Gene Expression Data

There is a need for efficient methods to cope with the rapidly increasing amounts of biomedical and genomic sequence data.

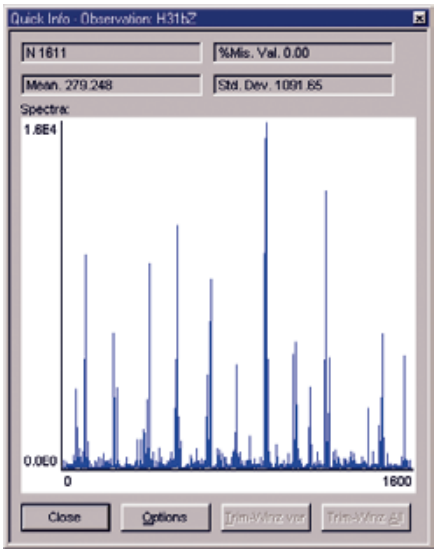
Multivariate projection techniques are ideally suited to this task to provide a graphical overview to find trends, outliers and patterns within the data.

Bioinformatics data can be divided into two categories, Analytical and Sequence based. In analytical bioinformatics, gene-chip arrays are used to identify changes in gene expression resulting from a treatment or associated with a disease state. Other data may come from LC-MS protein profiling or 2D gel electrophoresis. Multivariate techniques can rapidly pinpoint gene expression changes and anomalies.

Analysis of Gene-Chip data

Gene chips consist of short DNA strands (oligonucleotides) bound to a substrate. A fluorescently labelled genetic sample is applied to the chip and any strands of DNA in the sample with a complementary sequence will bind to the oligonucleotide probe. The binding is then indicated by fluorescence.

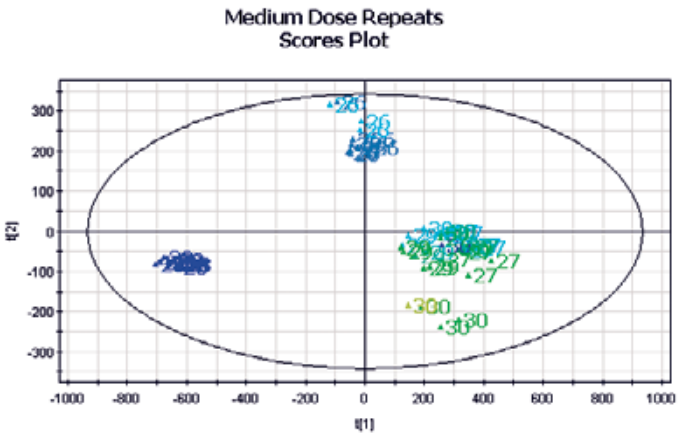
From a chemometric viewpoint, the data from gene chip arrays have the same characteristics and problems as conventional spectroscopy. The data sets are very large and often show large systematic variations. Pre-processing of the data is often necessary and the unique signal correction and scaling techniques within the SIMCA-P software may need to be applied.



"Spectrum" of gene expression shown with "Quick Info" Plot.

Data Overview

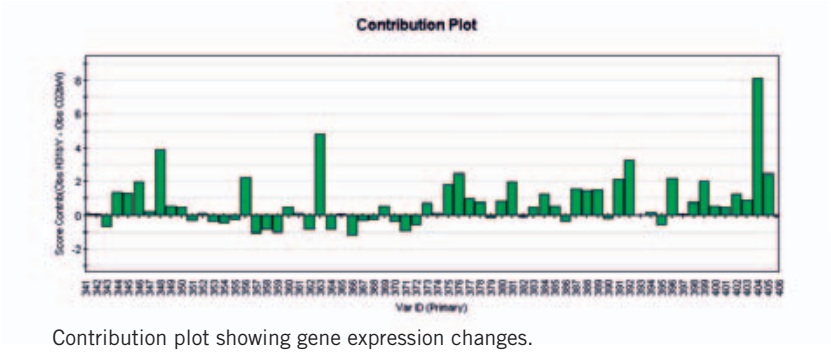
Principal Component Analysis (PCA) is very useful to get an overview of the data to determine genetic differences between treatments and that arising from experimental variation.



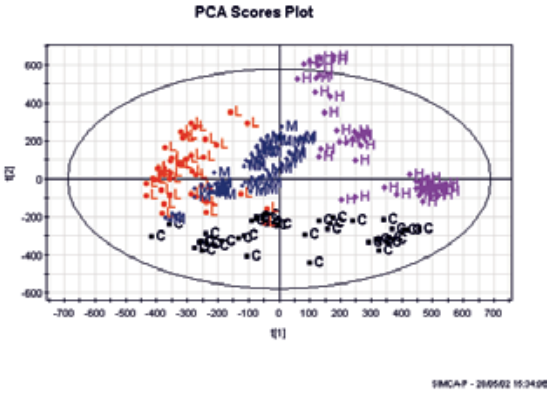
Plot showing experimental variation within a treatment group.

Discrimination between treatment groups

Principal Component Analysis (PCA) may be used to observe clustering within the data. The gene changes associated with these groupings can be seen using contribution plots.



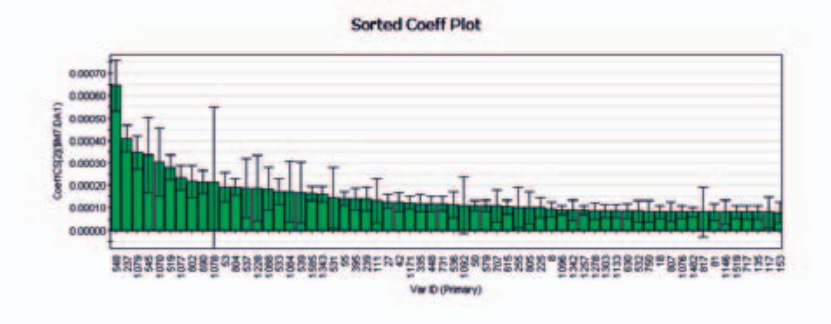
Contribution plot showing gene expression changes.

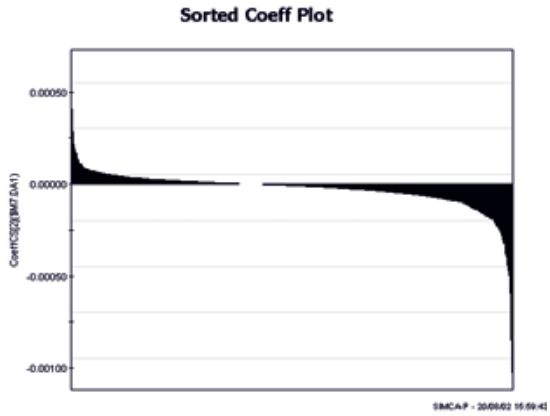


Plot showing differences between treatment groups High (purple), Medium (blue), Low (red) doses and Control (black).

Which genes have changed?

Partial Least Squares Discriminant Analysis (PLS-DA) can be used to observe changes in gene expression between groups and rapidly pinpoint which genes have been regulated up or down.





A comparison between high and low dosage groups is achieved using PLS-DA. The sorted Coefficient Plot shows the genes which have been up or down regulated.

Conclusions

- Analytical Bioinformatics data can be analysed by PCA and PLS-DA within SIMCA-P.
- PCA gives an overview of the data and highlights experimental variations and outliers.
- PCA contribution plots and PLS – Discriminant analysis can be used to determine the differences in gene expression between treatment groups.

SIMCA-P is our state-of-the-art "point and click" software for multivariate modelling and analysis. Huge data sets are quickly reduced to a few informative graphs.

UMETRICS

www.umetrics.com

In all countries:
Umetrics AB
Box 7960
SE-90719 Umeå
Sweden
Phone: +46 (0)90 184800
Fax: +46 (0)90 184899
Email: info@umetrics.com

In USA and Canada:
Umetrics Inc
17 Kiel Ave
Kinnelon, NJ 07405
USA
Phone: +1 973 492 8355
Fax: +1 973 492 8359
Email: info.us@umetrics.com

In UK:
Umetrics UK Ltd
Woodside House
Winkfield, Windsor
Berkshire, SL4 2DX, UK
Phone: +44 (0)1344 885605
Fax: +44 (0)1344 885410
Email: info.uk@umetrics.com