

GRAPHICAL EXPLORATION OF GENE EXPRESSION DATA USING CANONICAL PROJECTION METHODS AND GE-BIPLOTS

Cameron Hurst¹ and Janet Chaseling²

¹*Institute of Health and Biomedical Innovation, Queensland University of Technology*

²*Griffith School of Environment, Griffith University*

The development of DNA microarray technology represents a major breakthrough in the life sciences. It allows the examination of expression profiles of large numbers of genes simultaneously helping us understand gene function and interactions, and genetically characterize disease providing insights into their underlying pathology.

The data arising from microarray experiments typically involve tens of thousands of genes measured on a limited number of individuals. Analyses exploring associations and groupings of genes are mostly performed using methods of unsupervised learning such as k-means clustering, hierarchical clustering and self organizing maps. These methods are limited in that they are highly dependent on the clustering technique used and the number of clusters they employ or produce can be seen as subjective. Perhaps their biggest limitation however, is that most clustering techniques classify either genes or biological samples alone, and therefore provide little insight into gene-sample associations.

Recently the use of unconstrained projection methods has been proposed for the analysis of microarray data. These methods are well established in the multivariate statistics literature, and in the context of microarray analysis, typically involve projecting both genes and samples onto a low dimensional space using a gene expression (GE-) biplot. Where gene expression difference among individuals is dominated by between-class differences (e.g. disease types), such an approach allows the simultaneously differentiation among different groups of observations and identification of candidate genes differentially expressed among these groups.

In cases where between-group differences are not strongly associated with between individual (subject) differences however, unconstrained projection may actually mask between-group expression differences. In this research the use of canonical (or constrained) projection methods is demonstrated. These techniques involve projecting both genes and samples onto space representing optimal between-group variation helping identify gene expressions most different among *a priori* defined groupings of subjects. We demonstrate the use of a number of flexible canonical projection techniques and compare the results from these techniques with currently used methods (clustering and unconstrained projection methods).