Mini-Projet de Bioinformatique

Analyse et visualisation de données biologiques

Albert NDOUR

Juin 2025

1 Introduction

Ce document présente une série de mini-projets destinés aux étudiants en bioinformatique, couvrant diverses thématiques d'analyse de données génomiques. Chaque projet propose une introduction au domaine, les packages à utiliser, les sources de données et des suggestions d'analyses pratiques.

2 Mini-projets pour étudiants en Bioinformatique

2.1 Transcriptomique Spatiale

Titre : "Cartographie moléculaire des tissus : Analyse de données de transcriptomique spatiale"

Présentation du domaine :

— Introduction aux technologies de transcriptomique spatiale (Visium, MERFISH, Slide-seq, etc.)

— Importance de la localisation spatiale des transcrits dans la compréhension des fonctions tissulaires

Packages R/Python :

— `SpatialExperiment` et `Seurat` pour la gestion des données

— `spatialLIBD` pour l'analyse des données Visium

— `squidpy` (Python) pour l'analyse spatiale avancée

— `ggplot2` et `patchwork` pour la visualisation

Données :

— Données publiques Visium de 10X Genomics (cerveau de souris)

— Données spatialLIBD du cortex préfrontal humain

Projet pratique :

— Identification des couches corticales cérébrales par clustering des profils d'expression

— Analyse des gradients spatiaux d'expression de gènes

— Détection des structures tissulaires par analyse de voisinage spatial

Références et ressources :

— `SpatialExperiment` : https://bioconductor.org/packages/release/bioc/html/SpatialExperiment.html
— `Seurat` : https://satijalab.org/seurat/
— `spatialLIBD` : https://bioconductor.org/packages/release/bioc/html/spatialLIBD.html
— `squidpy` : https://squidpy.readthedocs.io/
— 10X Visium : https://www.10xgenomics.com/datasets (jeu de données "Mouse Brain Section")
— spatialLIBD : http://research.libd.org/spatialLIBD/index.html

2.2 Bulk RNA-seq

Titre : "Signatures transcriptomiques : Analyse différentielle et fonctionnelle de données RNA-seq"

Présentation du domaine :
— Principes du séquençage RNA-seq en vrac (bulk)
— Workflow d'analyse : contrôle qualité, alignement, quantification, analyse différentielle

Packages R :
— `DESeq2` pour l'analyse différentielle
— `edgeR` comme alternative d'analyse
— `clusterProfiler` pour l'enrichissement fonctionnel
— `EnhancedVolcano` pour la visualisation des résultats

Données :
— Données TCGA pour un type de cancer (ex : cancer du sein)
— Données de modèles murins de maladies inflammatoires

Projet pratique :
— Identification des gènes différentiellement exprimés entre conditions
— Enrichissement de voies biologiques et interprétation fonctionnelle
— Construction de signatures moléculaires prédictives

Références et ressources :
— `DESeq2` : https://bioconductor.org/packages/release/bioc/html/DESeq2.html
— `edgeR` : https://bioconductor.org/packages/release/bioc/html/edgeR.html
— `clusterProfiler` : https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
— `EnhancedVolcano` : https://bioconductor.org/packages/release/bioc/html/EnhancedVolcano.html
— TCGA : https://portal.gdc.cancer.gov/ (téléchargement direct)
— TCGAbiolinks : https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html
— GEO : https://www.ncbi.nlm.nih.gov/geo/ (GSE123813, modèle murin d'inflammation pulmonaire)

— ReCount3 : https://jhubiostatistics.shinyapps.io/recount3/

2.3 Single-cell RNA-seq

Titre : "Hétérogénéité cellulaire dévoilée : Analyse transcriptomique à l'échelle unicellulaire"

Présentation du domaine :

— Principes et technologies de séquençage unicellulaire
— Défis spécifiques : rareté des transcrits, effet de lots, dropout

Packages R/Python :

— `Seurat` (R) ou `Scanpy` (Python) pour l'analyse complète
— `sctransform` pour la normalisation
— `MAST` pour l'analyse différentielle
— `celldex` et `SingleR` pour l'annotation automatique

Données :

— Données PBMC 10X Genomics
— Atlas cellulaires publics (Tabula Muris, Human Cell Atlas)

Projet pratique :

— Clustering et identification de types cellulaires
— Reconstruction de trajectoires de différenciation avec `Monocle3` ou `Slingshot`
— Analyse d'interactions cellulaires avec `CellChat` ou `CellPhoneDB`

Références et ressources :

— `Seurat` : https://satijalab.org/seurat/
— `Scanpy` : https://scanpy.readthedocs.io/
— `sctransform` : https://github.com/ChristophH/sctransform
— `MAST` : https://github.com/RGLab/MAST
— `celldex` : https://bioconductor.org/packages/release/bioc/html/celldex.html
— `SingleR` : https://bioconductor.org/packages/release/bioc/html/SingleR.html
— `Monocle3` : https://cole-trapnell-lab.github.io/monocle3/
— `Slingshot` : https://bioconductor.org/packages/release/bioc/html/slingshot.html
— `CellChat` : https://github.com/sqjin/CellChat
— 10X Genomics PBMC : https://www.10xgenomics.com/resources/datasets
— Tabula Muris : https://tabula-muris.ds.czbiohub.org/
— Human Cell Atlas : https://data.humancellatlas.org/

2.4 Pseudo-bulk RNA-seq

Titre : "Du singulier au collectif : Analyse pseudo-bulk de données single-cell"

Présentation du domaine :
— Concept et utilité de l'approche pseudo-bulk
— Avantages par rapport aux analyses bulk et single-cell traditionnelles
Packages R :
— `muscat` pour la génération et l'analyse de données pseudo-bulk
— `aggregateBioVar` pour l'agrégation des données
— `DESeq2` pour l'analyse différentielle
Données :
— Données single-cell de populations immunitaires
— Données de cancers avec microenvironnement tumoral
Projet pratique :
— Agrégation des données unicellulaires par type cellulaire
— Analyse différentielle entre conditions pour chaque type cellulaire
— Comparaison des résultats avec les approches bulk et single-cell traditionnelles
Références et ressources :
— `muscat` : https://bioconductor.org/packages/release/bioc/html/muscat.html
— `aggregateBioVar` : https://bioconductor.org/packages/release/bioc/html/aggregateBioVar.html
— Jeu de données muscat : https://bioconductor.org/packages/release/data/experiment/html/muscData.html
— Données SingleCellExperiment : http://bioinfo.genyo.es/en/scRNAseq_datasets/

2.5 Multi-omique intégrative

Titre : "Intégration multi-omique : Analyse combinée du transcriptome, épigénome et protéome"
Présentation du domaine :
— Principes d'intégration de données multi-omiques
— Complémentarité des différentes couches d'information biologique
Packages R/Python :
— `MOFA` (Multi-Omics Factor Analysis)
— `mixOmics` pour l'intégration statistique
— `Seurat` et `Signac` pour l'intégration scRNA-seq et ATAC-seq
Données :
— Données TCGA multi-omiques (RNA-seq, méthylation, CNV)
— Données CITE-seq (RNA + protéines de surface)
Projet pratique :
— Intégration de données transcriptomiques et épigénétiques
— Identification de modules de régulation multi-omiques
— Prédiction de phénotypes à partir de signatures multi-omiques
Références et ressources :

— `MOFA` : https://biofam.github.io/MOFA2/
— `mixOmics` : https://www.bioconductor.org/packages/release/bioc/html/mixOmics.html
— `Seurat + Signac` : https://satijalab.org/signac/
— TCGA multi-omiques via TCGAbiolinks : https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html
— Jeu de données CITE-seq : https://cite-seq.com/resources/
— Dataset multi-omique : https://bioconductor.org/packages/release/data/experiment/html/curatedTCGAData.html

2.6 Déconvolution cellulaire

Titre : "Déchiffrer la complexité tissulaire : Déconvolution des profils d'expression"

Présentation du domaine :
— Principes des méthodes de déconvolution
— Applications en immunologie et oncologie

Packages R :
— `immunedeconv` pour la déconvolution immunitaire
— `CIBERSORT` et ses dérivés
— `MuSiC` pour la déconvolution basée sur des données single-cell

Données :
— Données bulk RNA-seq de tissus tumoraux
— Données de référence single-cell pour la construction de signatures

Projet pratique :
— Estimation des proportions de types cellulaires dans des échantillons tumoraux
— Comparaison de différentes méthodes de déconvolution
— Corrélation des infiltrats immunitaires avec les données cliniques

Références et ressources :
— `immunedeconv` : https://icbi-lab.github.io/immunedeconv/
— `CIBERSORT` : https://cibersort.stanford.edu/
— `MuSiC` : https://github.com/xuranw/MuSiC
— Jeux de données de référence : https://github.com/icbi-lab/immunedeconv/tree/master/data
— LM22 (CIBERSORT) : https://cibersort.stanford.edu/download.php
— ImmuCC de référence : http://202.120.223.180/immune/

2.7 Analyse de réseaux de co-expression

Titre : "Réseaux géniques : Analyse des modules de co-expression et inférence de régulateurs"

Présentation du domaine :

— Principes de la co-expression génique
— Approches d'inférence de réseaux de régulation

Packages R :

— `WGCNA` pour l'analyse de co-expression
— `CEMiTool` pour une approche simplifiée
— `GRNBoost2` (Python) pour l'inférence de réseaux

Données :

— Données longitudinales de réponse immunitaire
— Données développementales (ex : embryogenèse)

Projet pratique :

— Construction de modules de co-expression
— Identification de gènes "hub" et régulateurs clés
— Validation de modules par enrichissement fonctionnel

Références et ressources :

— `WGCNA` : https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/
— `CEMiTool` : https://bioconductor.org/packages/release/bioc/html/CEMiTool.html
— `GRNBoost2` : https://arboreto.readthedocs.io/en/latest/
— GEO GSE134515 (réponse immunitaire temporelle) : https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134515
— Données tutorielles WGCNA : https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/

3 Structure recommandée pour chaque mini-projet

1. Introduction théorique et contextualisation biologique
2. Présentation des méthodes d'analyse et packages
3. Tutoriel pratique avec code et données
4. Mini-projet avec questions de recherche spécifiques
5. Discussion des résultats et perspectives

4 Projet 1 : Analyse d'intégration multi-omique pour la médecine de précision

4.1 Description du sujet

Ce projet vise à explorer les méthodes d'intégration de différents types de données omiques (génomique, transcriptomique, protéomique) pour mieux comprendre les mécanismes de maladies complexes et identifier des biomarqueurs potentiels. Les étudiants apprendront à organiser, visualiser et analyser des ensembles de données multi-omiques pour extraire des informations biologiquement pertinentes.

4.2 Objectifs pédagogiques

— Maîtriser les concepts fondamentaux de l'intégration multi-omique
— Comprendre les défis liés à l'analyse de données omiques hétérogènes
— Appliquer des méthodes statistiques avancées pour l'intégration de données
— Interpréter les résultats dans un contexte biologique et clinique

4.3 Packages R à utiliser

— `MultiAssayExperiment` : Framework pour l'organisation et l'intégration de données multi-omiques
— `mixOmics` : Suite d'outils pour les analyses multivariées de données omiques
— `MOFA2` : Multi-Omics Factor Analysis pour intégrer plusieurs couches de données omiques
— `curatedTCGAData` : Accès facile aux données TCGA prétraitées
— `ggplot2` et `ComplexHeatmap` : Visualisation avancée des résultats

4.4 Sources de données

— The Cancer Genome Atlas (TCGA) : https://portal.gdc.cancer.gov/
— Jeu de données `curatedTCGAData` disponible via Bioconductor
— International Cancer Genome Consortium (ICGC) : https://dcc.icgc.org/

4.5 Références principales

1. Ramos, M., Schiffer, L., Re, A., Azhar, R., Basunia, A., Rodriguez, C., ... & Waldron, L. (2020). Software for the integration of multiomics data with genomic medicine : the Bioconductor package `MultiAssayExperiment`. Cancer Research, 80(16), 3268-3275.
2. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+ : a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology, 21(1), 1-17.
3. Rohart, F., Gautier, B., Singh, A., & Lê Cao, K. A. (2017). mixOmics : An R package for 'omics feature selection and multiple data integration. PLoS Computational Biology, 13(11), e1005752.

5 Projet 2 : Modélisation de la dynamique de transmission des maladies infectieuses

5.1 Description du sujet

Ce projet se concentre sur le développement et l'analyse de modèles mathématiques pour étudier la propagation de maladies infectieuses, un sujet particulièrement pertinent à l'ère post-COVID. Les étudiants exploreront différents types de modèles épidémiologiques (déterministes et stochastiques) et les appliqueront à des données réelles.

5.2 Objectifs pédagogiques

- Comprendre les fondements mathématiques des modèles épidémiologiques
- Implémenter des modèles à compartiments (SIR, SEIR, etc.) en R
- Estimer les paramètres des modèles à partir de données épidémiques réelles
- Simuler différents scénarios d'intervention et évaluer leur impact

5.3 Packages R à utiliser

- `deSolve` : Résolution d'équations différentielles ordinaires
- `EpiModel` : Framework pour la modélisation des épidémies
- `pomp` : Modélisation de processus partiellement observés via Markov
- `covidregionaldata` : Accès standardisé aux données COVID-19
- `ggplot2` et `shiny` : Visualisation et interfaces interactives

5.4 Sources de données

- Données COVID-19 de l'Organisation Mondiale de la Santé : https://covid19.who.int/data
- Johns Hopkins University COVID-19 Data Repository : https://github.com/CSSEGISandData/COVID-19
- Our World in Data COVID-19 Dataset : https://github.com/owid/covid-19-data

5.5 Références principales

1. Jenness, S. M., Goodreau, S. M., & Morris, M. (2018). EpiModel : An R package for mathematical modeling of infectious disease over networks. Journal of Statistical Software, 84(8), 1-47.
2. King, A. A., Nguyen, D., & Ionides, E. L. (2016). Statistical inference for partially observed Markov processes via the R package pomp. Journal of Statistical Software, 69(12), 1-43.

3. Funk, S., Abbott, S., Atkins, B. D., Baguelin, M., Baillie, J. K., Birrell, P., ... & Meakin, S. R. (2022). Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. Philosophical Transactions of the Royal Society B, 377(1857), 20210307.

6 Projet 3 : Analyses de survie et modèles à risques concurrents dans les études cliniques

6.1 Description du sujet

Ce projet aborde les méthodes modernes d'analyse de survie, en particulier les modèles à risques concurrents, essentiels pour évaluer correctement les résultats cliniques lorsque plusieurs événements peuvent se produire. Les étudiants apprendront à modéliser, analyser et interpréter des données de temps jusqu'à événement en présence de risques concurrents.

6.2 Objectifs pédagogiques

— Comprendre les concepts fondamentaux de l'analyse de survie
— Distinguer entre censure et événements concurrents
— Implémenter des modèles à risques concurrents et interpréter leurs résultats
— Développer des modèles prédictifs pour l'évaluation du risque

6.3 Packages R à utiliser

— `survival` : Analyses de survie standard (Kaplan-Meier, Cox)
— `cmprsk` : Analyses spécifiques aux risques concurrents
— `riskRegression` : Modélisation des prédictions de risque
— `survminer` : Visualisations avancées pour l'analyse de survie
— `pec` : Évaluation des performances prédictives

6.4 Sources de données

— Données `mgus2` (Monoclonal Gammopathy) du package `survival`
— Données de transplantation rénale du package `KMsurv`
— Framingham Heart Study via le package `survival` ou `frailtypack`
— Données cliniques de Mayo Clinic disponibles dans divers packages R

6.5 Références principales

1. Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. The Annals of Statistics, 16(3), 1141-1154.
2. Austin, P. C., Lee, D. S., & Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. Circulation, 133(6), 601-609.
3. Gerds, T. A., Kattan, M. W., Schumacher, M., & Yu, C. (2022). riskRegression : Risk Regression Models and Prediction Scores for Sur-

vival Analysis with Competing Risks. Journal of Statistical Software, 101(10), 1-31.

7 Bibliographie générale

— Gentleman R, Carey V, Huber W, et al. (2004). *Bioconductor : open software development for computational biology and bioinformatics.* Genome Biology, 5(10), R80.
— Amezquita RA, Lun AT, Becht E, et al. (2020). *Orchestrating single-cell analysis with Bioconductor.* Nature Methods, 17(2), 137-145.
— Huber W, Carey VJ, Gentleman R, et al. (2015). *Orchestrating high-throughput genomic analysis with Bioconductor.* Nature Methods, 12(2), 115-121.
— Stuart T, Satija R (2019). *Integrative single-cell analysis.* Nature Reviews Genetics, 20(5), 257-272.
— Love MI, Huber W, Anders S (2014). *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 15(12), 550.
— Langfelder P, Horvath S (2008). *WGCNA : an R package for weighted correlation network analysis.* BMC Bioinformatics, 9, 559.