Abstract

Inferring comprehensive regulatory networks from high-throughput data is one of the foremost challenges of modern computational biology. As high-throughput expression profiling experiments have gained common ground in many laboratories, different techniques have been proposed to infer transcriptional regulatory networks from them and much effort goes to the development of algorithms that infer the structure of transcriptional regulatory networks from this data. In this thesis, the large scale application of simulated gene expression data on network inference algorithms is evaluated and also a novel biclustering model is proposed within the framework of Probabilistic Relational Models.

In the first part of this thesis, a model, called SynTReN, is proposed for generating simulated regulatory networks and associated simulated microarray. This model addresses some of the limitations of previous implementations. Instead of using random graph models, topologies are generated based on previously described transcriptional networks, thereby allowing a better approximation of the statistical properties of real biological networks. The computational performance of our simulation procedure is linear in function of the number of genes, making simulation of large networks possible. 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 alone.

The second part of the thesis focuses on the description of an abstracted model of transcriptional regulation, namely by means of a biclustering model. We propose a probabilistic approach to identify overlapping regulatory modules, called *Pro*Bic, based on the framework of Probabilistic Relational Models. The model naturally deals with missing values and noise and thereby leads to a robust identification of biclusters. Both global and query-driven biclustering are combined within a single model-based approach that allows simultaneous identification of multiple and potentially overlapping biclusters. The powerful combination of Probabilistic Relational Models with an Expectation-Maximization approach allows *Pro*Bic to be easily extended to incorporate additional data sources, ultimately leading to the identification of regulatory modules with associated condition annotation, regulatory motifs and transcription factors.