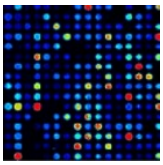
Séminaire – Math For Genomics

Séance du mercredi 3 avril 2019. 10h30.
Evry. IBGBI. LaMME.

Changepoint detection with kernels



Laboratoire de Mathématiques et Modélisation d'Évry
LaMME

université evry val-d'essonne

GENOPOLE

Morgane Perre-Jean (CEA, Evry)

Title : Statistical methods for DNA copy number segmentation in cancer studies

Normal cells have two copies of DNA, inherited from each biological parent of the individual. Changes in DNA copy numbers are a hallmark of cancer cells. Therefore, the accurate detection and interpretation of such changes are two important steps toward improved diagnosis and treatment. In tumor cells, parts of a chromosome of various sizes (from kilobases to a chromosome arm) may be deleted or copied several times. The analysis of copy number profiles measured from high-throughput technologies such as array-comparative genomic hybridization (array-CGH), Single Nucleotide Polymorphism array (SNP array) or high-throughput DNA sequencing data (WGS and WES) raises a number of statistical and bioinformatic challenges.

DNA copy numbers in tumor cells are piecewise constant along the genome. SNPs arrays and sequencing techniques provide both the DNA copy number and the heterozygosity at a large number of position along the genome. As a result, the signal is composed of two dimensions. We first present a method to simulate realistic datasets of DNA copy number profiles. Then, we present an algorithm using the two dimensions of the signal. We show the performance of change-point detection is improved if we use the two dimensions. We implement both methods and simulation framework into an R package named `jointseg`.

This is joint work with Pierre Neuvial and Guillem Rigaill.

Alain Celisse (Université de Lille. Laboratoire Painlevé)

Title : Kernseg: A new efficient change-points detection procedure for analyzing biological data

Summary : The main focus of this work is given to detecting homogeneous regions along the genome by means of a change-points detection approach. The typical signal we will consider is made of the total copy number (TCN) and the allele frequency (FracB). As we illustrate, our approach enjoys several assets: it allows for detecting changes not limited to the mean or the variance of the observations, it does not rely on any parametric distributional assumptions, and it allows for combining several biological time-series to get a higher detection power.

In this direction, our contributions are three-fold:
1- a generalization of classical change-points detection strategies dedicated for real-valued data,

2- a new formulation of the dynamic programming algorithm when combined with the use of reproducing kernels which reduces the memory complexity to $O(n)$ instead of $O(n^2)$ (where n is the number of observations) and the time complexity to $O(n^2)$ instead of $O(n^4)$ with a naive implementation,

3- a wide simulation study where our proposal is compared with state-of-the-art approaches for detecting changes along the TCN and FracB signals.

In particular, we observe that our approach (implemented in the R-package Kernseg) achieves a higher detection power especially in difficult settings (with a low purity).