Molecular signatures from gene expression data: algorithm development using R

David Casado and Ramón Díaz-Uriarte
Bioinformatics Unit, CNIO (Spanish National Cancer Centre)
Melchor Fernández Almagro 3
28029 Madrid
Spain.

April, 2004

Abstract

"Molecular signatures" or "gene-expression signatures" are used to model patients' clinically relevant information (e.g., prognosis, survival time) using expression data from coexpressed genes. Signatures are a key feature in cancer research because they can provide insight into biological mechanisms and have potential diagnostic use. However, available methods to search for signatures fail to address key requirements of signatures and signature components, especially the discovery of tightly coexpressed sets of genes.

We implement a method with good predictive performance that follows from the biologically relevant features of signatures. After identifying a seed gene with good predictive abilities, we search for a group of genes that is highly correlated with the seed gene, shows tight coexpression, and has good predictive abilities; this set of genes is reduced to a signature component using Principal Components Analysis. The process is repeated until no further component is found. Finally, to assess the stability of the obtained results, the bootstrap is used: biological interpretability is suspect if there is little overlap in the results from different bootstrap samples.

All the coding of the algorithm and its comparison with alternative approaches has been done using R and several packages available from CRAN (class —part of the VR bundle—, e1071) and Bioconductor (multtest), with a tiny bit of C++ code dynamically loaded into R. Right now, the predictive methods used include KNN and DLDA but, by using R, use of other methods is straightforward. This research and its code (released under the GNU GPL) is an example of the "turn ideas into software, quickly and faithfully" (Chambers, 1998, *"Programming with data"*) that is allowed by the S/R family of languages.