Contributed session 4

New results on Statistical significance for sequence analysis: Taking into account the length of the segment of local score. Where the local score is realized: a non-intuitive result.
Agnès LAGNOUX, Sabine MERCIER and Pierre VALLOIS

Biological sequence analysis and comparison are widely used tools. It is one main task of Bioinformatics to develop tools which allow to find significant sequence region or similarities in a data base. The key problem is to establish the p-value of the local score for a given sequence analysis or sequence comparison algorithm, a given scoring scheme, and a given null model. For the analysis case, the statistical significance of the local score under the independent null model have been largely studied from the 1990's with first the approximation of Karlin et al. [3]. Then the local score distribution has been given explicitly in 2001 in the independent model [4] and in a Markovian one in 2007. Here, we investigate how the length of the segment that realized the local score impacts the pvalue.
We also present a non-intuitive result on where the local score is realized. We observe on simulated examples that the p-value and the significant sequences differ whenever the length of local score is taken into consideration or not. When the length of sequences is large, the distribution of the pair (local score; length of the local score segment) can be approximated by the related pair defined in the Brownian motion setting. Here we present the following non-intuitive result in the case where the scoring is centered with unit variance. We actually prove that the probability that the local score is achieved “at the end” of a large sequence (on the last incomplete excursion of the Lindley process associated to the sequence under concern) is constant and approximatively equals 1/3 [1]. We also investigate numerically the more usual case, i.e. when the average score is non positive.

Continuous testing for Poisson process intensities
Franck Picard, Etienne Roquain, Anne-Laure Fougères, and Patricia Reynaud-Bouret

Next Generation Sequencing technologies now allow the genome-wide mapping of binding events along genomes, like the binding of transcription factors for instance. More generally, the field of epigenetics is interested in the regulation of the genome by features that are spatially organized. One open question that remains is the comparison of spatially ordered features along the genome, between biological conditions. An example would be to compare the location of transcription factors between disease and healthy individuals. We propose here to model the spatial occurences of genomic features in each condition by a Poisson process with a heterogeneous intensity on [0,1], and we restate the problem as the comparison of Poisson process intensites in continuous time. Contrary to global testing approaches that consist in testing whether the two intensities are equal on [0,1], we focus on a local testing strategy using scanning windows. Our method is based on kernel to build the test statistics, and on monte-carlo simulations to compute the p-value process. By using the continuous testing framework, we provide a procedure that controls the Family Wise Error Rate as well as the False Discovery Rate in continous time. We illustrate our method on experimental data, and discuss its extensions in the general framework of testing for Poisson process intensities.

Genome-wide generalized additive models
Alexander Engelhardt, Georg Stricker, Daniel Schulz, Matthias Schmid, Achim Tresch, and Julien Gagneur

Chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) is a widely used approach to study protein-DNA interactions. To analyze ChIP-seq data,

practitioners require combining tools based on different statistical assumptions and dedicated to specific applications such as calling protein occupancy peaks or testing for differential occupancies. Here, we present genome-wide generalized additive models (genoGAM), which brings the well-established and flexible generalized additive models framework to genomic applications using a parallelization-by-the-data strategy. We model ChIP-seq read count frequencies as products of smooth functions along chromosomes. Smoothing parameters are estimated from the data eliminating ad-hoc binning and windowing needed by current approaches. A peak caller based on GenoGAM fits is shown to outperform state-of-the-art approaches. The method also provides significance testing for differential occupancy. Application to a histone methylation profiling study in yeast shows controlled type I error rate and increased sensitivity over existing methods. Furthermore, applicability of genoGAM to DNA methylation data analysis is demonstrated.