snapCGH (segmentation, normalisation and processing of arrayCGH data) and methods for combining with gene expression information

Mike Smith, John Marioni, Natalie P Thorne, Simon Tavaré
Hutchison/MRC Research Centre, Department of Oncology, University of Cambridge, England

Array comparative genomic hybridisation (arrayCGH) is a technique allowing detection of copy number changes of DNA segments. A microarray is assembled by spotting DNA onto a substrate. Next, test and reference DNA, each labelled with a different fluorochrome, are hybridized to the array. The ratio of the hybridized intensities along the chromosomes provides a measure of the relative copy number between the test and reference samples.

Several statistical methods for the detection of copy number changes (segmentation) have been proposed, but currently the relative merits of each have only been briefly investigated. This is partly due to the disparity between the required formats of input data for the separate methods, which results in, at most only a couple of methods being employed on a given set of data. We have designed a new package (*snapCGH*), which defines a common object type allowing straightforward implementation of several different methods on a single dataset. The *snapCGH* package is also designed to be compatible with the widely used microarray package *limma*, allowing smooth transitions between each stage in the analysis.

The package also implements a new segmentation algorithm, called BioHMM, a heterogeneous hidden Markov model that enables the incorporation of biological data (e.g. distance between probes and probe quality) currently neglected by other segmentation methods. Additionally we have provided the facility to simulate arrayCGH data from multiple experimental platforms to allow more thorough comparisons of segmentation methods i.e. assessment of their performance with different array platforms.

We are also developing ways to incorporate gene expression information with arrayCGH data. These include methods of displaying both types of data alongside each other in meaningful ways and the dynamic retrieval of web-based resources initiated through user interaction with plotting functions.

We will present the methods we have developed for visualising and comparing segmentation methods. Addtionally we will give an overview and example of how users can combine and analyse arrayCGH and expression data together using the *snapCGH* package.

The latest version of the *snapCGH* package is available from:
http://www.bioconductor.org/packages/bioc/1.8/html/snapCGH.html