



Using biplots to interpret gene expression patterns in plants

Scott Chapman¹, Peer Schenk², Kemal Kazan¹ and John Manners¹

¹CSIRO Plant Industry, Long Pocket Laboratories, 120 Meiers Rd, Indooroopilly 4068, Australia and ²Cooperative Research Centre for Tropical Plant Protection, University of Queensland, Brisbane 4072, Australia

Received on May 22, 2001; revised on August 17, 2001; accepted on August 20, 2001

ABSTRACT

Summary: Plant biologists in fields of ecology, evolution, genetics and breeding frequently use multivariate methods. This paper illustrates Principal Component Analysis (PCA) and Gabriel's biplot as applied to microarray expression data from plant pathology experiments.

Availability: An example program in the publicly distributed statistical language *R* is available from the web site (www.tpp.uq.edu.au) and by e-mail from the contact.

Contact: scott.chapman@csiro.au

While cluster analysis is a popular method to assign genes to groups of similar gene expression (Eisen *et al.*, 1998; Celis *et al.*, 2000), only recently have Singular Value Decomposition (SVD) based methods (Eckart and Young, 1936) been applied to gene expression data (e.g. Alter *et al.*, 2000). Surprisingly, in presenting such results, Gabriel's biplot (e.g. Gabriel, 1971; Gabriel and Odoroff, 1990), has not yet been utilized in genomics. The biplot is the simultaneous interpretation of both rows (observations = genes) and columns (variables = treatments) of a reduced data matrix and is particularly useful where large numbers of rows and or columns are present.

SVD is one data-reduction technique to compute the principal components of a data matrix. The objective of Principal Component Analysis (PCA) is to determine a new coordinate system such that the 1st coordinate explains the maximal amount of variance in the data, while successive coordinates (totalling number of columns—1) explain maximal variance whilst being orthogonal to the first. The amount of information retained is given in the percentage of total variance explained by each successive component. Since this data transformation results in the first two or three components explaining the majority of the variance in the data, these components can be plotted in a familiar scatterplot style. The simultaneous plotting of both the row and column reduced information is known as a biplot, and together, the PCA and biplot enable

simple presentation of complex multidimensional data. Clustering the gene responses can also complement the PCA and biplot so that the variation in response within and between clusters can be visualized.

Schenk *et al.* (2000) subjected *Arabidopsis thaliana* leaves to either inoculation with the incompatible necrotrophic fungal pathogen *Alternaria brassicicola* (LOC) or treatments with defence regulators Salicylate (SA), Methyljasmonate (MJ) or Ethylene (ETH). The microarray study reported the Expressed Sequence Tags (ESTs) or genes that were substantially up or down regulated against control plants. We combined their data set on four treatments with unpublished data on a fifth treatment Systemic Disease Response (SYS), retaining 384 genes with at least 3-fold up- or down-regulation in at least one treatment and analyzed it with a script to utilize library functions in *R*, a freely available statistical computing and graphics environment (<http://cran.r-project.org/>).

Standardization of each column (treatment) requires subtraction of the column mean from each value followed by division of each value by the column standard deviation, so that the column has a mean of zero and a unit variance. In this data transformation, the SVD/PCA is the same as another method of PCA—computing eigenvalues from a correlation matrix of the data.

Gene responses over treatments were clustered into groups using another *R* library function for hierarchical clustering using squared Euclidean distance (Pythagorean distance) as the distance measure and Ward's method as the fusion criterion (Ward, 1963). The gene (row) by treatment (column) matrix of expression ratios was log transformed to base 2, column standardized, and analyzed by SVD to do PCA and determine scores for the gene effects and loadings that describe the relative contributions of the treatment effects in the decomposition of the data matrix.

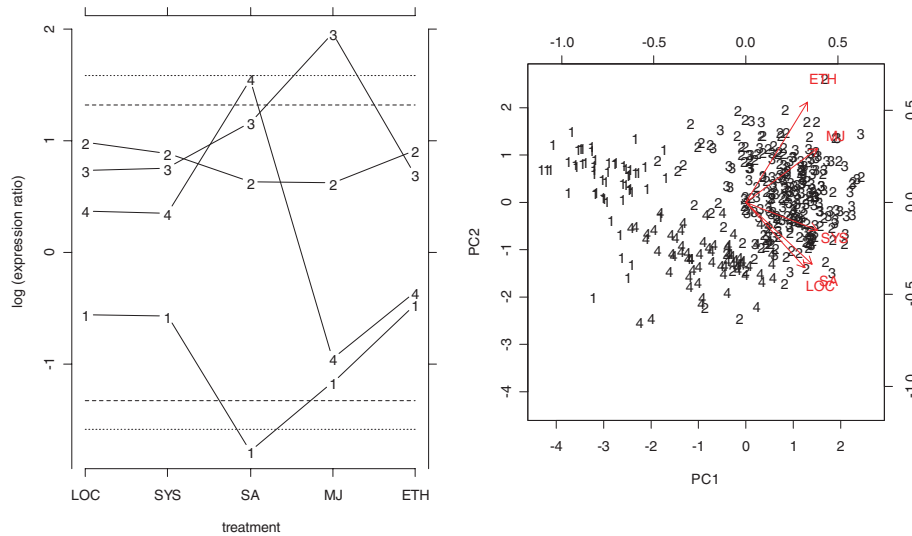


Fig. 1. (a) The mean responses (note scale is log base 2) of four gene clusters to five different treatments related to plant defence (with the dashed and dotted lines representing 2.5- and 3-fold up/downregulation, respectively) and (b) a biplot of the gene scores and treatment vectors. In both figures, the symbols represent genes belonging to the same cluster within a group of four.

BIPLOTS SHOW RELATIONSHIPS AMONG ALL OF THE GENES AND TREATMENTS

Figure 1a displays the mean responses of four gene cluster groups (accounting for approximately 80% of the variation among the 384 genes) in each treatment. The mean response of genes in groups 2 and 3 is upregulation in all treatments, while the mean response in group 1 is downregulation in all treatments. Genes in group 4 are upregulated in treatments LOC, SYS and SA, and downregulated in treatments MJ and ETH. Frequently, Figure 1a would be displayed with colours for different expression levels and alongside complementary cluster dendrograms (not presented here).

Complete explanations of the biplot interpretation, given different transformations of the data matrix, can be found elsewhere (e.g. Gabriel, 1971; DeLacy *et al.*, 1996; Gower and Hand, 1996). In Figure 1b, the 1st principal component (*x*-axis) was well correlated with the mean response of genes (symbols), explaining 48% of the variation. All treatments (vectors, with the data point being centred under the labels) have similar *x*-axis values and are positive. As the product of the gene score and the treatment vector determines any gene/treatment combination, genes to the right of Figure 1b tend to be upregulated (+ve for score by +ve for vector) in all treatments (e.g. group 3), while genes to the left of centre tend to be downregulated (-ve by +ve) in all treatments (e.g. group 1). Hence, the biplot shows the same information as the cluster means of groups 1 and 3 (Figure 1a), while retaining much of the information about

individual gene responses and simultaneously displaying the global expression effects.

In the 2nd axis (explaining an additional 17% of total variation), this particular example contrasts gene responses to different treatments. As for the 1st axis, treatment vectors with similar scores (and small angles between the vectors) have generated a similar pattern of responses among all of the genes taken together, e.g. ETH and MJ treatments with positive scores on the 2nd axis are well correlated with each other, while SA, LOC and SYS treatments that have negative scores on the 2nd axis are also well correlated with each other. The corollary is that treatments that are at approximately 90° to each other (e.g. ETH cf. SA, SYS or LOC) are poorly correlated in terms of the global patterns of gene response. Genes close to each other relative to the vectors have similar responses across the treatments, which can be confirmed by inspection of the mean cluster values in Figure 1a, e.g. in group 4, genes (with negative scores for the *y*-axis) were, on average, downregulated in ETH and MJ treatments, but were upregulated in SYS, SA or LOC. This same information can be read from the biplot in Figure 1b by noting that the group 4 genes were on the negative side of the origin with respect to the ETH or MJ vectors. In a practical sense, this set of genes (group 4), or a subset, could be considered as candidates to discriminate between unknown disease responses driven by these two different putative pathways (ETH, MJ) versus SA (Figures 1a and b).

With experience, the biplot can be rapidly interpreted

to explain the contrasts among many treatments and the relative values of the gene responses to these treatments for the entire experiment in a robust assessment of global gene response. For this analysis, we examined only highly up- or downregulated genes, though the entire dataset can be analyzed by PCA. The PCA and biplot can combine data over experiments to assist in both gene discovery, and the comparative molecular profiling of array data (including tissues, genotypes and experimental treatments). Understanding these types of displays will enable biologists to integrate multivariate methods with underlying theories of gene regulation and expression.

REFERENCES

- Alter, O. et al. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci.*, **97**, 10 101–10 106.
- Celis et al. (2000) Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett.*, **480**, 2–16.
- DeLacy, I.H. et al. (1996) Analysis of multi-environment trials—an historical perspective. In Cooper, M. and Hammer, G.L. (eds), *Plant Adaptation and Crop Improvement*. CABI, pp. 39–124.
- Eckart, C. and Young, G. (1936) The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211–218.
- Eisen, M. et al. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci.*, **95**, 14 863–14 868.
- Gabriel, K.R. (1971) The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, **58**, 453–467.
- Gabriel, K.R. and Odoroff, C.L. (1990) Biplots in biomedical research. *Stat. Med.*, **9**, 469–485.
- Gower, J.C. and Hand, D.J. (1996) *Biplots*. Chapman and Hall, London, pp. 277.
- Schenk, P. et al. (2000) Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. *Proc. Natl Acad. Sci.*, **97**, 11 655–11 660.
- Ward, J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.