Bayesian Smoothing Of MicroArray's Copy Number Data

Xavi Puig[1], Josep Ginebra[1] and Silvia Beà[2]

[1] Departament d'Estadística. Universitat Politècnica de Catalunya
[2] IDIBAPS. HospitaL Clínic. Universitat de Barcelona

Most human diseases, including cancer, are characterized by chromosomal alterations, which are defined as regions with an increase or decrease of genetic material. Identifying precisely such alterations will help in the localization of important regions that harbour crucial genes whose alteration may have a pivotal role in tumorigenesis, and may also help in identifying prognostic biomarkers. Recently, several microarray platforms (mainly CGH-array and SNP-array) were developed to detect copy number alterations of the whole genome in a single experiment. One of the current platforms is the 100K-SNP array, which contains more than 100,000 SNPs, and is widely used to detect alterations and loss of heterozigosity concomitantly. Alterations can be classified into four categories: gains of one copy, multiple copy gains (amplifications), loss of one copy, and loss of two copies (homozygous deletions) and loss of heterozigosity is considered positive or negative. However, the current methods and software for analyzing such data are not fully developed. The aim of the present work is to present a model that allows for an easy, automated and objective classification of alterations of 100K-SNP arrays. For this purpose we used the dataset of chromosome 18 (with 3,566 copy number measurements) from an aggressive human lymphoma. The measurements of fluorescence of each SNP intensity are used to infer copy number, but show a certain degree of variability. Moreover, this intrinsic noise often makes difficult to detect the boundaries of the alterations. Nevertheless, one expects that neighbour measurements should show similar magnitudes of copy number and consequently a smoothing tool is required for proper interpretation of the data. Usually tools based on non-parametric methods and hidden Markov models and intrinsic algorithm are used for this purpose. Instead, we propose a Bayesian smoothing method based on a conditional gaussian autoregressive model, in which one assumes that a measurement is similar to its neighbours. The idea of this smoothing is inspired on the model proposed by Besag, York and Mollie in 1991, which is very popular in spatial analysis. In a similar way, Brot and Richardson in 2006 proposed a method to analyze CGH-arrays. Instead, we analyze SNP-array. The Bayesian smoothing has the advantage that one disposes of a posterior distribution at each point measurement and one can use the probability that this measurement takes values larger than a fixed scalar to help classifying points between loss, no alteration, and gain. The level of smoothing can be naturally regulated, similar to bandwidth of kernel density, through prior distribution, but a better alternative is to use a hierarchical model and then estimate its posterior probability density. Due to the complexity of the SNP-array data it is necessary to show graphically the results in a nice and clear way simultaneously with the complementary information about loss of heterozigosity. Markov Chain Monte Carlo algorithms are used to estimate by simulating the posterior distribution of the parameters and hyperparameters of the model through WinBugs, while graphical representation of results is done through R. This analysis is also applicable to the rest of the chromosomes that configure the human genome. Moreover, such analysis could be performed in groups of patients that share a common type of cancer, which will be very useful to identify at high resolution level the minimal regions of gain and loss (including amplification and loss homozygous), that pinpoint important cancer genes. In this sense, we are currently working to reach this purpose.