Boosting Statistical Network Inference by Incorporating Prior Knowledge from Multiple Sources

Paurush Praveen
University of Bonn,
Bonn-Aachen International Center for IT,
Dahlmannstr. 2, 53113 Bonn
praveen@bit.uni-bonn.de

Holger Frölich
University of Bonn,
Bonn-Aachen International Center for IT,
Dahlmannstr. 2, 53113 Bonn
frohlich@bit.uni-bonn.de

1 Motivation

Statistical learning methods, such as Bayesian Networks, have gained a high popularity to infer cellular networks from high throughput experiments. However, the inherent noise in experimental data together with the typical low sample size limits their performance with high false positives and false negatives. Incorporating prior knowledge into the learning process has thus been identified as a way to address this problem, and principle a mechanism for doing so has been devised e.g. by Mukherjee and Speed, 2008 [1] and Fröhlich *et.al*, 2007 [2]. However, so far little attention has been paid to the fact that prior knowledge is typically distributed among multiple, heterogeneous knowledge sources (e.g. GO, KEGG, HPRD, etc.).

2 Method

Here we propose two methods for constructing an quantified informative network prior from multiple knowledge sources, namely- Gene Ontology terms, Protein domain data, existing protein interaction databases *etc*. Our basic assumption is that from each of these knowledge sources for each pair of putatively interacting molecules we can derive a confidence score, which is scaled between 0 and 1, where 0 means that an edge is extremely unlikely and 1 means it is highly likely. As an example consider similarity measures for gene products based on their GO annotation [3, 4].
The first model we propose is a Latent Factor Model (LFM) using Bayesian inference. It is based on the assumption that sources with correlated information are all derived from the true, but unknown biological network, which we represent by a latent variable. We then apply Markov Chain Monte Carlo to infer the latent variable. The model has certain similarities to the model by Weile *et.al.* [5] to integrate networks derived from several microarray data sets.
Our second model is the Noisy-OR model (NOM). It is a non-deterministic disjunctive relation between effects and it's causes (Pearl, 1988) [6]. The Noisy-OR model assumes that the relation among cause(s) and effect is non-deterministic, allowing the presence of the effect in absence of any of the modeled cause. Both models are compared to a naïve method (Independent Prior; IP), which assumes independence of knowledge sources, proposed by Gao *et.al.*, 2011 [7].

3 Results

Extensive simulation studies on artificially created networks as well as full KEGG pathways reveal a significant improvement of both suggested methods compared to the naïve model. The performance

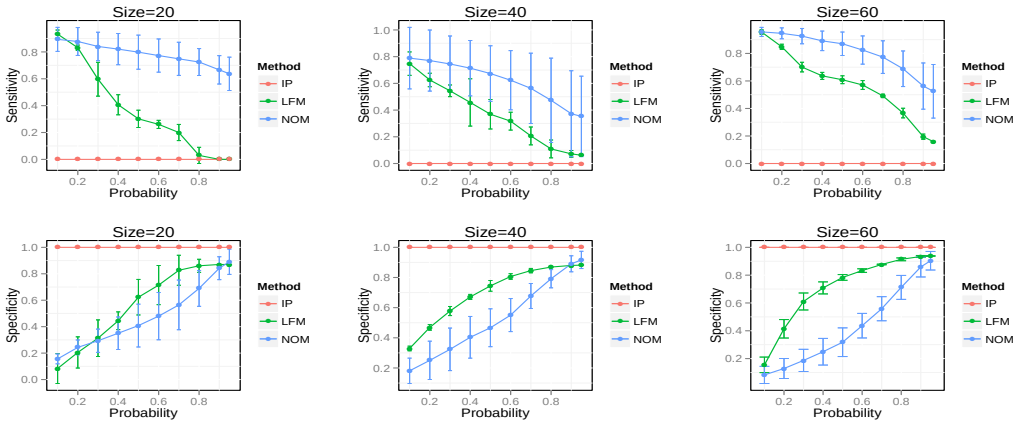

Figure 1: Performance of Latent Factor Model (LFM) and Noisy-OR Model (NOM) against the naïve Independent Prior (IP) method for sub-graphs of KEGG pathways of different number of nodes, generated by random walk model. Plotted are the average sensitivities and specificities for network reconstruction using the estimated prior only without any additional data (10 trials). The x-axis represents different edge-probability cutoffs.

of the latent factor model increases with larger network sizes, whereas for smaller networks the Noisy-OR model appears superior [Figure 1]. Furthermore, we show in our simulation studies that informative priors significantly enhance the reconstruction accuracy of Bayesian Networks. Finally, applying our method on two examples, one from breast cancer and one from murine stem cell development highlights the utility of our approach.

4 Discussion

We developed and proposed two methods to integrate different sources of biological information as prior knowledge for network inference via statistical learning. Our approach is based on the assumption of relatedness of biological data from several heterogeneous knowledge sources. Both, the Latent Factor Model and the Noisy-OR Model, significantly outperformed a recently proposed naïve Bayes approach working with independent priors. Specifically for larger networks the Latent Factor Model seems to offer an advantage compared to the Noisy-OR. Both methods were able to enhance the reconstruction accuracy of Bayesian Networks.

Acknowledgments

The work was partially supported by the state of NRW, Germany via B-IT Research School.

References

[1] Mukherjee, S. & Speed, T. P. (2008) Network inference using informative priors Proceedings of the National Academy of Sciences. *Proceedings of the National Academy of Sciences*, 105, 14313-14318

[2] Fröhlich, H.; Fellman, M.; Sultman, H.; Poustka, A. & Beissbarth, T. (2007) Large scale statistical inference of signaling pathways from RNAi and microarray data *BMC Bioinformatics*, 2007, 8

[3] Andreas Schlicker, Francisco S Domingues, Jörg Rahnenfhrer, Thomas Lengauer (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC bioinformatics* 7 (1) p. 302

[4] Fröhlich, H.; Speer, N.; Poustka, A. & Beissbarth, T. GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products *BMC Bioinformatics*, 2007, 8, 166.

[5] Weile, J.; James, K.; Hallinan, J.; Cockell, S. J.; Lord, P.; Wipat, A. & Wilkinson, D. (2012) Bayesian integration of networks without gold standards *Bioinformatics*

[6] Pearl, J. (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. *Morgan Kaufmann Publishers Inc, San Francisco* 1 edition.

[7] Gao, S. & Wang, X (2011) Quantitative utilization of prior biological knowledge in the Bayesian network modeling of gene expression data *BMC Bioinformatics*, 2011, 12, 359