Predicting Transcript Isoform Abundance from Gene-Level Expression Data

The main objective of the project is to train a model able to predict the transcript isoform abundance from either bulk RNAseq dataframes or (10X) single cell RNAseq experiments. This procedure would allow to take a step further in granularity of RNAseq and deconvolve transcriptomic experiments, going from gene-level resolution to isoform-level resolution. The primary objective of this project is to develop a computational model capable of predicting transcript-isoform abundances from gene-level expression measurements. The model will be trained and validated on matched datasets: each bulk or single-cell RNA-seq sample (gene-level) is paired with a corresponding SMART-seq derived isoform-level abundance profile. Successfully achieving this will unlock isoform-level resolution for large existing RNA-seq datasets, enabling deeper transcriptomic insight at scale.

Current standard RNA-seq experiments (bulk or droplet single-cell, e.g., 10x) typically provide gene-level expression matrices: rows correspond to samples or cells, columns to ~ 40,000 human genes. However, obtaining isoform-level quantification (via SMART-seq or long-read methods) remains technically challenging, costly, and thus less widely available. By leveraging the abundant gene-level data and learning the mapping to isoform-level abundances, this project aims to bridge that gap — enabling the inference of isoform usage where only gene-level data exists.

Data was generated by RNAseq experiments and came in the thoroughly annotated *.h5ad* format, with cell type or sample provenance is highlighted, and mainly consist of the gene-level expression matrix where each of the approximately 40k human genes is a column and the rows are samples (or cells).

The target is a vector of isoform abundances in the sample, and it was generated with an expensive procedure called SMART-seq.

The dataset is matched: each of the bulks will have its counterpart in the Smart-seq transcript isoform dataframe and this allows supervised learning.

A fully-connected neural network (feed-forward architecture) will be constructed and trained. Hyper-parameter tuning and iterative experimentation will compare different forms of input representation:

- raw input data with no transformation
- and on a dimensionality reduction of the input data (PCA or VAE)
- embeddings computed by pre-trained models available in literature, e.g., **scGPT**, **Geneformer**, **BulkFormer** [1, 2, 3]

To test the model, we strive to implement multiple strategies to assess the predictive capabilities of the model by testing with a train-valid. split with stratification according to cell type or sample origin.

Evaluation metrics proposed are:

- root mean squared deviation from the ground truth of the test samples;
- correlation of the predictions to the ground truth;
- eventually, classification error regarding the major class between predictions and ground truth.