Multi-Omics and Spatial Data Analysis with Factor Models

Scenario

A common analysis approach for high-dimensional omics datasets is the derivation of a low-dimensional embedding that captures the key sources of variation within the data. Such embeddings serve as a valuable foundation for various downstream analyses. With the increasing availability of datasets spanning multiple omics modalities, an essential challenge is identifying a unified data embedding that integrates information across all modalities. One powerful computational tool for this purpose is **Multi-Omics Factor Analysis (MOFA)** (Argelaguet et al., 2018), a framework that extends traditional factor analysis to handle multiple data modalities.

In this project, we will apply MOFA to a **chronic lymphocytic leukaemia (CLL)** dataset. This dataset includes ex vivo drug responses, transcriptomics, methylation, and somatic mutation data from approximately 250 patients. Using MOFA, we will derive a latent representation of the data and leverage it to predict clinical outcomes.

Additionally, we will demonstrate the application of **MEFISTO** (Velten et al., 2022)—an extension of MOFA designed for spatiotemporal data—by analysing spatial transcriptomics data from a Visium slide of the mouse brain.

Project Breakdown in Days

1. Getting started with MOFA
    a. Loading the CLL data
    b. Data preprocessing and overview
    c. Slides: Introduction to factor analysis and MOFA
    d. Discussion of MOFA input data / parameters
    e. Training a MOFA model
2. Understanding the output of MOFA
    a. Trained model properties (variance explained, factor correlation, number of active factors)
    b. Inferred factors and weights and their sparsity structure, use
        i. gene set enrichment analysis to give biological meaning to factors
        ii. correlation of factors with metadata to understand sample heterogeneity
    c. Bonus: additional analyses, prediction and imputation tasks
3. Analysing spatial (or temporal) data with MEFISTO
    a. Slides: Introduction to factor analysis with dependent observations
    b. Loading the Visium data
    c. Data preprocessing and overview
    d. Training a MEFISTO model
    e. Visualizing spatial factors
4. Extra day
    a. Time for questions
    b. Additional slides about Bayesian factor models
    c. Getting started with your own data

Expected Outcome

The participants will

- learn what factor models are and what they are useful for;
- learn how the popular factor model MOFA can be used to integrate multi-omics data;
- gain experience in using MOFA by analysing a clinical data set in depth;
- learn how the MEFISTO extension of MOFA can be used to analyse temporal and spatial data.

Datasets

CLL Dataset: Dietrich S, Oleś M, Lu J, Sellner L, Anders S, Velten B, Wu B, Hüllein J, da Silva Liberio M, Walther T (2018) Drug-perturbation-based stratification of blood cancer. *J Clin Invest* 128: 427–445
Visium Mouse Brain Dataset: https://www.10xgenomics.com/datasets
All data is available at https://github.com/florinwalter/ebi_mofa_workshop

Required Resources

- conda/mamba environment with relevant packages installed
- Jupyterlab
- If possible (not strictly required): GPU access

Detailed Project Plan

Day one – Monday
15:45 – 17:30
Slides: Project introduction
(more?)

Day two – Tuesday
14:15 – 17:30
Slides: Data handling (AnnData, MuData, CLL data set)
Tasks:

- Load data (build AnnData & MuData from csv files)
- Get data overview (views, missing values, data types)

Slides: Intuitive introduction to factor models and MOFA
Tasks:

- Train MOFA model on CLL data
- Make slides

Day three – Wednesday
16:30 – 18:00
Slides: MOFA recap, weights downstream analysis
Tasks:

- Inspect trained model properties (R2, number of factors, factor correlation, weights distribution)
- Perform gene set enrichment analysis
- Make slides

Day four – Thursday
16:30 – 18:00
Slides: MOFA recap, factor scores downstream analysis
Tasks:

- Correlate factor scores with covariates
- Bonus: prediction and imputation tasks
- Make slides

Day five – Friday
09:30 – 11:00
Wrap up, time for questions, preparation of presentation