A pipeline based on multivariate correspondence analysis with supplementary variables for cancer genomics

Christine Steinhoff[1,*], Matteo Pardo[2] and Martin Vingron[1]

[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestr 63-73, 14195 Berlin, Germany

[2] SENSOR Laboratory, CNR-INFM, Via Valotti 9, 25133 Brescia, Italy

***** christine.steinhoff@molgen.mpg.de

The development of several high throughput gene profiling methods, such as comparative genomic hybridization (CGH) and gene expression microarrays enables for studying specific disease patterns in parallel. The underlying assumption for studying both genomic aberrations and gene expression is that genomic aberration might effect gene expression either directly or indirectly. In cancer research, in particular, there have been a number of attempts to improve cancer subtype classification or study the relationship between chromosomal region and expression aberrations.
The intuitive way to analyze different data sources is separately and consecutively, e.g. first determine regions with copy number aberrations (possibly tissue or patients -specific) and then look for differentially expressed (onco)genes inside these regions [1]. There is a natural reason for integrating results rather than data: strong heterogeneity does not allow sensible alignments of the source data. Still, integrative approaches –where data are fused before their analysis- are preferable. Only recently, few integrative methods have been published [2]. Nevertheless, these approaches do not integrate covariate data like tumor grading, mutation status and other disease features. These features are frequently available and of interest for an integrative analysis.
We address these two problems, namely jointly analyzing different data sources and integrating supplementary categorical data. Furthermore, our approach can easily be applied to diverse data sources, even more than two, with and without supplementary patients' information.

We established a new data analysis pipeline for the joint visualization of microarray expression and arrayCGH data (aCGH), and the corresponding categorical patients' information. All computational analysis steps are programmed using R and Bioconductor. The pipeline comprises four parts: (a) data discretization, (b) binary mapping, (c) gene filtering, (d) multiple correspondence analysis. The first two steps transform data to a common binary format, a necessary step for jointly analyzing them. Filtering removes noise and redundancy by reducing the number of features (genes). We considered variance filtering, expression-aCGH correlation filtering and PCA loading on the first two principal components.
In the last pipeline step, we apply a method based on correspondence analysis, namely multivariate correspondence analysis with supplementary variables (MCASV) [3]. MCASV has been applied in the context of social sciences but to our knowledge has not been used in the context of biological high throughput data analysis. Features (expression and aCGH) and covariates (patients' information) are transformed into a common space. Vicinity between features and covariates can then be visualized and quantified. We e.g. determine genes that are correlated with covariates, possibly for interesting subsets of patients. In MCASV vicinity is measured by the angle intercurring between covariate and feature.

We applied our approach to a published dataset on breast cancer. Pollack et al. [4] studied genomic DNA copy number alterations and mRNA levels in primary human breast tumors. We were able to retrieve candidate genes that show strong association with grade 3 tumors and p53 mutant status. Candidate genes display significant enrichment of cancer related GO terms. Moreover there are interesting differences between genes selected starting from aCGH and expression data alone and genes selected by integrating the datasets.

1. Jacobs, S. et al. Genome-wide, high-resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using microarrays. *Cancer Res* **67**, 2544-2551 (2007).
2. Berger, J.A., Hautaniemi, S., Mitra, S.K. & Astola, J. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Trans Comput Biol Bioinform* **3**, 2-16 (2006).
3. Nenadic, O. & Greenacre, M. Multiple Correspondence Analysis and Related Methods. (Chapman & Hall/CRC, London; 2006).
4. Pollack, J.R. et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* **99**, 12963-12968 (2002).