cn.FARMS - a probabilistic model to detect DNA copy numbers

Djork-Arné Clevert[1, 4], Marianne Tuefferd[2], An De Bondt[3], Willem Talloen[3],
Hinrich W.H. Göhlmann[3] and Sepp Hochreiter[1]

[1] Institute of Bioinformatics, Johannes Kepler University Linz, 4040 Linz, Austria
[2] JE2492, Clinical Genomic Epidemiology, IFR69, University Paris XI, France
[3] Johnson & Johnson Pharmaceutical Research & Development. A division of Janssen Pharmaceutica, Beerse, Belgium
[4] Department of Nephrology and Internal Intensive Care, Charité University Medicine, Berlin, Germany

Motivation:

Existing pre-processing methods for DNA microarrays designed to detect copy number variations (CNVs) lead to high false discovery rates (FDRs). High FDRs misguide researchers especially in the medical context where CNVs are wrongly associated with diseases.

We propose a probabilistic model, cn.FARMS, for detecting CNVs. We suggest modeling the DNA copy numbers as a latent variable in our factor analysis model, which is optimized by Bayesian maximum a posteriori estimation. We originally developed this model for gene expression arrays but we extend our approach for copy number analysis. In comparison to other approaches we don't need to correct for sequence effects to remove the typical chromosome specific wave pattern correlated with GC content.

cn.FARMS can be used for single-locus and multi-loci CNV analysis. The latter constructs so-called "meta-probe sets" by before any summarization combining probes from adjacent probe sets to one measurement and then extracting the probabilistic level of agreement of the different loci on the copy number. CNV estimates with high mutual loci agreements yield considerably reduced FDRs.

Results:

We compared cn.FARMS on a HapMap Mapping250K_Nsp benchmark data set to CRMA, dChip, CNAG and CNAT v4. The aim described in this presentation is to distinguish males from females based on the X chromosome copy numbers, where males possess one copy and females two.

The ROC curve serves to compare the FDR for different true positive rates. For single loci cn.FARMS performs as well as the best method, CRMA, and for multi-loci cn.FARMS clearly outperforms its competitors.