Wiener Biometrische Sektion (WBS)
der Internationalen Biometrischen Gesellschaft
Region Österreich – Schweiz (ROeS)
http://www.meduniwien.ac.at/wbs/

WBS Sommer Seminar

Did the genomic data flood overrun the statistical levee?

Statistical approaches to analyse genomic data (*without drowning*)

Datum: Donnerstag, 4. Juli 2013
Ort: Hörsaal der Universitätszahnklinik der Medizinischen Universität Wien,
Sensengasse 2a, 1090 Wien
Plan siehe http://www.unizahnklinik-wien.at/de/patientinnen/anreise/
Beginn: 13:00 Uhr (s.t.)
Ende: 17:45 Uhr
Vorsitz: Georg Heinze (CeMSIIS, Med. Univ. Wien), Stephan Lehr (Baxter Innovations GmbH)

AGENDA

12.30 – 13 Registration and get-together

13 – 13.25 Stephan Lehr (Baxter Innovations GmbH, Wien**):**
Some practical aspects in design and analysis of biomarker studies

13.25 – 13.50 Andreas Gleiss (CeMSIIS, Med. Univ. Wien):
Test statistics for two-group comparisons of zero-inflated intensity values

13.50 – 14.15 Sonja Zehetmayer (CeMSIIS, Med. Univ. Wien):
Stopping rules for sequential trials in high-dimensional data

14.15 –14.40 Markus Jaritz (Research Intitute of Molecular Pathology (IMP), Campus Vienna Biocenter):
Next Generation Sequencing Data Analysis: From ChIP-Seq read islands to epigenomics information

14.40 – 15.05 Christoph Bock (CeMM, Austrian Academy of Sciences)
The Relevance of Next Generation Sequencing for Personalized Medicine

15.05 – 15.30 *Pause (Foyer)*

15.30 – 16.15 Lara Lusa (University of Ljubljana):
Class-imbalanced class prediction for high-dimensional data

16.15 – 16.40 Daniela Dunkler (Med. Univ. Wien):
Gene selection in microarray survival studies under non-proportional hazards

16.40 – 17.25 Harald Binder (Johannes-Gutenberg-University Mainz):
Regularized regression for omics data:
Why one size doesn't fit all, but you nevertheless should try

17:25 – 17:40 Georg Heinze (Med. Univ. Wien):
How the levee bears out against the flood: summary and discussion

18.00 – Informal debriefing in Altes AKH

Registration of attendance (free): per e-mail to franz.koenig@meduniwien.ac.at until 1 July 2013.
Please feel free to distribute the announcement to colleagues. The WBS runs a mailing list for announcing talks in the field of biostatistics. For subscription to the mailing list, send an e-mail to franz.koenig@meduniwien.ac.at as well.

WBS Sommer Seminar, 4. Juli 2013

Did the genomic data flood overrun the statistical levee?

Statistical approaches to analyse genomic data (*without drowning*)

Abstracts

Some practical aspects in design and analysis of biomarker studies

Stephan Lehr
Baxter Innovations GmbH, Vienna, Austria

The aim of this presentation is to set the scene for the workshop from a practical perspective. In particular, some definitions and guidelines will be sketched, possible pitfalls in the development of a prediction rule based on high-dimensional data will be demonstrated and clinical trial designs for predictive biomarker validation will be discussed.

Test statistics for two-group comparisons of zero-inflated intensity values

Andreas Gleiss[1], Mohammed Dakna[2] , Harald Mischak[2] and Georg Heinze[1]
[1] *Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University Vienna, Austria*
[2] *Mosaiques Diagnostics GmbH, Hannover, Germany*

Many experiments conducted in molecular biology compare intensity values obtained by micro-RNA transcriptomics, proteomics or metabolomics ('Omics') procedures between two groups of independent biological samples differing in an experimental condition or in the health status of the subjects. A special characteristic of such data is the frequent occurrence of zero intensity values caused by compounds (e.g., micro RNAs, peptides or metabolites) that cannot be detected in some samples. Zero intensities can arise either by true absence of a compound or by a signal that is below a technical limit of detection.
In the literature the distribution of observed signals is viewed either as a mixture of a binomial distribution (absence or presence of a detectable signal) and a continuous distribution (intensity if signal is present) or as a left-censored continuous distribution. While so-called two-part tests compare mixture distributions between groups, one-part tests treat the zero-inflated distributions as left-censored. We propose to employ a Left-Inflated Mixture model to combine these two approaches.
We perform a comparative simulation study using the setting of a typical omics experiment with the aim to test differential expression in several hundreds of compounds simultaneously. Both types of distributional assumptions as well as combinations of both are considered when comparing power and effect estimation across the various methods considered. We discuss issues of application using an example from proteomics.
We conclude that the considered tests generally show their strengths in scenarios satisfying their respective distributional assumptions. If it is a priori unclear which distributional assumptions best fit the data at hand then a two-part Wilcoxon test can be recommended. Assuming a log-normal distribution the Left-Inflated Mixture model provides direct estimates for the proportions of the two considered types of zero intensities.

Stopping rules for sequential trials in high-dimensional data

Sonja Zehetmayer
Section for Medical Statistics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University Vienna, Austria

Sequential trials have been proposed that allow for an early stopping of the trial in studies involving large scale hypothesis testing as in microarray experiments. To control the False Discovery Rate, multiplicity adjustment is required only for the number of hypotheses but not for the number of interim looks (under suitable assumptions asymptotically for an increasing number of hypotheses). In this talk we introduce novel stopping rules that stop a trial early if a certain success criterion is fulfilled based on the proportion of rejected hypotheses.

Next Generation Sequencing Data Analysis: From ChIP-Seq read islands to epigenomics information

Markus Jaritz
Research Intitute of Molecular Pathology (IMP), Campus Vienna Biocenter

Within the Epigenomics Gen-Au project, several hundred samples have been sequenced using ChIP-Seq technology since 2008. Despite powerful open source tools, many scripts, workflows and analysis strategies had to be developed. We present an overview of our analysis pipeline, addressing issues such as sequencing quality measurements, sample similarity, peak calling, peak overlaps and motif finding.

The Relevance of Next Generation Sequencing for Personalized Medicine

Christoph Bock
CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria
Max Planck Institute for Informatics, Saarbrücken, Germany

In my presentation, I will summarize the role of next generation sequencing for personalized medicine and highlight the relevance of bioinformatic and biostatistical methods for interpreting the vast amount of genome, epigenome and transcriptome data that are being generated at CeMM and at many genomics institutes world-wide. The talk will also discuss our ongoing work with the European BLUEPRINT project consortium (http://blueprint-epigenome.eu/) aimed at establishing comprehensive epigenome maps of hematopoietic cell types and various types of leukemia cells. I will conclude by outlining an integrated computational/experimental approach toward rational design of epigenetic combination therapies (Bock and Lengauer 2012 Nature Reviews Cancer), which we pursue in collaboration between the CeMM Research Center for Molecular Medicine and the Medical University of Vienna.

Contact: cbock@cemm.oeaw.ac.at and http://epigenomics.cemm.oeaw.ac.at.

Class-imbalanced class prediction for high-dimensional data

Lara Lusa and Rok Blagus
Institute for Biostatistics and Medical Informatics, Faculty of Medicine , University of Ljubljana, Slovenia

The goal of class prediction studies is to develop rules to accurately predict the class membership of new samples. The rules are derived using the values of the variables available for each subject: the main characteristic of high-dimensional data is that the number of variables greatly exceeds the number of samples. Frequently the classifiers are developed using class-imbalanced data, i.e., data sets where the number of samples in each class is not equal. Standard classification methods used on class-imbalanced data often produce classifiers that do not accurately predict the minority class; the prediction is biased towards the majority class. High-dimensionality poses additional challenges when dealing with class-imbalanced prediction. We show how class-imbalance impacts six types of classifiers, using simulated data and a publicly available data set from a breast cancer gene-expression microarray study. We also investigate the effectiveness of some strategies that are available to overcome the effect of class imbalance. We devote special attention to SMOTE, a popular oversampling technique and to boosting methods.

Gene selection in microarray survival studies under non-proportional hazards

Daniela Dunkler and Georg Heinze
Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University Vienna, Austria

Many studies aim at finding, among a huge number of candidate genes, those which are possibly linked to survival. With non-proportional hazards Cox regression (CR) could lead to under- or overestimation of effects. Given the large number of genes it is not feasible to determine the functional form of the time dependency for each gene. Moreover, researchers are often interested in a risk score, which could be used to identify high and low risk groups at the start of follow-up. Such a score should be interpretable irrespective of time-dependent effects.

We consider the odds of concordance, which are defined by OC=c/(1-c) where c is the concordance probability P(T1<T0), the probability that a person randomly selected from group G1 dies earlier than a person randomly selected from G0. Since gene expressions are not binary, we generalize OC to continuous data. Under proportional hazards, OC is estimated by the CR hazard ratio. Under non-proportional hazards OC can be obtained from weighted Cox regression (WCR) or a novel method based on conditional logistic regression, called concordance regression (CCR). The latter has the additional advantage that it has an attractive nonparametric analogue.

We demonstrate properties and aspects of application of these novel methods in the analysis of real and simulated studies. We also briefly discuss the possibility of their application in the context of regularized regression models with a high-dimensional predictor space. Concluding, OC and c are concise single numbers useful for clear decisions at time zero. They can be utilized to select genes irrespective of proportional hazards and are obtained by WCR or CCR. WCR and CCR are available as R packages coxphw and concreg, respectiveley, on CRAN.

Regularized regression for omics data: Why one size doesn't fit all, but you nevertheless should try

Harald Binder
Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center Johannes Gutenberg University Mainz, Germany

Modern "omics" techniques for molecular measurement, such as SNP microarrays or RNA-Seq, promise improved prognosis for patients or even personalized medicine. For statisticians this means the challenge of linking high-dimensional covariate vectors to clinical endpoints, such as survival, for identifying important molecular entities. While univariate techniques are popular for controlling false discovery rates, they do not directly provide risk prediction models. Regularized regression techniques, such as the lasso or componentwise boosting, allow to fit regression models with high-dimensional covariate vectors, providing automatic selection of a small number of potentially important molecular entities, e.g. SNP or gene signatures. Unfortunately, application to different types of molecular platforms is not straightforward. Potential pitfalls will exemplarily be illustrated for componentwise boosting in applications to SNP data and to RNA-Seq data. For both types of molecular measurements, the variance of covariates and potential choices for standardization are seen to critically affect signature selection and prediction performance. For the SNP data, there is preferential selection according to minor allele frequencies. Different variants of componentwise boosting are introduced for controlling of this effect. For the RNA-Seq data, the overall shape of the distribution heavily affects selection, and various types of pre-transformations are evaluated for addressing this. Despite these difficulties in handling, regularized regression is seen to be advantageous in dealing with correlation between covariates. This is specifically illustrated for SNP data with linkage disequilibrium structure, as contrasted to the results from univariate approaches. In summary, regularized regression techniques, such as componentwise boosting, are seen to be an attractive tool for omics data, if carefully adapted to the particular platform.