
Programme de Formation

Introduction to Machine Learning for analysis of biological data with R



Organisation

Durée : 21 heures

Mode d'organisation : Présentiel

Contenu pédagogique

Public visé

Researchers, engineers, PhD students and clinicians who need to analyze their own biological or clinical data (e.g. transcriptomics, scRNA-seq, clinical cohorts) and want to build a solid, reproducible analysis workflow in R.



Objectifs pédagogiques

Machine learning and multivariate statistics have become central to the analysis of biological and clinical data. This hands-on course offers a practical introduction to the complete analysis pipeline in R, from reproducible project setup and quality control to dimensionality reduction (PCA, UMAP), clustering (k-means, hierarchical), and regression modelling (linear and generalized linear models). The emphasis is on practical implementation and interpretation rather than mathematical formalism: participants learn to apply each method to real datasets, tune its parameters, and critically interpret the results. Examples are drawn from transcriptomics, clinical and ecological data to build transferable skills.

- Set up a reproducible R project and run analyses in an HPC environment
- Perform quality control, handle missing data, and normalize/scale biological datasets
- Apply and interpret dimensionality reduction methods (PCA, UMAP), including parameter tuning
- Implement unsupervised clustering (k-means, hierarchical) and determine the optimal number of clusters
- Build, diagnose, and interpret regression models (linear and generalized linear models)



Description

Day 1: R Essentials & Data Quality

Implement development best practices and prepare a robust dataset

- Morning: Environment and Syntax
 - Project Management: Directory structuring, reproducibility best practices, and version control



- HPC Environment: Introduction to High-Performance Computing (HPC) for executing R scripts at scale
- R refresher : Review of fundamental syntax and data structure manipulation (dataframes, plot)
- Afternoon: Data Preparation
 - Quality Control (QC): Outlier identification, missing data handling, and distribution assessment
 - Data Scaling: Normalization techniques, data transformation, and centering/scaling to prepare data for statistical models

Day 2: Dimensionality Reduction

Extract relevant signals and visualize high-dimensional datasets

- Morning: Principal Component Analysis (PCA):
 - PCA principle, implementation in R, and component interpretation (explained variance, biplots)
 - Hands-on practice
- Afternoon: PCA and non linear approach
 - PCA (Part 2): Advanced applications and hands-on practice
 - Introduction to UMAP (highly suited for omics data)

Day 3: Clustering & Linear Regression

Identify underlying structures (expression profiles) and introduce modeling

- Morning: Unsupervised Learning (Clustering)
 - K-means: Algorithm principles, determining the optimal number of clusters (elbow method, silhouette score)
 - Hierarchical Clustering: Distance metrics, linkage methods, and dendrogram interpretation
- Afternoon: Supervised Modeling
 - Clustering (Part 2): Advanced applications and hands-on practice
 - Linear Regression: Model building, assumption testing (residuals, homoscedasticity), and coefficient interpretation

Day 4: Advanced Modeling (Half-Day)

Expand statistical foundations into generalized models.

- Morning: Advanced Applications & Wrap-up
 - Advanced Regression Techniques: introduction to Generalized Linear Models (logistic regression, count models (Poisson / negative binomial))
 - Wrap-up: Q&A session and course review



Prérequis

Prior experience with R is required: participants should be comfortable importing a dataset, manipulating a data frame (filtering, subsetting), and producing basic plots. A working understanding of descriptive statistics is expected.



Modalités pédagogiques

Lectures, workshops, and case studies.



Moyens et supports pédagogiques

Ressources pédagogiques: files in PDF format will be made available to the trainee.

Equipements provided: digital resources ; to be brought by students: laptop.



Modalités d'évaluation et de suivi

Ongoing formative assessments throughout the course. A certificate of completion is issued at the end of the training.



Informations sur l'admission

Admission to this training is not subject to any examination, test or prior selection; enrolment is confirmed upon receipt of a complete application and validation by the training provider.



Informations sur l'accessibilité

Our organisation is committed to ensuring inclusive and equitable access to its training programmes, whether delivered online or in person, for all participants, including people with disabilities. A dedicated accessibility contact is available to assess individual needs and, where possible, implement appropriate pedagogical, technical and organisational adjustments.