Statistical and computational challenges for population-based segmentation of copy-number profiles

Franck Picard
Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558
Université Lyon 1, F-69622, Villeurbanne, France

Genes copy number is tightly regulated : two copies of each gene are generally present in diploid genomes, and deviations from this reference can cause massive disorders. Chromosomal aberrations range from massive rearrangements like in late stages of cancers, to point deletions/insertions, and assessing the genome-wide copy number profiles of populations has become a central task in cancer genomics and human genetics. The microarray technology has been applied to the measurement of genome-wide copy numbers is 2001, and this technology is now used in routine on groups of patients. The statistical task associated with the analysis of such data is to detect abrupt changes along the genome that correspond to gene copy number imbalances with repect to a reference genome. Segmentation has been a succesful strategy : it relies on a Gaussian regression model whose mean parameter changes abruptly at unknown coordinates along the genome. This model can be enriched to account for different states of genomic regions that can be amplied/normal or deleted for instance. These models are particularly efficient thanks to the Dynamic Programming algorithm that can be used to compute the best breaks position according to a least-square criterion. However, the quadratic algorithmic complexity of DP has limited their use on high-density arrays or next generation sequencing data. This complexity issue is particularly critical for the joint segmentation of many profiles. In this presentation we will introduce segmentation models and the application of dynamic programming to the associated estimation framework. Then we will present a population-based model for segmentation as well as its associated computational challenges that we solve by developping an at worst linear DP algorithm for segmentation, and by providing a parallel version of the algorithm for population-based data.

This is a joint work with Guillem Rigaill (Laboratoire Statistique et Génome, UMR CNRS 8071, USC INRA Université d'Evry, F-91037 Evry, France), et V. Miele (Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558 Université Lyon 1, F-69622, Villeurbanne, France).