Statistical Learning for Analyzing Functional Genomic Data

Axel Benner
Dept. of Biostatistics (C060)
German Cancer Research Center
Im Neuenheimer Feld 280
D-69120 Heidelberg
Email: benner@dkfz.de

An important topic concerning the statistical analysis of functional genomic data is multivariable predictive modelling, where the best prediction of a given outcome variable is sought. Since in microarray studies the number of predictor variables is much larger than the number of observations, standard statistical model building does not work properly. Statistical learning is a new approach to develop prediction models allowing the inclusion of all available data. Selection methods like boosting and regularization methods like penalized regression have been recognized as important statistical learning methods which can control for complexity.

Validation of the fitted models by using independent test samples, bootstrap resampling or cross validation is another important issue. The methods presented above enable for adaptive model selection by tuning their parameters and the set of variables included. Variable selection and choice of parameters is often done by minimization of the cross validated error rates. To estimate the prediction error at least double cross validation is necessary.

We illustrate and compare the different approaches using a data set on predicting survival for patients with acute myeloid leukemia. The results will be compared with respect to the prediction error and interpretability of the results.
