Methods for the Analysis of Gene Expression Data and Biochemical Networks

Ralf Zimmer

Chair for Practical Computer Science and Bioinformatics, Department of Computer Science,
Ludwig-Maximilians-Universität München, Germany

Abstract

Understanding the molecular interactions and the cellular processes of genes and proteins is a major challenge for bioinformatics in the 'post genome era'. A major application problem is the identification and validation of molecular targets for drug research in pharmaceutical industry. We present methods for the analysis of large scale expression data (i.e. microarray gene expression data) in the context of biochemical (i.e. metabolic and regulatory networks).

We represent molecular network data as formal models called Petri nets, which are well-suited for the task and have been studied for decades such that an extensive mathematical theory is available for their analysis. The Petri net representation of relevant molecular networks are obtained via information extraction from relevant metabolic and regulatory databases, via input from biological experts from the respective fields, and via text mining of scientific literature. Text mining methods have been fine-tuned for the generation of molecular networks based on genes and proteins in particular for certain human diseases. Based on such (possibly very large networks) appropriate algorithms have been developed to compare and statistically analyze expression data, in particular in the context of networks and pathways. An overall goal of these methods is to identify pathways involved in a certain disease state as measured by the expression values of genes and proteins in that state as compared to the normal state.

We have developed ProMiner, a text mining program, which allows to process very large texts (i.e. all Pubmed abstracts) with acceptable sensitivity and specificity (top rank at the recent text mining competition BioCreative) and ToPNet a pure Java software system, which can be used for interactive analysis and visualization of expression and network data in an integrated fashion. ToPNet also contains a set of algorithms for the combined analysis of data in the context of networks (i.e. pathway scoring, pathway search, significant area search, co-clustering of networks and data, and pathway queries). The ToPNet tool, including documentation and tutorials, is freely available for academic use via http://www.biosolveit.de.

The methods and the tools have been developed in collaboration with Aventis pharma Frankfurt and have been applied to several Aventis disease-related research projects. In particular, we have analysed data from osteoarthritis research with the goal to identify and model relevant disease pathways both for target identification and target validation.