Practical Lab Session: Gene Regulatory Network (GRN) Inference

Sergio Peignier

Abstract

This practical session guides you through the end-to-end pipeline of Gene Regulatory Network (GRN) inference. You will retrieve transcriptomic data, map transcription factors using AnimalTFDB, preprocess data using Scanpy, infer networks using different machine learning algorithms via GReNaDIne, and finally characterize network topology and functional enrichments.

1 Objectives

1. Acquire and filter transcriptomic data (bulk or single-cell) and prior knowledge databases.
2. Pre-process raw expression matrices in Python using Scanpy.
3. Perform GRN inference using multiple machine learning paradigms (Random Forests, Gradient Boosting, Linear Models, and SVMs) with the `GReNaDIne` library.
4. Analyze network communities and evaluate biological relevance via Gene Ontology (GO) enrichment.

2 Data Acquisition

Before writing code, establish your biological context:

- **Expression Data:** Download a clean expression matrix for your species of interest from the EBI Expression Atlas (bulk) or Single Cell Expression Atlas (pay attention to the dataset size).
- **Transcription Factors (TFs):** Download the list of TFs and cofactors for your target species from AnimalTFDB 4.0. Restricting candidate regulators to known TFs reduces false positives.

3 Pre-processing with Scanpy & TF Mapping

Load your dataset into Python, execute quality control, normalize, log-transform, select highly variable genes (HVGs), and annotate known regulators.

4 GRN Inference & Method Comparison

Using the GReNaDIne library (or gxn), model the inference task using different machine learning frameworks. To do so, you can work in groups to test different models such as Random Forests, SVMs and Linear models.

5 Network Characterization & Functional Enrichment

After inferring the network, convert top-ranked edges into a graph structure to evaluate modules:

1. **Community Detection:** Apply clustering algorithms (e.g., Leiden or Louvain via NetworkX/igraph) to uncover functional sub-networks.

2. **GO Enrichment:** Test whether identified communities are enriched for specific Gene Ontology terms to validate biological significance.