Martin Sill
Dr. sc. hum.

Robust Biclustering Analysis of Microarray Data by Sparse Singular Value Decomposition Incorporating Stability Selection

Promotionsfach: Biostatistik (DKFZ)
Doktormutter: Prof. Dr. Annette Kopp-Schneider

The discovery of new molecular subclasses is a common aim in the analysis of high-dimensional gene expression data arising from microarray experiments. For example, in cancer research some tumors are assumed to form heterogeneous molecular subclasses according to the underlying biological processes. In order to find such subclasses and to identify new marker genes and possible targets for cancer treatment, clustering methods are applied. Traditional one-way clustering methods such as hierarchical clustering are often not appropriate to analyze high-dimensional data sets and therefore new clustering concepts are needed.

Over the past decade, the biclustering concept emerged in the field of microarray data analysis. The biclustering concept describes the simultaneous clustering of the rows and the columns of a data matrix. Several biclustering approaches have been published for the analysis of high-dimensional gene expression data. Despite of huge diversity regarding the mathematical concepts of the different biclustering methods, many of them can be related to the singular value decomposition (SVD). Recently, a sparse SVD approach (SSVD) has been proposed to reveal biclusters in gene expression data. In contrast to traditional one-way clustering, where several methods for assessment of cluster stability are known, similar approaches for biclustering are not yet available. In this thesis, a new biclustering method that combines the SSVD with stability selection is proposed. Stability selection is a subsampling-based variable selection that allows to control Type I error rates. The newly developed S4VD algorithm incorporates this subsampling approach to find stable biclusters, and to estimate the selection probabilities of genes and samples to belong to the biclusters. So far, the S4VD method is the first biclustering approach that takes the stability of biclusters regarding perturbations of the data into account.

In a simulation study, the S4VD algorithm outperformed the SSVD algorithm and two other SVD-related biclustering methods in recovering artificial biclusters in simulated data and in being robust to noisy data. Application of the S4VD algorithm to two cancer related microarray dataset revealed biclusters that correspond to coregulated genes associated with known cancer subtypes. A gene set enrichment analysis showed that the genes associated with the biclusters belong to significantly enriched cancer-related Gene Ontology categories. Moreover, marker genes for the different cancer subtypes showed high selection probabilities of belonging to the corresponding biclusters. Using the capabilities of the R statistical software, implementations of the S4VD algorithm, the SSVD algorithm and additional visualization functions have been made available. Furthermore, the S4VD algorithm has been integrated into a recently developed interactive visualization software for the interactive analysis of gene expression data.