Poster J-53

Using Synthetic Gene Expression Data to Assess Characteristics of Gene Network Inference Algorithms

Authors:
Tim Van den Bulcke (*ESAT-SCD, K.U.Leuven*)
Koenraad Van Leemput (*ISLab, Universiteit Antwerpen*)
Thomas Dhollander (*ESAT-SCD, K.U.Leuven*)
Bart De Moor (*ESAT-SCD, K.U.Leuven*)
Piet van Remortel (*ISLab, Universiteit Antwerpen*)
Kathleen Marchal (*CMPG, K.U.Leuven*)

Short Abstract: We report the application of synthetic datasets to three well-known gene network inference algorithms, providing insight in the impact of different aspects of the expression data on the quality of the inferred network. Synthetic data reveals operational characteristics of inference algorithms unlikely to be discovered using biological microarray data only.

Long Abstract:
The development of algorithms to infer the structure of gene regulatory networks based on expression data is an important subject in Systems Biology research. This paper reports the application of SynTReN, an existing network generator which samples topologies from existing biological networks and uses Michaelis-Menten and Hill enzyme kinetics to simulate gene interactions. This research describes the use of synthetic data to provide insight into how the quality of the inferred network is affected by different aspects of the expression data. Three well-known inference algorithms were subjected to these analyses: Aracne, Genomica and SAMBA. Each of the algorithms exhibited a different behavior to varying parameters of the synthetic data. We analyzed the impact of the network size, network topology, type and degree of noise, availability of expression data and interaction types between genes. The results show the added value of synthetic data in revealing operational characteristics of inference algorithms which are unlikely to be discovered by means of biological micro-array data only.

The validation of gene network inference algorithms requires benchmark data sets for which the underlying network is known. Since experimental data sets of the appropriate size and design are usually not available, there is a need to generate well-characterized synthetic data sets that allow thorough testing of learning algorithms in a fast and reproducible manner. The generated data should be both biologically plausible and computationally easy to generate for different parameters.

We illustrate the application of SynTReN, which is an existing generator of models of transcription regulatory networks and associated expression data. Instead of using random graph models, SynTReN uses topologies that are generated based on previously described source networks, allowing better approximation of the statistical properties of biological networks. No attempt was made to explicitly fine-tune the parameters of the inference algorithms, since the main goal of this work was to observe the effect of different properties

of the data on the inference procedure in a qualitative sense, and not a quantitative performance comparison of the algorithms themselves.

The following setup was used throughout the experiments: in a first step, a synthetic gene interaction network is generated based on a chosen network topology and interaction types. Next, expression data is generated that corresponds to the gene interactions dictated by the network. This involves setting various levels of noise and structuring of the resulting dataset in experiments, and samples per experiment. For each experiment, a number of simulated micro-array datasets are produced. The resulting datasets are used as input to the different network inference algorithms, which produce a candidate network of genes or gene modules, depending on the algorithm. In a final step, both the original network topology and the inferred candidate are compared by calculating several performance metrics from derived adjacency matrices of the algorithm outcome.

Three different types of algorithms were applied to the different synthetic datasets. Genomica reconstructs a cell's regulatory network from expression data as a network of interacting modules. The algorithm takes gene expression data as input and searches for a partitioning of genes into modules and an associated regulation program for each module. The output of the algorithm is a list of gene modules. Each module consists of a set of co-regulated genes, their regulators, and the conditions under which regulation takes place.

SAMBA is a bi-clustering algorithm that groups genes by means of a clustering of similar expression patterns in the input data over a subset of input conditions. The algorithm is based on a graph theoretic approach and statistical modeling of the data. The subsets of genes that jointly respond to specific conditions can be interpreted to form a module network.

Aracne is the third inference algorithm used in this study. It differs from the previous two algorithms in the sense that it explicitly infers a gene interaction network instead of a module network. It reconstructs gene regulatory networks from microarray expression profiles on the basis of mutual information between the genes.

Experiments show that the topology of the network can have a strong impact on the performance of an inference algorithm, which should be taken into account when evaluating inference algorithms using synthetic datasets. It is encouraging to note that for two of the algorithms tested in this experiment, Genomica and Aracne, the inference results are better for topologies that are known to be biologically more plausible.

Other experiments show that inference performance drops as network size increases, even if enormous amounts of data are available. This decay however is not as drastic as what might be expected. In order to infer larger networks with high confidence, it seems that complementary approaches are needed such as adding additional data sources or incorporating domain knowledge.

We observed a clear effect of different types of noise on the inference performance of all algorithms. Experiments show that noise is an important, and sometimes a required factor during inference. All three noise types available in the SynTReN generator provoked a qualitatively different inference behavior, which supports their adoption in generators for synthetic data in general.

A substantial but decreasing benefit of supplying more expression datasets was observed when trying to infer interaction networks. The algorithms tested behave differently, sometimes reaching a maximum performance where adding more arrays becomes pointless. Reaching the inference quality plateau requires enormous amounts of data relative to the size of the inferred network.

The results show the added value of synthetic data in revealing operational characteristics of inference algorithms which are unlikely to be discovered by means of biological micro-array data, and make a strong case for computer models of biological systems in leveraging Systems Biology research.