A Multivariate Approach to Integrating Datasets using made4 and ade4

by Aedín C. Culhane and Jean Thioulouse

The public microarray repositories, ArrayExpress and the GeneExpression Omnibus (GEO), now contain over 100,000 microarray gene expression profiles (Table 1). This is a considerable data resource.

However the average number of arrays per study is only between 30 and 40 (Table 1). Given that the number of features (genes) on microarrays now exceeds 50,000, this presents a considerable dimensionality problem. Low case to feature ratio is likely to remain an issue, as cost and availability of biomaterial, such as biopsy tissue, are often limiting. As a result, meta-analysis or merging data from multiple studies is attractive.

Table 1. Public Microarray Databases[a]

Database	Arrays	Studies
ArrayExpress[b]	44,602	1,487
GEO[c]	87,073	2,353

[a]Statistics: ArrayExpress (June 2006), GEO (5 July 2006)
[b]http://www.ebi.ac.uk/arrayexpress/
[c]http://www.ncbi.nlm.nih.gov/geo/

Unfortunately, matching of variables (gene probes) from different microarray technologies is challenging. Numerous microarray platforms have been developed and a number of studies have reported disappointingly low correlations between different technologies. Matching of probes by their DNA sequence reduces cross-platform inconsistency (Carter et al., 2005), and functions to perform sequence matching are available in the Bioconductor package **matchprobes**. EnsEMBL alignments of DNA probe sequences to the human and other genomes can be retrieved using the package **biomaRt**. However even the performance of matched probes may vary across platforms. These differences may be due to real biological effects where probes on different platforms detect different splice variants or homologues of a gene.

A different approach is simply to examine genes or cases with covariant trends across matched datasets. We have described the application of co-inertia analysis (CIA) for visualization and analysis of such trends across microarray datasets (Culhane et al., 2003). Functions to perform these analyses are provided in the Bioconductor package **made4** (Culhane et al., 2005). **made4** is an extension to **ade4** (Thioulouse et al., 1997; Chessel et al., 2004), an extensive R package for multivariate analysis of ecological data.

Our multivariate approach for cross-platform analysis of microarray data may be easily applied to heterogeneous datasets. Increasingly, microarray experiments are performed in parallel with proteomics, metabolomics or other high throughput array technologies. Typically the identity of peaks or spots in proteomics or metabolomics data is unknown. Therefore mapping probes, spots or peaks across datasets is not possible. In analysis of these data, we are simply exploring features (peaks or spots) that have similar trends across datasets and are correlated with a covariate of interest. CIA is suitable for such an analysis.

We will describe the application of CIA to cross-platform visualization of microarray data and other functions in **made4** and **ade4** for multivariate analysis of biological datasets.

Co-inertia analysis

CIA is a multivariate analysis method that describes the relationship between two data tables (Dray et al., 2003). It can be used on quantitative, qualitative or distance matrices. Classical methods, like principal component (PCA) or correspondence analysis (CA), aim at summarizing a table by searching orthogonal axes on which the projection of the sampling points (cases) have the highest possible variance. This characteristic ensures that the associated graphs (factor maps) best represent the initial data (see Figure 1).

To extract information common to two tables, canonical analysis (CANCOR, Gittins, 1985) searches successive pairs of axes (one for each table) with a maximum correlation. The problem is that this analysis may lead to axes with high correlation, but low percentages of explained variance. This means that it will be difficult to give a biological interpretation to these axes. To overcome this difficulty, CIA searches for pairs of axes with maximum covariance (instead of correlation). This ensures that CIA axes will have both a high correlation and also good percentages of explained variance for each table. Computations are based on the cross table between the variables of the two tables. The importance of each axis is given by the percentage of total co-inertia, which is similar to the percentage of explained variance for canonical axes.

CIA has been successfully applied to visualization of cross-platform relationships between microarray datasets (Culhane et al., 2003). CIA is an attractive approach as it can be applied to data where the number of variables (genes) far exceeds the number of cases, as seen in microarray data. Given data with such low sample size, CANCOR cannot be used

and canonical correspondence analysis (Ter Braak, 1986) is reduced to a plain CA (Dray et al., 2003).

Monte-Carlo tests can be used to check the significance of the relationship between the two tables. The method consists of performing many random permutation of the cases (arrays), followed by the recomputation of the total co-inertia. By comparing the total co-inertia obtained in the normal analysis with the co-inertias obtained after randomization, one can estimate the probability of the observed relationship between the tables.

PCA or CA of microarray data

PCA and CA are well suited to exploratory analysis of microarray data, and complement popular clustering approaches. While clustering investigates pairwise distances among objects highlighting fine relationships, PCA and CA examine the variance of the whole dataset highlighting general trends and gradients (reviewed by Brazma and Culhane, 2005). To perform a PCA or CA on a microarray dataset using **made4**, use the function **ord**.

To illustrate we apply CA to a microarray gene expression profiling study of 4 childhood tumors (NB, BL-NHL, EWS, RMS; Khan et al., 2001). A subset of these expression data (khan$train, 306 genes x 64 cases), a factor describing the class of each case (khan$train.classes, length=64) and a data frame of gene annotation are available in dataset khan in **made4**.

```
library(made4)
data(khan)
dataset = khan$train
fac =     khan$train.classes
geneSym = khan$annotation$Symbol

results.coa <- ord(dataset, type="coa")
par(mfrow= c(1,2))
plotarrays(results.coa, classvec=fac)
plotgenes(results.coa, genelabels= geneSym)
```
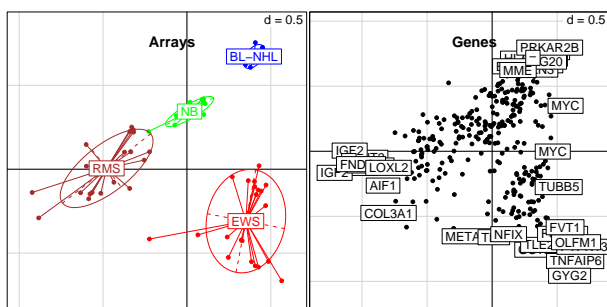

Figure 1: CA of a 306 gene subset of Khan dataset (Khan et al., 2001). **A)** Plot of arrays **B)** Plot of genes. The further a gene and case are projected in the same direction from the origin, the stronger association between that gene and case (gene is upregulated in that array sample).

Cross-platform analysis using CIA

To perform CIA, objects in the dataset must be "matchable". For example, where multiple studies are performed on the same samples, CIA can detect co-varying patterns across datasets. If the cases are matched, there is no constraint to match the variables (genes) and the number of variables in each dataset may differ.

In Example 2, we examine a panel of 60 cell lines from the National Cancer Institute (NCI60) that have been subjected to gene expression profiling using Affymetrix (Staunton et al., 2001) and spotted cDNA (Ross et al., 2000) arrays. We apply **cia** to subsets of these 2 datasets which are available in **made4**.

```
data(NCI60)
names(NCI60)
  [1] "Ross"    "Affy"    "classes" "Annot"
fac = NCI60$classes[,2]
results.cia = cia(NCI60$Affy, NCI60$Ross)
par(mfrow=c(1,2))
plotarrays(results.cia, clabel=0)
plotarrays(results.cia, clabel=0,
          classvec=fac)
```

In Figure 2, matched cases are joined by a line. If two cases (arrays) have similar profiles, they will be projected close together. Therefore, the shorter the length of connecting line the greater the correlation. In Figure 2B, one green case is represented by a long line, indicating a large cross-platform difference between the two expression profiles for this cell line. This may suggest a quality issue in one dataset.
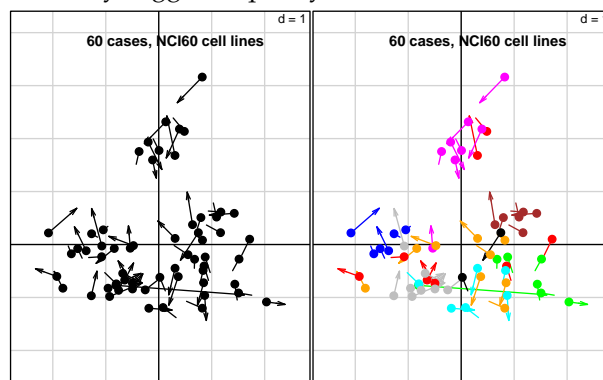

Figure 2: CIA of NCI60 datasets Affy (closed circles) and Ross (arrows). **B).** Same as A) but cases are colored by class (cancer cell line phenotype). Further details and interpretation in Culhane (2003).

CIA using "matched" genes

Equally CIA could be performed on variables (genes). Visualization of matched genes across platforms is often useful when there is a one:many match

of gene probes. On older arrays, a gene was generally only represented by one probe, but on recent microarrays a gene maybe represented by 5 or more probes (or probesets). In Example 3, we examine microarray studies of acute lymphoblastic leukemia (ALL) using older hu6800 (Golub et al., 1999) or more recent u95av2 (Chiaretti et al., 2004) Affymetrix arrays. These datasets are available in Bioconductor packages **golubEsets** and **ALL**.

```
library(affy)
library(ALL)
data(ALL)
ALL.fac <- substring(ALL$BT,1,1)
library(golubEsets)
data(Golub_Train)
golub <- Golub_Train[,1:27]  #ALL data
golub.data <- exprs(golub)
#footnote 1
golub.data[] <- as.double(golub.data)
golub.fac <-golub$T.B.cell
```

We performed a *t*-test on the Golub data, using **rowttests** in the **genefilter** package, to select genes which were significantly (P<0.001) associated with T-cell or B-cell ALL.

```
library(genefilter)
ttests <- rowttests(golub.data,golub.fac)
nsignf <- sum(ttests$p.val < 0.001)
topGeneInd <- order(ttests$p.val)[1:nsignf]
ttests.signf <-
    rownames(golub.data)[topGeneInd]
```

There were 109 significant gene probes on the hu6800 arrays, which were matched to genes on the u95av2 using **biomaRt**.

```
library(biomaRt)
mart <- useMart("ensembl", mysql=TRUE)
mart <- useDataset("hsapiens_gene_ensembl",
                   mart)
pRef <- getBM(attributes="affy_hg_u95av2",
              values=ttests.signf,
              filters="affy_hugenefl",
              mart=mart)
anyNA <- function(x) any(is.na(x))
pRef <- pRef[!apply(pRef, 1, anyNA), ]
dupSet <- function(x, a)
    subset(x, a %in% a[duplicated(a)])
pMany <- dupSet(pRef, pRef$affy_hugenefl)
```

Of the 109 hu6800 probesets, 96 mapped to 133 u95av2 probesets. Therefore 29 hu6800 probesets mapped to more than 1 u95av2 probesets. These 29:66 "one to many" matches were examined using **cia**.

```
hu6800set <-
    exprs(golub[pMany$affy_hugenefl, ])
u95av2set <-
    exprs(ALL[pMany$affy_hg_u95av2, ])

cia.out <- cia(t(hu6800set), t(u95av2set))
coordVar1 <- cia.out$coinertia$co
coordVar2 <- cia.out$coinertia$li
par(mfrow=c(2, 2))
plotarrays(cia.out, sub="Genes")
plotarrays(cia.out, clabel=0,
           classvec=pMany$affy_hugenefl)
plotarrays(coordVar1, classvec=golub.fac)
plotarrays(coordVar2, classvec=ALL.fac)
```

In Figure 3 we observe that the probesets selected using the older hu6800 platform, do appear to discriminate B and T cells expression profiles on u95av2 arrays. However it appears that only a few gene probes contribute a significant amount of variance across both datasets.
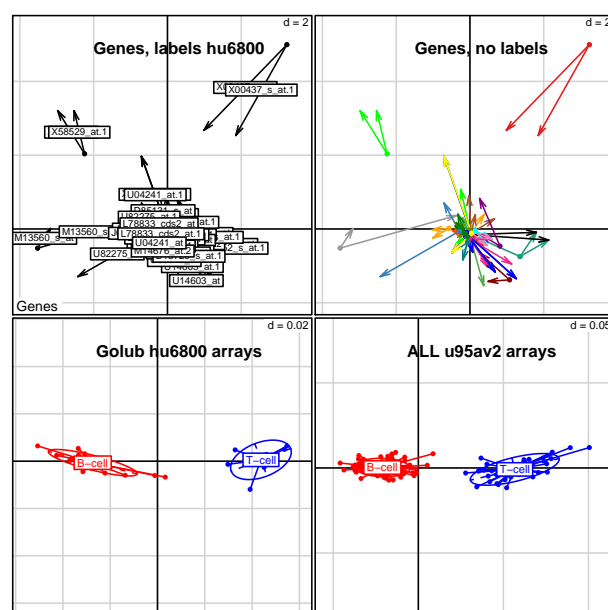

Figure 3: CIA of a set of genes in golub and ALL datasets. Projection of probesets **A)** with and **B)** without hu6800 probe labels, and arrays **C),D)**.

One nice feature of this analysis is that one:many probeset matches are clearly visualised. For example, M13560_s_at, X58529_at, X00437_s_at have 2 matches on the hgu95av2 platform. We can see that only one M13560_s_at u95av2 probeset matches has a high loading on the B-cell end of axis 1 (horizontal), indicating that the expression of only one of these probesets agrees with the older hu6800 array.

Combining microarray data with gene sequence information

We have also used CIA to integrate microarray data with counts of motif occurrences in gene promoters, to discover which promoter motifs are most associated with the main patterns of gene expression in a dataset (Jeffery et al., 2007). We have also extended this approach using between group analysis (Culhane et al., 2002), a supervised method where groupings of arrays or tissues of a-priori interest are contrasted with the rest. Using between group CIA, we identify gene motifs (or other gene features) that are most associated with a gene expression classifier (Jeffery et al., 2007).

Using **ade4** codon usage may be investigated using internal correspondence analysis, a variant of between groups and within groups analyses (Lobry and Chessel, 2003). CIA has also been applied to study the relationships between amino-acid physico-chemical properties and protein composition (Thioulouse and Lobry, 1995). These analyses maybe facilitated using the **seqinr** package (Charif et al., 2005). **seqinr** is an interface between R and the ACNUC (Gouy et al., 1985) sequence retrieval system for nucleotide and protein sequence databases such as GenBank, EMBL and SWISS-PROT.

Although we have only described analysis of 2 tables, many other multivariate analysis methods are available in **ade4**, which are easily extended using the duality diagram (class **dudi**) (Chessel et al., 2004). There are several functions for analysis of three-way or multiple tables (class **ktab** class, and functions **sepan**, **statis**, **pta**, **mcoa**, **mfa**, **foucart**). Distance matrices can be integrated in this framework through principal coordinates analysis (**dudi.pco** function), and the **kdist** class in the case of k distance matrices measured on the same individuals.

GUIs : ade4TkGUI, Rweb

The **made4** package was created to ease the use of multivariate data analysis of microarray gene-expression data. Indeed, it has two main advantages: it is an interface between the **ade4** package and Bioconductor data objects and classes, and it provides wrapper functions to simplify the use of multivariate analysis functions implemented in **ade4**.

Another approach to the simplification is the use of a graphical user interface (GUI). A new package (**ade4TkGUI**) has been developed using the **tcltk** package to provide a GUI to **ade4**. This GUI has two special features. The first one is a centralized graphical display of **ade4 dudi** objects). The second one is a dynamic view of factor maps, allowing exploration of sample and variable sets by way of zooming, panning, and searching on labels. An Rweb interface to **seqinr** and **ade4** multivariate analysis is also available (`http://pbil.univ-lyon1.fr/Rweb/Rweb.general.html`).

Summary

The Bioconductor package **made4** facilitates multivariate analysis of microarray data, and builds on extensive experience of multivariate data analysis in ecology. Multivariate data analysis methods provide many useful tools to extract meaningful biological information from these large data sets. Sometimes, these methods are overlooked because they are thought to be complicated and subject to barely met application hypotheses. This is partly true in the framework of the Gaussian approximation model of multivariate analysis. But the geometric model (for example Le Roux and Rouanet, 2004), and the duality diagram (Holmes, 2006) lift most of these assumptions.

Bibliography

A. Brazma and A.C. Culhane. Algorithms for gene expression analysis. In M.J. Dunn, L.B. Jorde, P.F.R. Little, and S. Subramaniam, editors, *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. John Wiley and Sons, London, 2005.

S.L. Carter, A.C. Eklund, B.H. Mecham *et al.* Redefinition of affymetrix probe sets by sequence overlap with cdna microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics*, 6: 107, 2005.

D. Charif, J. Thioulouse, J.R. Lobry *et al.* Online synonymous codon usage analyses with the ade4 and seqinr packages. *Bioinformatics*, 21(4):545–7, 2005.

D. Chessel, A. Dufour, and J. Thioulouse. The ADE4 package – I: One-table methods. *RNews*, 4(1):5–10, June 2004.

S. Chiaretti, X. Li, R. Gentleman *et al.* Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–8, 2004.

A.C. Culhane, G. Perrière, E. C. Considine *et al.* Between-group analysis of microarray data. *Bioinformatics*, 18(12):1600–8, 2002.

A.C. Culhane, G. Perrière, and D.G. Higgins. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4:59, 2003.

A.C. Culhane, J. Thioulouse, G. Perrière *et al.* Made4: an R package for multivariate analysis of gene expression data. *Bioinformatics*, 21(11):2789–90, 2005.

S. Dray, D. Chessel, and J. Thioulouse. Co-inertia analysis and the linking of ecological tables. *Ecology*, 84:3078–3089, 2003.

R. Gittins. *Canonical analysis, a review with applications in ecology. Vol.12 of Biomathematics.* Springer- Verlag, Berlin, 1985.

T.R. Golub, D.K. Slonim, P. Tamayo *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

M. Gouy, C. Gautier, M. Attimonelli *et al.* ACNUC–a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Comput Appl Biosci*, 1(3):167–72, 1985.

S. Holmes. Multivariate analysis: The french way. In D. Nolan and T. Speed, editors, *Festschrift for David Freedman*. IMS, Beachwood, OH, 2006.

I. B. Jeffery, S. F. Madden, P. A. McGettigan *et al.* Integrating transcription factor binding site information with gene expression datasets. *Bioinformatics*, 23(3):298–305, 2007.

J. Khan, J.S. Wei, M. Ringner *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, 7(6):673–9, 2001.

B. Le Roux and H. Rouanet. *Geometric Data Analysis.* Kluwer Academic Publishers, Dordrecht, 2004.

J. R. Lobry and D. Chessel. Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *J Appl Genet*, 44(2):235–61, 2003.

D.T. Ross, U. Scherf, M.B. Eisen *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*, 24(3):227–35, 2000.

J. E. Staunton, D.K. Slonim, H.A. Coller *et al.* Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci USA*, 98(19):10787–92, 2001.

C. Ter Braak. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 69:1167–1179, 1986.

J. Thioulouse and J. Lobry. Co-inertia analysis of amino-acid physico-chemical properties and protein composition with the ade package. *Comput Appl Biosci*, 11(3):321–9, 1995.

J. Thioulouse, D. Chessel, S. Dolèdec *et al.* ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing*, 7(1):75–83, 1997.

Aedín C. Culhane
Department of Biostatistics and Computational Biology,
Dana-Farber Cancer Institute & Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA.
aedin@jimmy.harvard.edu

Jean Thioulouse
Biométrie et Biologie Evolutive,
CNRS & Université Lyon 1, France.
jthioulouse@biomserv.univ-lyon1.fr

Using amap and ctc Packages for Huge Clustering

by Antoine Lucas and Sylvain Jasson

Introduction

Huge clustering is often required in the field of DNA microarray (DeRisi et al., 1997) analysis. A new use of clustering results appears with presentation and exploration software like *TreeView* (Eisen et al., 1998).

DNA microarray is the most appropriate method for high throughput gene studies, allowing expression evaluation of vast gene numbers in different cells types or conditions. From a technical point of view, microarray analysis first needs image processing (for example *Imagene* (http://www.biodiscovery.com), *BZScan* (Lopez et al., 2004) or *ScanAlyze* (Eisen et al., 1998)) that gives large tables of data, followed by statistical processing including data normalization.

A main goal of microarray analysis is to detect co-regulated genes presenting similar expression profiles, which can be achieved by various classification techniques. In this area, hierarchical clustering is of special interest as it allows multi-scale cluster visualization.

Some R extensions provide efficient clustering tools (mainly: **stats** and **cluster**; Struyf et al., 1997). The packages **amap** and **ctc** aim to complete the set of clustering tools for R with:

- Additional features to standard clustering functions.

- A novel PCA method, robust to extreme values.