

C12

A comparative study of the stability of gene expression profiles built by univariate or multivariate methods and of their dependence on the choice of training data

M Zucknick, E A Stronach, H Gabra, S Richardson

Imperial College, London, United Kingdom

One application of gene expression arrays is to derive molecular profiles which discriminate well between two classes of tumour samples. To be interpretable, these profiles should be reproducible which is assessed by their performance on validation data independent of the training samples used to build the profile. Microarray data contain thousands of gene variables, but usually only a few dozen samples. Consequently, the selection of genes into a signature can be influenced by which samples are included in the training data. The question, how stable signatures are, when the training data are varied, is of great interest.

Molecular profiles are often built by selecting genes based on a ranked univariate measure like correlation between gene and response. An alternative is to use multivariate sparse penalised regression, e.g. the lasso. Contrary to the univariate approach, lasso regression will select only one representative gene out of a group of highly correlated genes, favouring a more diverse selection of discriminatory genes.

Both approaches require the “tuning” of a parameter, which is achieved by independent cross-validation. For the univariate technique this parameter is the number of genes to be selected, and for penalised regression it is a penalty parameter controlling the number of variables entering the multivariate model.

We compare the stability of signatures derived from both methods for well-established cancer gene expression datasets: The data are repeatedly randomly split into training and validation data and the agreement between resulting signatures is assessed. Our results show the benefit of using a multivariate method.