Summary of prediction for Challenge 2

Participant: Vu Hong Ai (vuhongai)

Since the prediction of proportion of 5 states of different knock-out solely based on the gene names, my approach is to learn the minimal representation of individual genes based on the provided single-cell RNAseq and use this representation to predict the 5-state proportion by training simple neural network.

Different steps to building the prediction model include:

1. Learn the minimal representation of gene (gene embedding)

In recent study [1], researcher presented an unsupervised method (scETM) to reconstruct the scRNA-seq by learning gene and topic embedding. A notable modification in the original code to adapt to this challenge is the size of gene embedding reduced from 400 to 32 dimension. I saw no significant reduction of negative log-likelihood and adjusted rand index (ARI) on held-out samples. The 32-dimension gene embedding vector of all 15077 genes were then extracted for the next steps.

2. Train a neural network to predict the 5-state proportion

By training the model in step 1, I obtained an embedding vector of 32-dimension for each gene. A preprocessing step by StandardScaler (sklearn) was first applied to embedding vector of all 15077 genes before training the model. The models in this step will map the 32-dimension vector gene embedding to 5-dimension vector output, only contain fully connected dense layers, with both l1 and l2 regularization to stabilize the training. The loss is a combination of *kl_divergence* and *mean_absolute_error* losses. Hyperparameter details can be found in *step2_neural_network.ipynb* notebook. Each model was trained 5 times with Early Stopping callback. The prediction on all 15077 genes of selected "good" models were used to be filtered in the step 3.

3. Sample the "best" prediction from different models.

Since the predictions of different neural network in step 2 fluctuates considerably, it is necessary to select the most frequent output(s) rather than directly averaging all predictions. Here, I selected 2 predictions from 2 best performed models for each k-fold training, based on MAE of validation set (max_mae_val < 0.1), resulting a list of 20 predictions (vector of 5) for each gene. I then calculated the mean_absolute_error of all pairs within these 20 predictions, select and average the prediction(s) with highest number of similar prediction outputs (max_mae < 0.06) for submission.

1. Zhao, Y., Cai, H., Zhang, Z., Tang, J. & Li, Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature communications* **12**, 5261 (2021).