Multivariate analyses of biological 'omics data using the R package mixOmics

The objective of this workshop is to introduce the fundamental concepts of multivariate dimension reduction methodologies. Those methods are particularly useful for data exploration and integration of large data sets, and especially in the context of systems biology, or in research areas where statistical data integration is required. Each methodology that will be presented during the course will be applied on biological "omics" studies including transcriptomics, metabolomics and proteomics data sets using the R package mixOmics (http://mixomics.org/).

Organized and sponsored by National Polytechnic Institute of Toulouse (INP)

Dates 9-10 November 2017, 9-5pm

Practical information The workshop is free of charge for all participants. Priority will be given to INP students, external postgraduate students and early career researcher. The workshop includes tuition, course material. The workshop excludes tea/coffee and lunch during the breaks

Location: INP Toulouse, France (TBA)
Maximum number of participants external to INP: 15

Registration Fill in the survey at: https://www.surveymonkey.com/r/XTDHYRY Registration close on **27th October**.

Contact mixomics@math.univ-toulouse.fr

Prerequisite and requirements

We require from the trainees **a good working knowledge in R programming** (e.g. handling data frame, perform simple calculations and display simple graphical outputs) **to fully benefit from the workshop.**
Participants are requested to **bring their own laptop**, having installed the software RStudio http://www.rstudio.com/and the R package mixOmics (instructions will be provided prior to the training).

More details on the covered topics

1. *Key methodologies in mixOmics and their variants*
    A. Exploration of one data set and how to estimate missing values
    B. Identification of biomarkers to discriminate different treatment groups
    C. Integration of two data sets and identification of biomarkers
    D. Repeated measurements design
    E. Introduction to the integration of more than two data sets
2. *Review on the graphical outputs implemented in mixOmics*
    A. Sample plot representation
    B. Variable plot representation for data integration
    C. Other useful graphical outputs
3. *Case studies and applications*

Five case studies will be analysed using the methods presented above

The following statistical concepts will be introduced: covariance and correlation, multiple linear regression, classification and prediction, cross-validation, selection of diagnostic or prognostic markers, l_1 and l_2 penalties in a regression framework.
Each methodology will be illustrated on a case study (theory and application will alternate).
If all participants are French native speakers, the workshop might be taught in the French language. Material and slides are written in English.

Note that mixOmics is not limited to biological data only and can be applied to other type of data where integration is required.

Target group

The course is intended for data analysts in the fields of bioinformatics, computational biology and applied statistics with some statistical knowledge and **a good working knowledge in R**. It will be particularly useful to those interested in:
1. Exploring large data sets.
2. Selecting features with methods implementing LASSO-based penalisations.
3. Using graphical techniques to better visualise data.
4. Understanding and/or applying multivariate projection methodologies to large data sets.

Anticipated learning outcomes

After completion of this workshop, participants will be able to
1. *Understand fundamental principles of multivariate projection-based dimension reduction technique.*
2. *Perform statistical integration and feature selection using recently developed multivariate methodologies.*
3. *Apply those methods to high throughput biological studies, including their own studies*.

Some relevant publications

- Rohart F., Gautier, B, Singh, A and Lê Cao, K. A. mixOmics: an R package for 'omics feature selection and multiple data integration. On bioRxiv (in press in PLoS Comp Biol).
- Singh A, **Gautier B**, Shannon C, Vacher M, **Rohart F**, Tebbutt S, **K-A. Lê Cao**. DIABLO – multi-omics data integration for biomarker discovery. *Manuscript available in bioRxiv*.
- K-A. Lê Cao*, ME Costello*, VA Lakis, **F Bartolo**, **XY Chua**, R Brazeilles, P Rondeau. (2016) MixMC: Multivariate insights into Microbial Communities.*PLoS ONE*11(8): e0160169 [link]
- Günther P., Shin H., Ng R.T., McMaster W. R. , McManus B. M. , Keown P. A. , Tebbutt S. J. , **Lê Cao K-A.** (2014), Novel multivariate methods for integration of genomics and proteomics data: Applications in a kidney transplant rejection study, *OMICS: A journal of integrative biology*, [link].
- **González I., Lê Cao K.-A.,** Davis, M.D. and **Déjean S.** (2013) Insightful graphical outputs to explore relationships between two 'omics' data sets. *BioData Mining* 5:19.
- **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*, 22:253. [link]

Instructors

Dr Kim-Anh Lê Cao (University of Melbourne, Brisbane Australia) was awarded her PhD in 2008 at Université de Toulouse, France. She then moved to Australia as a postdoctoral fellow at the University of Queensland, Brisbane Australia. Since the beginning of her Ph.D Kim-Anh has initiated a wide range of valuable collaborative and research opportunities in both statistics and molecular biology. Her main research focus is on variable selection for biological data (`omics' data) coming from different functional levels by the means of multivariate dimension reduction approaches. Since 2009, her team has been working on developing a statistical software dedicated to the integrative analysis of `omics' data, to help researchers make sense of biological big data.

Kim-Anh is senior lecturer at the University of Melbourne, and regularly runs statistical training workshops and short series seminars as well as mixOmics multi-day workshops (12 mixOmics workshops, totalling 20 days since 2014). More details on Kim-Anh current research projects: http://sysgen.unimelb.edu.au/research/research-groups/computational-biostatistics-methods-le-cao

Sébastien Déjean was awarded his PhD in Applied Statistics in 2002 at Université de Toulouse, France after spending 4 years in a Biometry lab at INRA (French National Institute for Agricultural Research). He then has been working at the Toulouse Mathematics Institute as a research engineer. He works in close collaboration with researchers working in different areas, such as high-throughput biology, chemistry and information retrieval. Sébastien is an expert in statistical data analysis and he contributes to the development of several R packages including mixOmics as a core member.

More details on Sébastien's current research projects: http://perso.math.univ-toulouse.fr/dejean/

Additional tutors

Dr David Rengel (data analyst, Laboratory of Plant-Microbe Interactions, National Institute for Agricultural Research)
Yannick Lippi (Mathematics and computing, ToxAlim, National Institute for Agricultural Research)

Local organizing committee

Prof Christèle Robert Granié (Génétique, Physiologie et Systèmes d'Elevage, National Institute for Agricultural Research)
Dr Sébastien Déjean (Toulouse Mathematics Institute, University of Toulouse)
Dr Cecile Bonnefont (National Polytechnic Institute of Toulouse)

More details and background

Multivariate dimension reduction approaches are useful exploratory tools to get a first understanding of large and complex data sets. These approaches are extremely efficient to compute and highly flexible: they can either fit into a regression or a classification framework. The latest developments in that exciting area of research include feature selection and statistical integration of several data sets to maximise the covariance or correlation between heterogeneous sources of data.

Principal Component Analysis (PCA) is the oldest and most popular multivariate method that uses projections to reduce data dimension. The tutorial will first introduce key concepts in multivariate dimension reduction with PCA, and then present innovative approaches for statistical data integration.

Data integration is often required in a systems biology context, when experiments are performed on the same individuals or subjects but at different molecular levels. These so-called ‘omics data (transcriptomics for the study of transcripts, proteomics for proteins, metabolomics for metabolites, etc) often come from different technology platforms or sources. Combining such heterogeneous data sets can lead to the discovery of important biological insights, provided that relevant information can be extracted while integrating those large data sets.

Our tutorial will focus on some key multivariate methodologies to select relevant and meaningful biological features during the integration process. Each methodology will be illustrated on real biological studies transcriptomics, metabolomics, proteomics data). Other types of biological data can also be analysed or integrated. We will further illustrate mixOmics analyses for 16S data (microbiome data), time course ‘omics data and meta-analyses if there is sufficient interest from the audience.