Building prognostic gene signatures with RNA-Seq data

Isabella Zwiener[1,2], Barbara Frisch[2], Harald Binder[2]

[1]Center for Thrombosis and Hemostasis (CTH), University Medical Center Mainz
[2]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center Mainz

One major aim of RNA-Seq experiments, in addition to finding differentially expressed genes, is building prognostic signatures. If the differentially expressed genes shall be validated using e.g. qRT-PCR, we may be interested in a short list of the most important genes. A penalized regression model with automated variable selection can be used for both: it leads to sparse models including only a small number of differentially expressed genes and it can be used for prediction.

Before applying a penalized regression model we will have a closer look at the specific properties of RNA-Seq data. As RNA-Seq produces count data, the expression data are skewed and include some extreme outliers. Furthermore, RNA-Seq data show a mean-variance dependency, i.e. genes having a large mean expression value tend to have larger variances. In penalized regression models the selection probability of a covariate depends on the variance and covariates with large variances are preferred. To account for this, standardization of covariates has been proposed. Standardization includes estimation of variances, which will be affected by skewness and outliers. The question arises, if a transformation of the data is necessary to achieve prognostic signatures with higher sensitivity, specificity and better prediction performance.

In a simulation study we compare sensitivities and specificities of different transformations on RNA-Seq data. Chosen transformations include the default of standardization, the log transformation, a variance-stabilizing transformation, and rank-based transformations. To build a penalized model we use componentwise likelihood-based boosting. It can handle the high dimensionality of RNA-Seq data and allows adjusting for clinical covariates. The simulations are based on real RNA-Seq data including extreme outliers, skewed distributions, and the typical mean-variance dependency of count data. Furthermore, we will apply all transformations on real data from patients with kidney renal clear cell carcinoma to predict their survival times.

The results of the simulation study show, that transforming the data has a large impact on sensitivity and specificity. The rank-based transformations perform well in all simulation scenarios. The default of standardizing covariates performs quite well in some situations, but in other situations also even worse than not standardizing covariates at all. In the application example we will see that the prediction error aslo depends on the transformation used.

This illustrates that we should be aware of the specific properties of RNA-Seq data before building a penalized regression model. Sensitivity, specificity and prediction error depend on the choice of a suitable transformation.