Integration methods for microarray data.

Risso D[(1)], Bisognin A[(2)], Romualdi C[(2)].

[(1)] Department of Statistical Sciences, University of Padua, via C. Battisti 241, 35121 Padova
[(2)] Department of Biology, University of Padua, via U. Bassi 58/B, 35121 Padova

Motivation

The increasing amount of publicly available gene expression datasets gives the opportunity to combine information arising from independent studies. In this context, meta-analysis is a challenging field for microarray data analysis. Two key issues, in conducting a meta-analysis on microarray datasets, are
i) the integration, through appropriate data normalisation, of expression levels – which increases sample size and efficiency of inference – and
ii) the combination of statistical tests using an a posteriori approach. For what concerns the first issue, recent studies show that different datasets, even if derived from the same technology (e.g. Affymetrix GeneChip), cannot be merged at a raw level. Unfortunately, analysis of variance like models, that would allow the inclusion of e.g. laboratory effect, cannot be used because of the small sample size. Often, one chooses only one or two experiments from a series of different datasets. In this context lab effect cannot be estimated: data transformation aiming at reducing this effect could represent a valid alternative.

Methods

We applied different data transformations, suitable for single channel arrays, in order to merge data from different studies. We compared these techniques using either descriptive or inferential approaches. As far as concerns descriptive analyses, we observe
i) sample groups resulting from a cluster analysis and
ii) gene expression profiles correlation among different studies after transformation. For the inferential approach we consider a cross-validation knn algorithm to study the ability of the normalized integrated datasets to predict new observations.

Results

Three datasets pertaining to breast cancer (with relapse events used for the inferential step) (GSE1456, GSE3494, GSE7390) have been used for the comparative evaluation of the selected transformations. First of all we show that, even if derived from the same platform (Affymetrix GeneChip known to be highly reproducible), the integration of raw expression data (with rma performed independently for each dataset) is not efficient, and that, on the contrary, data transformation reduces the inter-study variability increasing the correlation among studies. Secondly we show that data transformation is highly dependent from data distribution, making difficult the identification of the best technique.

Contact : davide@stat.unipd.it