrate relationships, like Pka to Raf, Raf to Mek and Mek to Erk. The edge between Pip3 and Plcg has a relationship of recruitment leading to phosphorylation. Some of the interactions shown are indirect and may involve specific phosphorylation sites of the signalling molecules. Our objective was to test if this regulatory network could be reverse-engineered from postgenomic data of the type available at a plant research institute like SCRI.

Sachs et al. (2005) have applied intracellular multicolour flow cytometry experiments. Each independent sample in the data consists of quantitative amounts of each of the 11 phosphorylated molecules, simultaneously measured from a population of over 5000 cells. This provides a rich data set from which Sachs et al. (2005) successfully reconstructed the gold-standard network of Figure 1 with Bayesian networks.

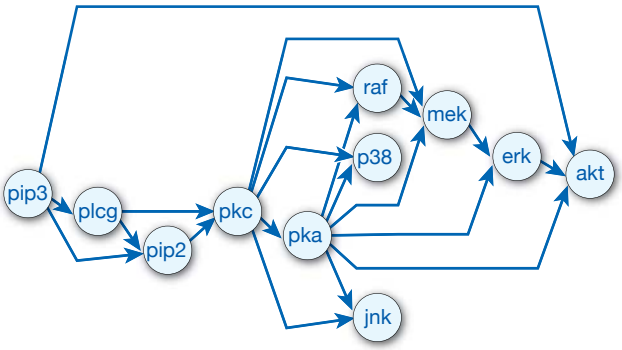

Figure 1 Raf signaling pathway. The graph shows the currently accepted Raf signaling network, taken from Sachs *et al.* (2005). Nodes represent phosphorylated proteins and phospholipids, edges represent interactions, and arrows indicate the direction of signal transduction.

Unfortunately, the number of experimental conditions which can be explored with a microarray experiment at a plant institute is usually much smaller than at a medical research institute. To reflect this practical constraint we down-sampled the data from Sachs et al. (2005) to five non-overlapping sets of 100 experimental conditions. We then evaluated the accuracy achieved with the network reconstruction method, and

investigated to what extent the systematic integration of biological prior knowledge can improve the results. The source of prior knowledge we chose was the KEGG pathways database (Kanehisa and Goto, 2000), which represents our current knowledge of the molecular interactions and reaction networks related to metabolism, a variety of cellular processes, and different diseases. We extracted all pathways from KEGG that contained at least one pair of the 11 proteins/phospholipids included in the Raf–Mek–Erk pathway. We formulated our prior knowledge as a matrix containing the relative proportions of pairwise molecular interactions among all the pathways extracted.

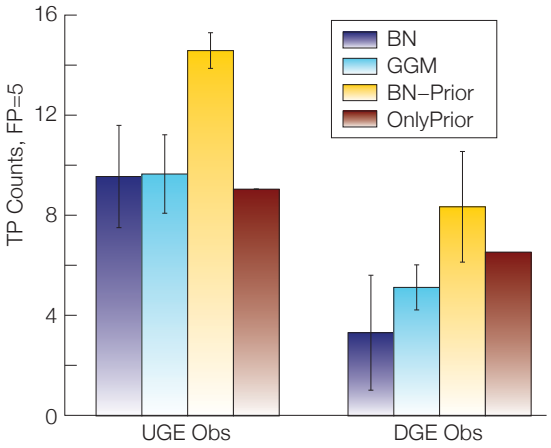

Figure 2 Reconstruction of the Raf-Mek-Erk signalling pathway. The figure shows the means across five data sets of the number of true interactions inferred for a fixed number of 5 spurious edges. Each evaluation was carried out twice: with and without taking the edge direction into consideration (DGE and UGE respectively). The blue bars represent the results obtained from the data only, Bayesian networks (BN) and graphical Gaussian models (GGM); brown bars from the prior knowledge alone; yellow bars from the Bayesian approach, which systematically integrates the data with the prior knowledge from KEGG. The error bars show the respective standard deviations computed from the five replications.

The results are shown in Figure 2. The proposed Bayesian inference scheme clearly outperforms the methods that do not include the prior knowledge from the KEGG database. It also clearly outperforms the prediction that is solely based on the KEGG pathways alone without taking account of the cytometry data. The histograms indicate the accuracy one can typically expect to achieve with the amount of data included in our study, rising from a recovery of about 50% to 75% of the interactions in the pathway at a cost of incurring about 5% of false interactions.

This method is sufficiently generic that it can be applied directly to plant data in the study of plant regulatory networks.

References

Husmeier, D., Dybowski, R. and Roberts, S. 2005, Probabilistic Modeling in Bioinformatics and Medical Informatics. Advanced Information and Knowledge Processing. Springer, New York.

Kanehisa, M. and Goto, S. 2000, KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research, **28**, 27–30.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. and Nolan, G.P. 2005, Protein-signaling networks derived from multiparameter single-cell data. Science, **308**, 523–529.

Werhli, A.V. and Husmeier, D. 2007, Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. Statistical Applications in Genetics and Molecular Biology, **6**,15.