GCB 2013
Workshop Statistical Methods in Bioinformatics
10.09.2013, Göttingen

Speakers:

13:15-13:40	Rainer Spang	Modeling signaling networks with unknown unknowns
13:40-14:05	Lars Kaderali	Individual Cell Population Context in RNAi Data Analysis
14:05-14:30	Holger Fröhlich	Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics
14:30-14:50	Maral Saadati	Statistical Challenges of High Dimensional Methylation Data
14:50-15:10	Manuela Zucknick	Integrated risk-prediction modelling based on multiple genomic data sources with Bayesian variable selection models
15:10-15:35	Coffee Break	
15:35-16:00	Yinyin Yuan	Integrative Modeling of Cancer-Immune Interactions in Triple-Negative Breast Cancers
16:00-16:25	Julien Gagneur	Inferring causal molecular intermediates from omics data in the context of genetic and environmental variations
16:25-16:50	Pratyaksha Wirapati	Challenges in development of practical omics biomarkers for risk prediction
16:50-17:15	Klaus Jung	Global Tests for Expression Data of Molecular Subsets in High-Throughput Experiments

Abstracts:

Modeling signaling networks with unknown unknowns

Rainer Spang
Institute for Functional Genomics, University of Regensburg

In cancer, mutations interrupt and modulate the propagation of signals in both signalling and gene regulation networks. The modulation can be different from patient to patient. Importantly, it can determine whether a targeted drug will be effective for this patient or not. In a personalized medicine setting we can not reconstruct entire networks but need to focus on some key features of them. Moreover we can only generate very limited amounts of data for an individual patient. I will present a simple inference method “No-Conan” that partially reconstructs network features in a way that can not be confounded by unobserved confounders.

Individual Cell Population Context in RNAi Data Analysis

Lars Kaderali
Institute for Medical Informatics and Biometry (IMB), Dresden University of Technology

High-content, high-throughput RNA interference (RNAi) offers unprecedented possibilities to elucidate gene function and involvement in biological processes. Microscopy based screening allows phenotypic observations at the level of individual cells. It was recently shown that a cell's population context significantly influences results. However, standard analysis methods for cellular screens do not currently take individual cell data into account unless this is important for the phenotype of interest, i.e. when studying cell morphology.

We present a method that normalizes and statistically scores microscopy based RNAi screens, exploiting individual cell information of hundreds of cells per knockdown. Each cell's individual population context is employed in normalization. We present results on two infection screens for hepatitis C and dengue virus, both showing considerable effects on observed phenotypes due to population context. Using a cell-based analysis and normalization for population context, we achieve improved sensitivity and specificity not only on a individual protein level, but especially also on a pathway level. This leads to the identification of new host dependency factors of the hepatitis C and dengue viruses and higher reproducibility of results.

Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics

Holger Fröhlich
Algorithmic Bioinformatics, Bonn-Aachen International Center for IT, University of Bonn

Predictive, stable and interpretable gene signatures are generally seen as an important step towards a better personalized medicine. During the last decade various methods have been proposed for that purpose. However, one important obstacle for making gene signatures a standard tool in clinics is the typical low reproducibility of signatures combined with the difficulty to achieve a clear biological interpretation. For that purpose in the last years there has been a growing interest in approaches that try to integrate information from molecular interaction networks. We here propose a technique that integrates network information as well as different kinds of experimental data (here exemplified by mRNA and miRNA expression) into one classiffier. This is done by smoothing t-statistics of individual genes or miRNAs over the structure of a combined protein-protein interaction (PPI) and miRNA-target gene network. A permutation test is conducted to select features in a highly consistent manner, and subsequently a Support Vector Machine (SVM) classifier is trained. Compared to several other competing methods our algorithm reveals an overall better prediction performance for early versus late disease relapse and a higher signature stability. Moreover, obtained gene lists can be clearly associated to biological knowledge, such as known disease genes and KEGG pathways. We demonstrate that our data integration strategy can improve classiffication performance compared to using a single data source only. Our method, called stSVM, is available in R-package netClass on CRAN (*http://cran.r-project.org*).

Statistical Challenges of High Dimensional Methylation Data

Maral Saadati, Axel Benner
Dept. Biostatistics, German Cancer Research Center (DKFZ), Heidelberg

With the fast growing field of epigenetics comes the need to better understand the intricacies of methylation data analysis. Challenges arise from the fact that methylation values (so-called beta values) are proportions between 0 and 1, often from a bimodal distribution with peaks close to 0 and 1. Therefore, the majority of standard statistical approaches do not apply. The logit transformation into so-called m-values is a common approach to circumvent this problem and allow the use of common statistical methods. However, it can be observed that the transformation from beta to m-values does not necessarily result in an approximately normal distribution. Often bimodality, asymmetry and heteroscedasticity are conserved even after transformation. We give an overview and discussion of methods suggested in the recent years that attempt to address the characteristics of methylation data in certain research questions. For example, beta regression models with fixed and random effects for screening of ”differential” methylation between groups while adjusting for confounders, model based clustering to derive methylation phenotypes using mixtures of beta distributions, random forests for classification and prediction of patient survival in high dimensional settings. Our goal is to sensitise researchers to the challenges and issues that arise from this type of data as well as to present possible solutions.

Integrated risk-prediction modelling based on multiple genomic data sources with Bayesian variable selection models

Manuela Zucknick
Dept. Biostatistics, German Cancer Research Center (DKFZ), Heidelberg

Risk prediction based on clinical and molecular information is fundamental in translational cancer research. The statistical analysis of high-throughput omics data allows the identification of prognostic and predictive biomarkers, which can aid the development of targeted therapies.

We present a Bayesian hierarchical regression model with variable selection (BVS) as an alternative to well-known sparse regularisation methods such as lasso regression for risk prediction modelling based on high-dimensional omics input data. A common application in cancer research is the prognosis of patient survival or the prediction of therapy response with simultaneous feature selection. In addition to gene expression data, high-throughput technologies are also available for many other types of genomic data, and today clinical researchers are routinely collecting genome-wide data from various sources on the DNA- and RNA-level. If data from several sources are available for the same set of biological samples, they can be analysed together in an integrative manner, with the aim of providing a more comprehensive picture of the disease biology and improving the performance of feature selection for risk prediction models.

BVS models are very flexible in their setup and are naturally well-suited to extensions allowing the integration of additional data sources. We will present such an extension for the integration of copy number variation or methylation data with gene expression data. We investigate model behaviour and the influence of prior specifications through simulation studies. The model is illustrated in applications in translational oncology.

Integrative Modeling of Cancer-Immune Interactions in Triple-Negative Breast Cancers

Yinyin Yua
Institute of Cancer Research, London, UK

The interplay between cancer and immune cells has important prognostic implications for breast cancers. Recent studies have revealed diverse spatial distributions and heterogeneous molecular profiles of immune cells in breast tumors. To model both the spatial context and molecular profiles of cancer-immune interactions, we proposed to integrate omics data with quantitative immune morphology. Our image analysis tool automates cell-by-cell classification in Hematoxylin & Eosin (H&E) tumor slides and establishes spatial relationships between cancer and immune cells. By using a kernel estimate of cancer density and other statistical tools, we modeled the spatial context derived from pathological images and molecular profiles derived from microarrays to jointly characterize cancer-immune interactions. We applied our approach on three independent subsets of 52, 37 and 56 oestrogen receptor (ER)-negative, Her2-negative (Triple-Negative) breast tumors from the METABRIC study, all with H&E images, DNA copy number, and RNA expression microarray data. We demonstrated that image-based modeling of cancer-immune interactions led to the discovery of prognostic biomarkers independent of standard clinical factors in all three subsets. Integration of image-based biomarkers with RNA expression data further helped delineate immune modules enriched with relevant immune pathways and a PTEN oncogenic signature, indicating an interplay between immunity and important cancer suppressors. The characterization of immune morphology and cancer genomic background has provided interesting hypothesis for how triple-negative breast cancer evades immune surveillance.

Inferring causal molecular intermediates from omics data in the context of genetic and environmental variations

Julien Gagneur
Gene Center, Ludwig-Maximilians-University Munich

Dissecting the molecular mechanisms that link genotype to phenotype promises to deliver the necessary insights to develop drugs tailored to the genetic background and life circumstances of the patient. Information from interventional data is scarce, and hence the challenge resides in developing causal inference strategies to exploit the breadth of observational population-level genetic and molecular profiling data being generated.

Here we investigated to what extent environmental perturbations, combined with genetic variations, facilitate causal inference in molecular networks. Using yeast as a model system, we carried out joint profiling of fitness and gene expression of a genetically diverse population in 5 environmental contexts. We developed novel inference techniques to predict molecular functional intermediates with an environment-specific role on growth. Our approaches leverages on ubiquitous genotype-environment interactions, exploiting the rich statistical independencies they imply. Technically, we build on Bayesian model comparisons, assessing the statistical evidence that a particular transcript carries a mediating role between genetic signal and its environment-specific effect on phenotype. We applied the approach to genome-wide identified transcripts specific for each environment-specific growth QTLs. Comprehensive independent test using the genome-wide deletion collection confirmed the majority of the 400 top-ranking model predictions. Our results show that exploiting condition-specific genetic effects substantially increases the predictive accuracy over approaches based on genetic or environmental variations alone.

Together, these results have wide-ranging implications for the design of clinical omics studies and their integrated analysis across multiple contexts. Furthermore, the dataset is a unique resource to test different inference strategies as, for the first time, large-scale perturbational data for matching conditions is available.

Challenges in development of practical omics biomarkers for risk prediction

Pratyaksha Wirapati
Bionformatics Core Facility, Swiss Insitute of Bioinformatics, Lausanne

In the past decade, many gene expression signatures had been proposed as biomarkers in cancer. However, only a handful managed to survive more extensive validations and used in practice. For some type of cancer, large sample size can be obtained by combining publicly available datasets to obtain more reliable signatures. Nevertheless, there are still challenges in translating such signatures into a specific, customized diagnostic platforms for daily clinical practice. Some of clinical requirements are: (1) ability to process single-sample data, rather than batches, with proper accounting of measurement uncertainty (2) inclusion of already well-established conventional clinical information and biomarkers into the prediction system, (3) relating the assay score to actual risk in the target population (4) relating the projected risk to clinical utility measures, for more flexible and individualized clinical decision making. I will present a prototypical system that attempts to address the above issues, based on a current project in development of early-stage lung cancer prognostic marker, where signatures learned from a compendium of diverse expression microarray platforms need to be translated into practical FFPE assay technology, and incorporated into an “electronic cancer nomogram”.

Global Tests for Expression Data of Molecular Subsets in High-Throughput Experiments

Klaus Jung
Department of Medical Statistics, University Medical Center Göttingen

Global tests are a common tool for correlating the expression data of a molecular set of features with some response variable. In DNA microarray experiments for example, global tests are used to test whether the mean expression profiles of a certain subset of genes differ between two groups of samples. The global test approach is in contrast to that of comparing each feature individually between the two groups. Since a specified subset is typically related to a certain biological function or cellular pathway, the biological interpretation can be facilitated by the result of a global test.

Methodologically, the difficulty of a global test is the high-dimensional setting of the expression data, since even a subset of features is usually larger than the sample size. Depending on the structure of the covariance matrix, classical multivariate test can fail to maintain a specified significance level. Most global test approaches use permutation procedures to correct for this bias.

While global test are already developed and used for the analysis of gene expression data, they are rarely adapted for the analysis of other high-throughput data such as proteomics or epigenomics data. In this regard, we present a new global test procedure for the analysis of high-throughput protein expression data. The new methods were evaluated in a simulation study and applied to protein expression data from 2-D gel electrophoresis.

Apart from their applicability to a only one type of expression data, global test are also useful for integrating different types of ‘omics’ data. We present an approach of co-analyzing individual microRNAs with their specific set of target mRNAs [1]. The latter one is analyzed by means of global test procedures .

[1] Artmann S, Jung K, Bleckmann A and Beißbarth T (2012): Detection of simultaneous Group Effects in microRNA Expression and related Target Gene Sets. PLoS ONE, 7, e38365.