CN-FARMS: a probabilistic model to detect DNA copy numbers

1Institute of Bioinformatics, Johannes Kepler University Linz 4040 Linz, Austria; 2Johnson & Johnson Pharmaceutical Research & Development, Janssen Pharmaceutica N.V., Beerse, Belgium; 3Department 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 latent variable model, cn.FARMS, for array-based CNV analysis which controls the FDR without loss of sensitivity. At a DNA region, cn.FARMS constructs a model by a Bayesian maximum a posteriori estimation where the unobserved, latent variable represents the copy number that is measured by observed genetic markers (probes). The latent variable’s prior prefers parameters which represent the null hypothesis, (same copy number for all samples), from which the posterior can only deviate by a high information content in the data. The more probes agree on the region’s copy number, the less is the uncertainty about the latent variable’s value, the higher is the information content.

Results: We compared cn.FARMS on a HapMap Mapping250K_Nsp and SNP60K.0 benchmark data set to CRMAv2 and dChip. The comparison is based on the sex-determination on the X chromosome, where males possess one copy and females two. The ROC curve serves to compare the FDR for different true positive rates. In both experiments cn.FARMS yielded the best classification results.

Availability: This approach is publicly available in R at http://www.bioinf.jku.at/software.

CN-analysys as a Three-Step Pipeline

The Model

Our approach is based on a linear model with Gaussian noise. Denote the actually observed sum of allele A and B to be the mean normalized and log-transformed PMs by $z$ and the copy number variation in the hybridization mixture by $x$. Then we assume that the log-observations $z$ depend on the true copy number variation $x$ via

$$z = x + e,$$

where

$$x, e \in \mathbb{R}^n,$$ with $s = n = 400, 1000, 2000$, $e \sim \mathcal{N}(0, \Psi)$.

In equation (1), $x$ models the latent factor in the data $z$ while $e$ models the independent noise in each probe of each array. According to the model, the observation vector $z$ is Gaussian distributed:

$$z \sim \mathcal{N}(0, \lambda^2 + \Psi).$$

Model parameters were estimated with the expectation-maximization (EM) algorithm, modified to maximize the Bayesian posterior

$$p(\lambda, \Psi \mid z) \propto p(z \mid x, \lambda, \Psi) p(\lambda, \Psi).$$

Sparse Overcomplete Representation

Meta-Probe Sets

Datasets

We compared cn.FARMS on a HapMap Mapping250K_Nsp and SNP60K.0 benchmark data set to CRMA and dChip. The aim described in this poster is to distinguish males from females based on the X chromosome copy number, where males possess one copy and females two. The benchmark data is publicly available at http://ftp.hapmap.org.

References

W. Talloen, D. A. Clevert, S. Hochreiter, A. Mutterer, L. Bernien, S. Kais, H. Göhlmann: Calling probe sets information for the informative re-annotation of all SNPs on high-density SNP chips. Bioinformatics 2007 23(7):i55-i62


Acknowledgment

I acknowledge the support of the ISCB, Department of Energy, and National Science Foundation in the form of an International Travel Grant, which enabled me to attend this conference.