Discrimination and scoring using small sets of genes for two-sample microarray data

Comparison of gene expression for two groups of individuals form an important subclass of microarray experiments. We study multivariate procedures, in particular use of Hotelling's T-2 for discrimination between the groups with a special emphasis on methods based on few genes only. We apply the methods to data from an experiment with a group of atopic dermatitis patients compared with a control group. We also compare our methodology to other recently proposed methods on publicly available datasets. It is found that (i) use of several genes gives a much improved discrimination of the groups as compared to one gene only, (ii) the genes that play the most important role in the multivariate analysis are not necessarily those that rank first in univariate comparisons of the groups, (iii) Linear Discriminant Analysis carried out with sets of 2-5 genes selected according to their Hotelling T-2 give results comparable to state-of-the-art methods using many more genes, a feature of our method which might be crucial in clinical applications. Finding groups of genes that together give optimal multivariate discrimination (given the size of the group) can identify crucial pathways and networks of genes responsible for a disease. The computer code that we developed to make computations is available as an R package. (c) 2006 Elsevier Inc. All rights reserved.