Cancer Immunotherapy Grand Data Science Challenge: Challenge 3

Team "St3p"

Abstract

Contents

1 Problem Statement

The goal of this challenge is to develop new metrics for ranking perturbations in terms of their capacity to move cells from an undesired to a desired state. Propose a statistic s(·) that summarizes the gene expression distribution P_i obtained from knocking out gene i as well as a scoring function that takes in P_0, Q and the predicted statistic $\hat{s}(P_i)$ and outputs the score of knocking out gene i (where a larger score indicates a better perturbation).[1]

1.1 Baseline Definitions

Let P_0 denote the empirical gene expression distribution of the unperturbed cells, i.e., P_0 is a distribution in 15,077-dimensional space. Similarly, let P_i denote the gene expression distribution of the cells obtained by knocking out gene i. Let Q denote the desired cell state proportion vector, i.e., Q is a 5-dimensional vector of probabilities that add up to 1.

2 Proposed Method

To prepare the data, we preserve the data preparation pipeline described in the introductory videos (c.f. Figure 1), within initial feature selection based on variance across cell states, then dimensionality reduction via principal component analysis (PCA), and finally additional dimensionality reduction, community detection and labeling based on UMAP embeddings, K-nearest neighbors, and Leiden community detection.

2.1 Statistic

We propose as a relevant statistic a cluster quality metric that represents whether our clusters are both internally dense and maximally separated from each other.

[1] Additional desiderata of the proposed statistic and scoring function are: 1) scoring function should ideally depend on P_0; 2) scoring function should ideally take uncertainty into account given by the different sample sizes for the perturbations in the training dataset; 3) different perturbations lead to different growth rates and thus result in a different number of cells; the scoring function may also take into account the predicted number of cells resulting from a guide and favor perturbations with a large growth rate; 4) scoring function could take into account the classification boundaries of each cell state; 5) it may be helpful to use a more informative statistic than the cell state proportion vector that could for example take into account the classification boundaries of each cell state.

To compute this statistic, we first look at each at each cluster individually, and compute a density metric, which is defined as the square of the distances from the cluster's centroid to each data point within the cluster, then taking the square root. A higher number here represents lesser cluster density.

We then sum the cluster quality for each of the clusters identified within the UMAP space ($N_clusters$ is assumed to be 5, but it is not strictly required). A higher number here represents lesser cluster density.

After summing the individual cluster quality metrics, we add the average distance between each pair of centroids, to capture the degree to which clusters are separated from each other, within UMAP space. A higher number here represents lesser cluster quality - capturing both density within clusters, and separation between them.

(Conceptually, a cluster of greater quality can be more confidently labeled given *a priori* knowledge of the role of one of its constituents - e.g. if tcf7 is more highly expressed in one cluster, vs. all other clusters, we conclude that this cluster represents the progenitor state - but that confidence is lesser if the cluster is not internally dense, and maximally separated.)

2.2 Scoring Function

Our scoring function is then defined as the proportion of cells in your desired state(s), divided by the quality of the clusters that you were able to create based on the data from your perturbation.

(If, for example, your objective function was to maximize the proportion of progenitor, effector, and cycling cells, and the estimated state proportion vector from your perturbation, Q_i, was [0.2, 0.1, 0.3, 0, 0.4], the proportion of cells in your desired state would be 0.7.)

This effectively corrects the degree to which you've satisfied your objective function (which you can define in any way you'd like), for the degree to which you were able to cleanly reduce your expression vectors into clean estimates of the gene expression programs, and groups of gene expression programs, that define your T-cell states.

(In addition, to capture the statistical confidnece you'd gain by having generated large numbers of cells in your Perturb-seq experience, you could multiple the number of cells measured - the number that survive quality control filters - by the proportion of cells in your desired state(s), before dividing by the clustering quality metric.)

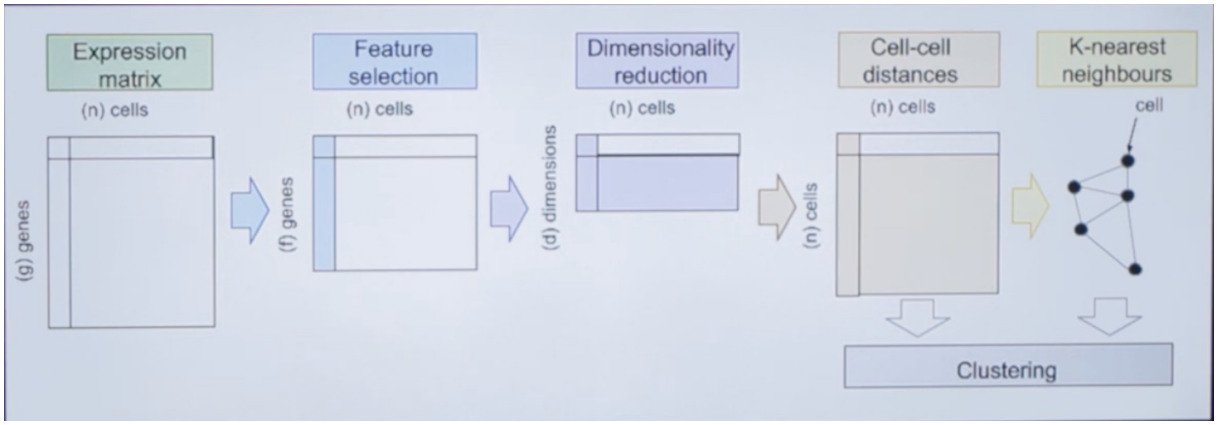

Figure 1: Initial Data Preparation Pipeline