Contributed sessions 1

A New Framework for the Identification of Genomic Structural Variants Using Joint Alignment of Reads

Anish M.S. Shrestha, Martin C. Frith, Kiyoshi Asai, and Hugues Richard

The problem of aligning a query sequence to a reference sequence is a well-studied, fundamental problem in computational biology, and has been seen as a basic tool in genomic studies. However, the recent shift of paradigm in genomics, propelled by the advances in DNA sequencing technologies, has changed the point of view. Genomic variants are no longer represented by single (possibly long) query sequence but rather by a population of shorter reads covering each position in the query. Surprisingly, despite this change the current practice is still to independently align each of these reads to the reference. This necessitates cumbersome and often ad-hoc downstream procedures in order to consolidate information obtained from the set of alignments. We present a probabilistic framework that extends traditional scoring schemes and alignment techniques to jointly align a population of reads to a reference, in order to detect genomic rearrangements. We demonstrate the advantages of our method by applying it to the problem of identifying long deletions in a query genome with respect to a reference, when the query sequence information is in the form of reads from a high-throughput sequencing machine.

Factorization of count matrices with application to gene expression profile analysis

Ghislain DURIF, Franck PICARD, Sophie LAMBERT-LACROIX

We propose a Gamma-Poisson factor model, based on generalized PCA. In particular, the data matrix Xnp is supposed to depend on latent factors or components and its entries are supposed to follow a Poisson distribution, which is adapted for counts. The matrix np of Poisson intensities is factorized into a product of two parameter matrices UVT where UnK and VnK respectively quantify the observation and variable contributions to the K latent factors. To account for the covariance structure within the data matrix, and especially the possible correlation between covariates (e.g. genes), we introduce gamma priors onto the entries of the parameter matrices U and V. This constitutes a more complete and flexible model than for instance Non-Negative Matrix Factorization, that is based on a Poisson model and that assumes independence between covariates.

The Gamma-Poisson distribution also model over-dispersion, which often characterizes NGS data.

Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces

Nils Ternès, Federico Rotolo, Georg Heinze, Stefan Michiels

Stratified medicine seeks to identify gene signatures predicting whether a patient will benefit from a treatment. We evaluated several approaches to identify such signatures using high-dimensional Cox models in randomized clinical trials (RCT).

We investigated five approaches: lasso penalty on biomarker main effects and biomarker-by-treatment interactions (full-lasso), with possible biomarker-specific weights (adaptive lasso); control of main effects by principal components or ridge penalty, and lasso penalty on interactions (PCA+lasso or ridge+lasso); lasso within a ‘modified covariates’ regression model (Tian et al. 2014). We performed simulations under null and alternative scenarios with n=500 patients and p=500 biomarkers and we varied the number of true main effects and of treatment-modifiers. We also proposed two novel measures of treatment effect prediction for gene signatures to evaluate the interaction strength: a between-arms difference in C-indices and a Wald-based interaction statistic. We finally illustrate the methods using gene expression data on adjuvant chemotherapy in a large data set of breast cancer patients.