Unsupervised joint analysis of arrayCGH, gene expression data and supplementary features

Steinhoff C[(1)], Pardo M[(1,2)], Vingron M[(1)]

[(1)] Max Planck Institute for Molecular Genetics, Berlin
[(2)] SENSOR Laboratory, CNR-INFM, Brescia

Motivation

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 and then look for differentially expressed (onco)genes inside these regions. 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 analysisare preferable. Only recently, few integrative methods have been published. Nevertheless, these approaches do not integrate covariate data like tumor grading, staging, age, 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.

Methods

We established a new data analysis pipeline for the joint visualization of microarray expression and arrayCGH data (aCGH), and the corresponding categorical patients' information. This pipeline comprises four parts, detailed below:
(a) data discretization,
(b) binary mapping,
(c) gene filtering,
(d) multiple correspondence analysis with supplementary variables.
All computational analysis steps are programmed using R and Bioconductor. (a) We propose three different approaches for the discretization of expression data: Probability of Expression, POE, ordinary fold change and DNAcopy. The different discretization procedures each focus on a different biological objective. For arrayCGH we use standard discretization with DNAcopy. (b) Discretized expression and arrayCGH data, and categorical supplementary data are mapped into a binary space by transforming each of the three data matrices to its corresponding indicator matrix. (c) For many applications it is customary to remove noise and redundancy from omics data by reducing the number of features (genes). We considered variance filtering, expression-aCGH correlation filtering and PCA loading on the first two principal components. (d) In the last step, we apply a method based on correspondence analysis, namely multivariate correspondence analysis with supplementary variables (MCASV). 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.

Results

We applied our approach to a dataset by Pollack et al. (2002), where they studied genomic DNA copy number alterations and mRNA levels in primary human breast tumors. Our results confirm common knowledge on breast cancer, namely that ERBB2 amplification is clearly associated with p53 status mutant; that elevated expression of ERBB2 can be associated with advanced tumors, most prevalently with ER (estrogen receptor) status minus; and finally that MYC ove-

rexpression and amplification are strongly related and show association with advanced tumors (while the contrary, underexpression and loss, are rather associated with not very advanced tumors). Furthermore, we were able to retrieve new candidate genes that show strong association with ER status, tumor grade and p53 mutant status. Candidate genes display significant enrichment of cancer related GO terms. Moreover, there are interesting differences between genes selected from aCGH and expression data alone and genes selected by integrating the datasets. For example, for genes related to ER status negative, only by integrating the two data types we get (with correlation filtering):
i) genes with a strong genomic localization (on chromosome 17) and
ii) significant enrichment of cancer related GO terms “cell substrate adhesion“, „cell matrix adhesion“ and „integrin mediated signaling pathway”. With variance filtering we get the GO term ‘regulation of apoptosis’.

Contact : steinhof@molgen.mpg.de