

Gene expression

MADE4: an R package for multivariate analysis of gene expression data

Aedín C. Culhane^{1,*}, Jean Thioulouse², Guy Perrière² and Desmond G. Higgins¹¹Bioinformatics, Conway Institute, University College Dublin, Dublin 4, Ireland and ²Laboratoire de Biométrie et Biologie Évolutive, Université Claude Bernard, Lyon 1, 43, bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

Received on February 9, 2005; revised and accepted on March 16, 2005

Advance Access publication March 29, 2005

ABSTRACT

Summary: MADE4, microarray **ade4**, is a software package that facilitates multivariate analysis of microarray gene-expression data. MADE4 accepts a wide variety of gene-expression data formats. MADE4 takes advantage of the extensive multivariate statistical and graphical functions in the R package **ade4**, extending these for application to microarray data. In addition, MADE4 provides new graphical and visualization tools that aid in interpretation of multivariate analysis of microarray data.

Availability: The R package MADE4 is available from Bioconductor <http://bioinf.vcd.ie/software> and from Bioconductor <http://www.bioconductor.org>

Contact: aedin.culhane@ucd.ie

Supplementary information: MADE4 is well documented. There are tutorials, in the form of vignettes, which describe typical analyses. In addition, the MADE4 manual provides descriptions and examples for each function.

1 INTRODUCTION

The aim in writing microarray **ade4** (MADE4) was to provide a simple-to-use tool for multivariate analysis of microarray data. Multivariate approaches have been applied very successfully in the analysis of microarray data. Principal component analysis (PCA) has been shown to be useful in exploratory analysis of linear trends in data (Raychaudhuri *et al.*, 2000; Crescenzi and Giuliani, 2001). Fellenberg *et al.* (2001) described the application of correspondence analysis to study the association between microarray samples and genes in a reduced dimensional space. A group ordination approach was applied successfully to classification and class prediction of microarray samples (Culhane *et al.*, 2002). More recently, Culhane *et al.* (2003) employed a two-table coupling method (co-inertia analysis, CIA) to examine covariant gene-expression patterns between microarray datasets from different platforms.

Although PCA is available in several R packages, including **stats** and **amap**, the R package **ade4** contains many additional multivariate statistical methods including methods for analysis of one-data matrix, coupling of two-data matrices or multi-table analysis, http://cran.univ-lyon1.fr/doc/Rnews/4Rnews_2004-1.pdf (Thioulouse *et al.*, 1997; Chessel *et al.*, 2004). These latter methods for integrating multiple datasets make this particular package

very attractive for analysis of microarray data. MADE4 is developed as an extension to **ade4** to facilitate input and analysis of microarray data. In order to provide this functionality, MADE4 is integrated with Bioconductor (Gentleman *et al.*, 2004), probably the most popular microarray analysis software, which contains numerous packages for preprocessing, normalization, gene filtering and analysis of microarray data.

2 DATA INPUT

MADE4 accepts a wide variety of gene-expression data input formats, including Bioconductor **AffyBatch**, **exprSet**, **marrayRaw**, and standard R matrix formats (**data.frame** or **matrix**). MADE4 will automatically recognize these data formats, and no additional data processing is required.

3 MULTIVARIATE ANALYSIS

The function **ord** simplifies running ordination methods such as principal component, correspondence or non-symmetric correspondence analysis. It provides a wrapper which calls each of these methods in **ade4**.

```
results.coa <- ord(data, type = 'coa')
```

4 BETWEEN GROUPS ANALYSIS (BGA)

Between-group analysis (BGA) is a supervised classification method (Culhane *et al.*, 2002). The basis of BGA is to ordinate the groups rather than the individual samples. In tests on two microarray gene-expression datasets, BGA performed comparably to a range of supervised classification methods, including support vector machines and artificial neural networks (Culhane *et al.*, 2002). An attractive feature of BGA is that it is not limited by the large number of genes relative to the number of samples typical of microarray data. BGA of a dataset can be performed using the function **bga**. The projection of test data on BGA axes can be assessed using the function **suppl**. Leave-one-out cross validation can be performed using **bga.jackknife**.

```
results.bga <- bga(data, classvector)
```

5 CO-INERTIA ANALYSIS (CIA)

Co-inertia analysis (CIA) has been applied to the cross-platform comparison of microarray gene-expression datasets (Culhane *et al.*,

*To whom correspondence should be addressed.

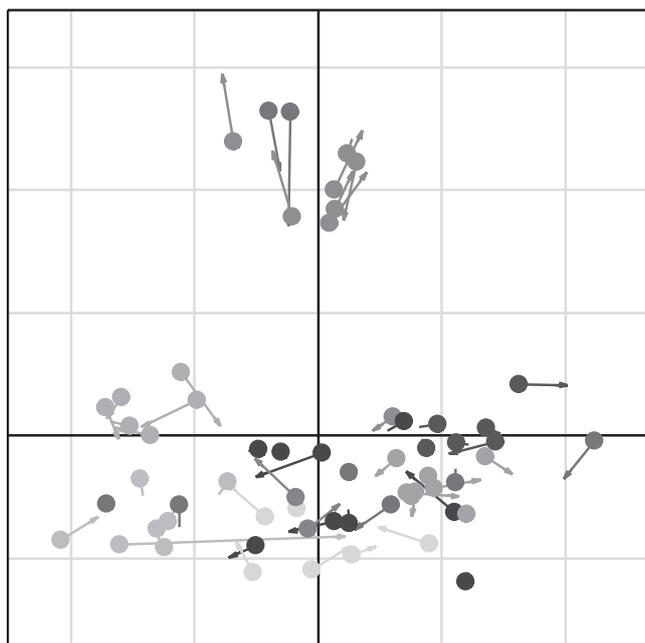


Fig. 1. Visualization of results of a CIA of two microarray datasets. Gene expression in 60 cell lines were assessed using Affymetrix and spotted microarrays, and the covariance between these was examined using CIA. The plot shows the projection of the two sets of microarray samples. Each sample of the 60 cell lines is represented by two co-ordinates, the spotted microarray (arrow) and Affymetrix space (closed circle), which are joined by a line. The length of the line is proportional to the divergence between the two samples. See Culhane *et al.* (2003) for more details.

2003). CIA is a multivariate method that identifies trends or co-relationships in multiple datasets which contain the same cases or variables. That is, either the rows or the columns of a matrix must be 'matchable'. CIA can be applied to datasets where the number of variables (genes) far exceeds the number of samples (arrays) (Fig. 1).

```
results.cia<-cia(dataset1, dataset2)
```

6 VISUALISATION OF RESULTS

There are many functions in MADE4 to visualise results. The simplest way to view results produced by `ord`, `bga` or `cia` is to use `plot`. Microarray samples (or genes) can be colour coded if a vector of class membership is given.

In addition, there are functions for drawing 1D and 3D plots. For example, the function `html3D` produces output which can

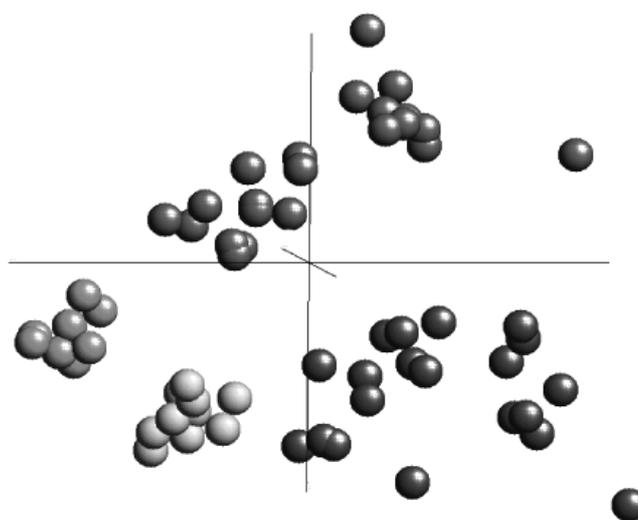


Fig. 2. Visualization of three axes of an ordination of microarray data. The 3D plot was produced using `html3D` and `Rasmol`.

be visualized using `jmol`, `Rasmol` or `chime` (Fig. 2), providing a free and very useful interface for colouring, rotating, zooming and manipulating 3D graphs.

ACKNOWLEDGEMENTS

We would like to acknowledge the assistance of Dr Florent Baty, Ian Jeffery and Ailís Fagan in testing development versions of MADE4.

REFERENCES

- Chessel,D. *et al.* (2004) The ade4 package—I: One-table methods. *Rnews*.
- Crescenzi,M. and Giuliani,A. (2001) The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data. *FEBS Lett.*, **507**, 114–118.
- Culhane,A.C. *et al.* (2002) Between-group analysis of microarray data. *Bioinformatics*, **18**, 1600–1608.
- Culhane,A.C. *et al.* (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, **4**, 59.
- Fellenberg,K. *et al.* (2001) Correspondence analysis applied to microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 10781–10786.
- Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
- Raychaudhuri,S. *et al.* (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*, 455–466.
- Thioulouse,J. *et al.* (1997) ADE-4: a multivariate analysis and graphical display software. *Stat. Comput.*, **7**, 75–83.