

Evaluation of a Gene Network Extraction Method on Synthetic Data

Edwin D. de Jong and Arno Siebes

Large Distributed Databases Group, ICS, Utrecht University.

Introduction In recent years, a number of methods for the extraction of gene interaction networks from data have been introduced, e.g. [2, 1]. A question arising from this development is how the effectiveness of the different available network extraction methods may be compared. This effectiveness can be tested to some degree by comparing the interactions identified by a method with databases of known biological interactions, such as BIND, HPRD, KEGG and Reactome. However, to measure the accuracy of a network extraction method as accurately as possible, the true underlying network must be known. Since current biological knowledge of genetic interactions is incomplete, the use of a *generator* of gene expression data is promising. Here, we use a recent generator named SynTReN [3] to evaluate a network extraction method.

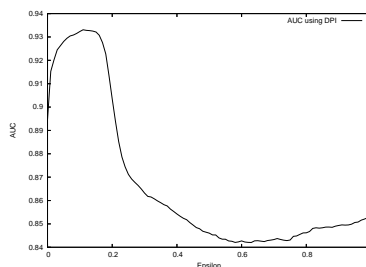
SynTReN SynTReN generates synthetic transcriptional regulatory networks and produces synthetic gene expression data that corresponds to the network. The structure of the networks is based on known biological gene networks. To generate expression data, Michaelis-Menten and Hill interaction kinetics are employed. The networks produced by the generator have been shown to approximate genuine biological networks more closely than do those of random graph models [3].

Methods The recent ARACNE algorithm was shown to successfully extract genetic regulatory networks from human B cell expression data [1]. A central element of ARACNE is the Data Processing Inequality (DPI) from signal theory, which is employed to remove indirect interactions.

We employ SynTReN data to evaluate the effectiveness of DPI in identifying indirect interactions. Given the expression data, the initial set of interactions is identified based on the mutual information between each pair of genes. DPI operates by considering each set of three genes, and removing the interaction that has the lowest mutual information, provided it is lower by a multiplicative factor of at least ϵ ; see [1] for details. By varying epsilon, the aggressiveness of DPI can be regulated; $\epsilon = 0$ means the lowest of the three interactions is always removed, while $\epsilon \geq 1$ removes no interactions. The quality of networks is evaluated by measuring the area under the ROC curve (AUC).

Results and Conclusions It is found that for a range of values of the ϵ parameter, the AUC can be substantially improved. When no DPI is used, the AUC is equal to the values at the right end of the graph, i.e. around 0.85. Using DPI, an area under the curve of over 0.93 can be achieved, representing a substantial improvement. While even a zero tolerance factor leads to improvement (AUC = 0.89), the best performance is achieved for $0.05 < \epsilon < 0.15$. For high values of ϵ , e.g. around 0.6, a negative effect on performance can be observed. Our conclusions are twofold: the use of a network generator is effective in evaluating network extraction methods, and the DPI procedure can effectively improve reconstruction accuracy.

Acknowledgements: The authors would like to thank the authors of the SynTReN generator, especially Piet van Remortel, Koenraad Van Leemput, and Tim Van den Bulcke, for kindly providing the source code of the generator and valuable assistance with its use.



References

1. K Basso, AA Margolin, G Stolovitzky, U Klein, R Dalla-Favera, and A Califano. Reverse engineering of regulatory networks in human b cells. *Nature Genetics*, 37(4):382–390, 2005.
2. AJ Butte and IS Kohane. Unsupervised knowledge discovery in medical databases using relevance networks. In *Proceedings of the AMIA Symposium*, pages 711–715, 1999.
3. Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchall. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(43), 2006.