Differential Network Analysis and Validation Strategies for High-dimensional Oncological Genetic Data

by Miriam Lohr

Summary

In the past two decades gene expression measurements have become a common tool to get insights into cancer biology. Detecting differences between tumor and normal tissue or tissue of distinct disease stage will help to identify molecular mechanisms suitable as possible drug targets. While discovering single differentially expressed genes has become a standard concept in microarray analysis, the detection of differential interaction networks is still more challenging.

In this thesis, our goal is to improve statistical methods for high-dimensional data to gain deeper insights into cancer biology by analyzing gene expression data. We focus on two topics. First, the large number of available gene expression datasets is used to validate differentially expressed genes or biomarkers in cancers. Second, genes are not considered alone but in interaction networks that might change during disease progression or between different disease stages. We detect differential interaction networks from gene expression data by testing gene sets derived with and without biological prior knowledge.

Using microarray or RNA-seq technology expression of thousands of genes are measured simultaneously which requires adjustment for multiple testing. By considering multiple gene expression datasets significant findings may be validated. Hence, a tradeoff between strict adjustment for multiple testing and validation of results must be determined. We propose two new approaches to validate biomarkers derived from high-dimensional data. The first strategy combines an exploratory screening for markers with a common meta-analysis of validation datasets. The second approach based on sequential validation of considered datasets. By successively reducing the number of genes through the validation steps less adjustment for multiple testing is required. Both approaches are applied to breast and non-small cell lung cancer datasets to detect differentially expressed genes and prognostic markers, respectively. Afterwards, the results are compared to the findings obtained by a common meta-analysis. We show that our approaches are able to detect markers described to be relevant in literature that are missed by the standard meta-analysis and strict adjustment for multiple testing.

Furthermore, we propose a framework for the detection of differential interaction networks. Graphical Gaussian Models which base on partial correlations form the basis of our approaches. The estimation of partial correlations requires a good estimation of the covariance matrix, because partial correlations can be directly derived from the inverse of the covariance matrix. If we deal with genetic data, we often have the more genes that should be incorporated into a network than available observations. Hence, the sample covariance matrix has not full-rank and cannot be inverted. Therefore, we use a linear shrinkage approach for the covariance matrix that guarantees the desired properties.

First, a strategy that combines pre-defined gene sets derived by biological prior with a Gene Set Enrichment Analysis in Gene Ontology (GO) gene sets is proposed to generate hypothesis networks that can be afterwards tested for differential interaction structure. In addition, an Algorithm for Differential Network Gene Selection (DiNGS) is introduced for that purpose. This flexible algorithm is analyzed for stability by stratified bootstrapping and applied to gene expression data of a breast cancer cohort. Starting with a suitable pair of genes a forward selection that aims to build a differential interaction network is performed. The hypothesis of a differential interaction network can be tested by the application of permutation tests based on ordinary or partial correlations. A simulation study is conducted to explore the properties of the 14 proposed permutation tests in terms of holding the α-level if no systematical differences are present and power when we embed changes of interactions into the network of one of the two groups. Surprisingly, our simulations reveal that permutation tests based on ordinary correlations have a higher power than tests using partial correlations to detect differences in gene interaction networks.