Analysis of CGH arrays using MCMC with Reversible Jump: detecting gains and losses of DNA and common regions of alteration among subjects

Oscar M. Rueda[1], Ramon Diaz-Uriarte[1]

[1]Statistical Computing Team, Structural and Computational Biology Programme, Spanish National Cancer Center (CNIO), Melchor Fernández Almagro 3, Madrid, 28029, Spain

Abstract

Copy number variation (CNV) in genomic DNA is linked to a variety of human diseases (including cancer, HIV acquisition and progression, autoimmune diseases, and neurodegenerative diseases), and array-based CGH (aCGH) is currently the main technology to locate CNVs. To be immediately useful in both clinical and basic research scenarios, aCGH data analysis requires accurate methods that do not impose unrealistic biological assumptions and that provide direct answers to the key question "What is the probability that this gene/region has CNAs?". Current approaches fail, however, to meet these requirements.

We have developed RJaCGH, a method for identifying CNAs from aCGH. We use a non-homogeneous Hidden Markov Model fitted via Reversible Jump Markov Chain Monte Carlo, and we incorporate model uncertainty through Bayesian Model Averaging. RJaCGH provides an estimate of the probability that a gene/region has CNAs while incorporating inter-probe distance. Using Reversible Jump we do not need to fix in advance the number of hidden states, nor do we need to use AIC or BIC for model selection. We presented a first version of our model at UseR two years ago. Since then, we have explored different approaches to improve convergence and speed-up computations, including usage of Gibbs sampling vs. Metropolis-Hastings, delayed rejection, and coupled parallel chains.

Based on the output from RJaCGH, we have also developed two probabilistically-based methods for the indentification of regions of alteration that are common among samples. Our methods are unique and qualitatively different from existing approaches, not only because of the use of probabilities, but also because they incorporate both within- and among-array variability and can detect small subgroups of samples with respect to common alterations. The two methods emphasize different features of the recurrence (sample heterogeneity, minimal required evidence for calling a common region) and, thus, will be instrumental in the current efforts to standardize definitions of recurrent or common CNV regions, cluster samples with respect to patterns of CNV, and ultimately in the search for genomic regions harboring disease-critical genes.

We will discuss the statistical features of our models, as well as the implementation of RJaCGH, including the combined usage of R and C, and different approaches for improving speed and decrease memory consumption.