

# Data to Diamonds: Multivariate Datamining Leads to Concise Gene Signatures for Disease Classification

CSIRO Bioinformatics

In recent years there has been significant growth in the data volumes generated expression experiments. The advent of high throughput *gene chips* (Affymetrix, cDNA, . . .), has led to the capture of massive amounts of information in a single experiment. Currently, experimenters can measure the expression levels for 5,000 to 30,000 genes for each sample. Protein expression technology is not far behind.

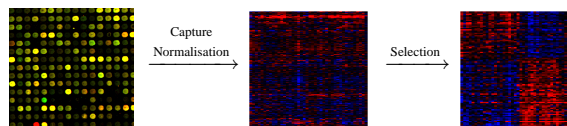
CSIRO Bioinformatics has developed a data analysis technique which rapidly sifts through very large numbers of expression measurements to identify the genes that form the best predictive set.

## Data volumes, data problems

Of course, access to such large quantities of data raises interesting new questions on how it should be analysed.

1. In many cases, scientists are interested simply in finding relationships among the samples or genes measured, with no particular target property in mind. This *unsupervised* approach has led to many interesting studies based on various forms of cluster analysis (hierarchical, k-means, self organising maps).
2. In other cases, there may be a *design* amongst the samples, and the experimenter is interested in which and how genes are affected by the differences in samples; for example, cell cycle experiments where the samples are taken through time.
3. A third form of analysis has a target property of the samples in mind and is interested in which genes might be used to predict or explain that property. Examples are; finding the genes whose expression levels diagnose the presence of a disease, predicting the outcome of a particular treatment given gene expression, and predicting the survival time of a particular patient with known gene expression.

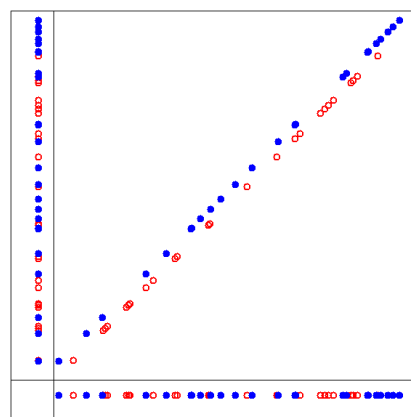
In most cases of this latter type of analysis, the majority of the genes are not relevant. It is for this *supervised* data analysis problem that CSIRO Bioinformatics has designed the *Gene-Rave* methodology.



*cDNA microarray images become gene expression levels, which are then further selected.*

## Gene-Rave — an integrated analysis

Most supervised methods require some form of preselection of genes, and this is usually achieved by considering the association with the target of each gene in turn. This form of preselection can have serious drawbacks in that *ad hoc* methods often need to be used, depending on the nature of the target. Further, if genes are relevant only in combination, gene-by-gene selection can miss them. *Gene-Rave* overcomes this problem by integrating the modelling of the target and gene selection into a single process.



*Two genes separate classes, but neither does on its own.*

*Gene-Rave* models target sample properties using a *Generalised Linear Models* framework. This family of models encompasses:

- two and multiclass classification (eg. disease or not disease, disease subtype, relapse or remission) using logistic and multinomial regression;
- continuous numeric regression targets (eg. minimum residual disease, LC50) using Gaussian, Poisson or

Gamma regression;

- censored survival targets (eg. months of survival from diagnosis) using Cox's proportional hazards regression.

Gene selection is achieved by use of Bayesian prior built into the model. The prior formalises the idea that most genes have zero weight in the model, and as the model fitting process proceeds genes, are very rapidly eliminated from the model.



*The plots above show model weight for more than 4,000 genes, at four different iterations. Gene-Rave rapidly eliminates genes from the model.*

## Advantages

The parsimonious models constructed in this way

- usually contain very small sets of genes that have predictive performance equal to or better than much larger sets identified by existing techniques.
- In addition, the technology is extremely fast required only a few minutes to analyse a few hundred arrays with more than 12,000 genes on standard PC hardware.
- This makes it possible to analyse very large datasets, and use computer intensive statistical validation techniques, such as cross validation and permutation testing, to verify the results.

## Verification

**Cross validation** is used to get estimates of prediction error. It involves dividing the data samples into  $v$  groups. Each group is removed from the data and the model fit to the remaining  $v - 1$  groups. The model

is then used to predict the left-out group. This process is repeated for each of the groups and the overall prediction error assessed.

**Permutation tests** are used to assess the significance of a model. By randomly permuting the target labels or values of the samples and rebuilding the model the likelihood of achieving a result by chance can be evaluated.

## CSIRO Bioinformatics

CSIRO Bioinformatics is a wholly owned subsidiary of the CSIRO, one of the worlds largest research organisations. CSIRO Bioinformatics has statisticians and software engineers with extensive experience in data mining in areas as diverse as remote sensing, insurance and medical instrumentation.

CSIRO Bioinformatics aims to provide statistical consulting and specialised software development services to the global biotechnology industry and to collaborate with leading international research groups.

## Contact

For further information contact,

Glenn Stone, PhD, AStat  
CSIRO Bioinformatics

Locked Bag 17  
North Ryde, NSW 1670  
Australia

Tel: +61 2 9325 3259, Fax: +61 2 9325 3200

Glenn.Stone@csiro.au

www.bioinformatics.csiro.au

