Unravelling 'omics' data with the R package mixOmics

Kim-Anh Lê Cao, Sébastien Déjean, Ignacio González

Unravelling 'omics' data with the R package `mixOmics`

K-A. Lê Cao[a], **I. González**[b] and **S. Déjean**[b]

[a]Queensland Facility for Advanced Bioinformatics
University of Queensland
4072 St Lucia, QLD, Australia
k.lecao@uq.edu.au

[b]Institut de Mathématiques
Université de Toulouse et CNRS
UMR 5219, F-31062 Toulouse, France
ignacio.gonzalez@math.univ-toulouse.fr
sebastien.dejean@math.univ-toulouse.fr

Mots clefs : multivariate statistics, data integration, high-throughput biological data.

Recent advances in high throughput 'omics' technologies enable quantitative measurements of expression or abundance of biological molecules of a whole biological system. The transcriptome, proteome and metabolome are dynamic entities, with the presence, abundance and function of each transcript, protein and metabolite being critically dependent on its temporal and spatial location.

With `mixOmics`, we are currently establishing a global analytical framework to extract relevant information from high throughput 'omics' platforms such as genomics, proteomics and metabolomics. Specifically, the statistical methodologies developed and implemented in the R package focus on the so-called multivariate projection-based approaches, which can handle such large data sets, deal with multicollinearity and missing values. These methodologies enable dimension reduction by projecting these large data sets into a smaller subspace, to capture the largest sources of variation in the biological studies. These techniques enable exploration, visualisation of the data and lead to biological insights.

Principal Component Analysis (PCA) is a commonly used dimension reduction technique to highlight expression patterns that might be due to biological variation, or systematic platform bias in a single data set. Recently, we have proposed another variant based on Independent Component Analysis (IPCA) [1]. By applying Lasso penalisation [2] on the PCA or IPCA components, we further reduce the dimension of the data by selecting the relevant information (the measured biological entities) related to the biological study. Both approaches are unsupervised, i.e. the focus is to identify the genes, proteins or metabolites with similar information without taking into account experimental knowledge on class labels of the samples. A supervised approach was also developed based on Partial Least Square Discriminant Analysis (PLS-DA) [3] to select discriminative biological entities across several groups of samples.

Whilst single omics analyses are commonly performed to detect between-groups difference from either static or dynamic experiments, the integration or combination of multi-layer information is required to fully unravel the complexities of a biological system. Data integration relies on the currently accepted biological assumption that each functional level is related to each other.

Therefore, considering all the biological entities (transcripts, proteins, metabolites) as part of a whole biological system is crucial to unravel the complexity of living organisms.

To that purpose, we have developed integrative approaches, such as regularized Canonical Correlation Analysis (rCCA) [4], sparse PLS [5,6] to highlight or understand the relationship between two types of biological entities. We have demonstrated on several biological studies that this integrative analyses of large scale omics datasets could generate new knowledge not accessible by the analysis of a single data type alone.

All methodologies are implemented in `mixOmics` along with S3 methods for an easy use of the package and an easy interpretation via graphical tools [6]. Our website gives more information about the methodologies and how to use the package (http://www.math.univ-toulouse.fr/~biostat/mixOmics/). For the non R specialist, a web application was also developed and made available to the research community (http://mixomics.qfab.org).
In this presentation, I will cover the recent developments of `mixOmics`, illustrate the use of the methodologies to various biological studies and demonstrate the usefulness of the graphical tools to give biological meaning to the obtained results.

Références

[1] Yao, F., Coquery J. and Lê Cao K.-A. (2012). Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinformatics*, 13:24.
[2] Tibshirani, R. (2007). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1): 267-288.
[3] Lê Cao K.-A., Boitard, S. and Besse, P. (2011). Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12:253.
[4] González I., Déjean S., Martin P.G.P., Gonąlves 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), pp 173-199.
[5] Lê Cao K.-A., Rossouw, D., Robert-Granié C., Besse, P. (2008). A Sparse PLS for Variable Selection when Integrating Omics data. *Statistical Applications in Genetics and Molecular Biology*: Vol. 7 : Iss. 1, Article 35.
[6] Lê Cao K.-A., Martin P.G.P, Robert-Granié C., Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* 10: 34.
[7] Lê Cao K.-A., González I. and Déjean S (2009). integrOmics: an R package to unravel relationships between two omics data sets. *Bioinformatics* **25**(21):2855-2856. *Note: the package has since been renamed mixOmics.*