BIOINFORMATICS

Vol. 00 no. 00 2009
Pages 1–2

Supplemental data of integrOmics: an R package to unravel relationships between two omics data sets

Kim-Anh Lê Cao [1]*, Ignacio González [2] and Sébastien Déjean [3]

[1]The University of Queensland, Institute for Molecular Biosciences and ARC Centre of Excellence in Bioinformatics, Brisbane QLD 4072 Australia, [2]Plateforme Biopuces, Genopôle Toulouse Midi-Pyrénées, Institut National des Sciences Appliquées, F-31077 France, [3]Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse et CNRS, F-31062 France.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: With the availability of many 'omics' data, such as transcriptomics, proteomics or metabolomics, the integrative or joint analysis of multiple datasets from different technology platforms is becoming crucial to unravel the relationships between different biological functional levels. However, the development of such an analysis is a major computational and technical challenge as most approaches suffer from high data dimensionality. New methodologies need to be developed and validated.

Results: `integrOmics` efficiently performs integrative analyses of two types of 'omics' variables that are measured on the same samples. It includes a regularized version of Canonical Correlation Analysis to enlighten correlations between two data sets, and a sparse version of Partial Least Square regression that includes simultaneous variable selection in both data sets. The usefulness of both approaches has been demonstrated previously and successfully applied in various integrative studies.

Availability: `integrOmics` is freely available from http://CRAN.R-project.org/ or from the website companion (http://math.univ-toulouse.fr/biostat) that provides full documentation and tutorials.

Contact: k.lecao@uq.edu.au

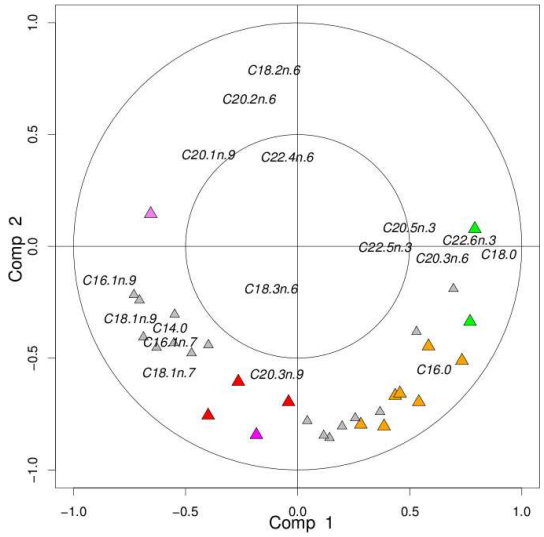

Fig. 1. Examples of variable visualization output with `integrOmics`.

VISUALIZATION OUTPUTS

Representation of the variables

Graphical representation of the variables enable a better understanding of the correlation structure between the two data sets.

The (selected) variables can be represented by projecting them on correlation circles (González *et al.*, 2008), as shown in Fig. 1. On this graphic the coordinates of the variables are obtained by computing the Pearson correlations between the X and Y variables and the score vectors for the different dimensions of rCCA or sPLS. As both type of variables are of variance 1, their projections on the plane defined by the two score vectors are inside a circle of radius 1, which is centered at the origin (*correlation circle*). In Fig. 1 two circles of radius 0.5 and 1 were plotted to reveal the correlation structure of the variables. Variables highly correlated (co-regulated) will be projected in the same direction and close to the circle of radius 1. Fig. 1 simultaneously represents fatty acids and transcripts from the `nutrimouse` data set included in the package. Triangles of the same color represent genes that belong to the same metabolic pathway (orange = fatty acid catabolism, red = lipid biosynthesis, purple = xenobiotic metabolism and green = detoxification, see González *et al.* (2009) for more details). As a result, variables which are highly correlated (fatty acids and transcripts) or co-regulated (transcripts only or fatty acids only) are projected in the same direction and cluster together. This type of graphic gives complementary information to the network graph illustrated in the main paper.

```
> data(nutrimouse)
> X = nutrimouse$gene
> Y = nutrimouse$lipid
> nutri.spls <- spls(X, Y, ncomp = 2, keepX =
+   c(15, 15), keepY = c(12,12))
> plotVar(nutri.spls, keep.var = TRUE)
```

*to whom correspondence should be addressed

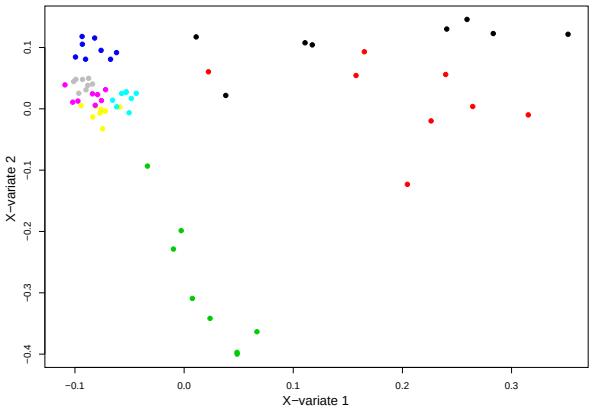

Fig. 2. Examples of sample visualization output with `integrOmics`.

Representation of the samples

Scatter plots of the score (latent) vectors from the first dimensions allow the user to identify similarities between the samples. Often, these similarities (clusters of samples) were found to have a biological meaning (González *et al.*, 2009; Lê Cao *et al.*, 2009). Fig. 2 displays the sPLS latent vectors computed on the `liver.toxicity` data included in the package. `liver.toxicity` consists of $p = 3116$ gene expressions and $q = 10$ clinical measurements on $n = 64$ rats that were exposed to different doses of acetaminophen. The colors represent the different doses and times of necropsies: green, red, black and blue for the times 48h, 24h, 18h and 6h (for both 1500 mg/kg and 2000 mg/kg lethal doses), yellow, purple, turquoise and gray for the times 48h, 24h, 18h and 6h (for both 50 mg/kg and 150 mg/kg low doses).

```
> data(liver.toxicity)
> X <- liver.toxicity$gene
> Y <- liver.toxicity$clinic
> toxicity.spls <- spls(X, Y, ncomp = 2,
+   keepX = c(50, 50), keepY = c(10, 10))
> plotIndiv(toxicity.spls)
```

Dimensions 1 and 2 of the sPLS highlight very well the different experimental conditions between the samples as well as their similarities: dimension 1 tends to separate the moderately to severely toxic doses of acetaminophen whereas dimension 2 separates the exposure time (6h, 18h, 24h vs. 48h). This graphic combined with the variable representations gives a valuable insight into the biological study, see Lê Cao *et al.* (2009) for more details.

REFERENCES

González, I., Déjean, S., Martin, P. G. P., and Baccini, A. (2008). CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software*, **23**(12).

González, I., Déjean, S., Martin, P. G. P., Gonçalves, O., Besse, P., and Baccini, A. (2009). Highlighting relationships between heteregeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, **17**(2).

Lê Cao, K.-A., Martin, P., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, **10**(34).