Modeling Array CGH data: Comparison of various approaches

Course Project for STAT675 (Rice University) or GS010103 (GSBS)

Objective: Array-based CGH (aCGH) is a high-throughput technique to measure DNA copy number alterations (CNA) across the genome. These are measures of the intensity rations of test and reference DNA samples and provide information about the number of copies in DNA. See `http://www.nimblegen.com/products/cgh/index.html` for details about the technical aspects of such experiments. This projects aims to compare different methods available to analyze such data.

You can use the following steps as guidelines for the project.

1. One of most popular methods used for analyzing aCGH data is the CBS algorithm of Olshen et al (2004). The method and software available for download are available at `http://www.mskcc.org/mskcc/html/18551.cfm`

2. Hidden Markov Models (HMM) are especially useful in modeling such data. Two methods in this regard are by Fridlyand et al. (2004) and Guha et al. (2007). Both papers use HMM's to model aCGH data, with the latter proposing a Bayesian method. The Fridlyand paper and sofware can be downloaded from `http://www.biostat.ucsf.edu/janef`. The Guha paper is available at `http://www.bepress.com/harvardbiostat/paper24`. A MATLAB software for the implementation of the Guha paper is available at: `http://www.mathworks.com/products/bioinfo/demos.html?file=/products/demos/shipping/bioinfo/acghhmmdemo.html`.

3. You can use the Pancreatic adenocarcinoma data from Alguirre et al.,(2004) available at `http://genomic.dfci.harvard.edu/Pancreas_CDNA_Data.htm` which was also analyzed by Guha et al.

4. You should critically evaluate all the methods. Discuss the merits and demerits of each procedure.

References

Aguirre, A.J., Brenman, C., Bailey, G., Sinha, R., Feng, B., Leo, C., et al. (2004). High-resolution characterization of the pancreatic adenocarcinoma genome. PNAS 101, 9067-9072.

Fridlyand J., Snijders A., Pinkel, D., Albertson D. G. and Jain, A. N. Application of Hidden Markov Models to the analysis of the array CGH data. (Special Genomic Issue of Journal of Multivariate Analysis, June 2004, V. 90, pp. 132-153)

Guha S., Y. Li and D. Neuberg (2007). Bayesian Hidden Markov Modeling of Array CGH Data. To appear in Journal of the American Statistical Association.

Olshen, A.B., Venkatraman, E.S., Luciito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5, 557-572.