Pathobiochemical signatures of cholestatic liver disease in bile duct ligated mice

Statistical analysis (Matthias Koenig)

Introduction

This document gives an overview over the planned analysis.
The following naming is used

factor : one of the measured quantities over time, i.e. either

i) gene expression of a single gene (e.g. Actb RNAa);
ii) a biomarkers (i.e. ALT, albumin)
iii) one of the histological markers (e.g. BrdU-positive Kupffer cells)

time point: a single value from the measured time points {**0h** (control), **6h**, **12h**, **18h**, **30h**, **2d**, **5d**, **14d**}

time period: a period from the defined periods {**0h** (control), **initial**={6h, 12h}, **perpetuation**={18h, 30h, 2d}, **progression**={5d, 14d}

Correlation/Regression analysis

Correlation analysis between two factors will be implemented based on a method for correlation analysis in time-course gene expression data {Son2008}. The method is based on standard Pearson/Spearman correlation coefficients taking the temporal ordering of values into account. For control/comparison all analysis will be performed in parallel with standard Pearson and Spearman regression analysis.

Clustering analysis of the factors will be performed based on these calculated correlation coefficient (in interval [-1,1]) and presented in a cluster/correlation matrix with associated cluster dendrogram (replaces figure 7). Figure 8 and 9 will be reduced to present the analysis of the regression value for the complete time course (and within periods).

The complexity of the plots will be reduced.

Son, Young Sook, and Jangsun Baek. "A modified correlation coefficient based similarity measure for clustering time-course gene expression data." Pattern Recognition Letters 29.3 (2008): 232-242.

Decision Trees

The Problem of prediction (classification) of individual time points or time periods from a given set of factors (predictors) is a classical (supervised) classification problem. These problem will be solved with a classifier/predictor based on decision trees.

The prediction problem of time points is special due to the existence of a temporal ordering in the time point/period classes (i.e. later periods/time points come after earlier ones), resulting in ordinal prediction classes. Consequently, an analysis approach for ordinal classification based on decision trees will be applied using extended binary classification {Frank2001, Li2006}. For control/comparison simple decision trees on nominal data will be calculated in parallel.

Confidence intervals of predictions will be calculated based on bootstrap methods. Classifiers will be evaluated based on ROC curves (bootstrap validation with proper split in training (4 mice) and test dataset (1 mouse).

Frank, Eibe, and Mark Hall. A simple approach to ordinal classification. Springer Berlin Heidelberg, 2001.
Li, Ling, and Hsuan-Tien Lin. "Ordinal regression by extended binary classification." Advances in neural information processing systems 2006: 865-872.

Significance analysis

The significance analysis between different timepoints within a factor (supplement 4) will be implemented similar to the current analysis, i.e. test between 2 time points/periods for difference (analog to t-test). The necessary corrections for multiple testing will be employed. These corrected p-values will be used for significance calls (insignificant differences are not reported).