Challenge 3: Scoring Function for Proposed Perturbations

With the purpose of developing a new scoring function to evaluate perturbations in terms of their capacity to move cells from an undesired to a desired state, we are going to propose a statistic $s(\cdot)$ which could summarize both gene expression and state proportion at the individual cell level.

Let P_i denote the gene expression distribution of cells obtained by knocking out gene i, and P_0 denote the empirical gene expression distribution of unperturbed cells. Both P_i and P_0 are 15,077-dimensional vectors. Similarly, let R_i [1] denote the cell proportion vector of cells obtained by knocking out gene i, then R_i is a 5-dimensional vector of probabilities that add up to 1. And Q is the desired cell state proportion vector.

In order to measure perturbations more comprehensively, we would like to process P_i and R_i separately at first, leading to sub-statistic s_i^P and s_i^R, then get the two parts involved when constructing $s(\cdot)$ in the end.

1. Measurement of Gene Expression - s_i^P

Though P_i contains complete information about the full gene expression distribution obtained by knocking out gene i, it would require a large sample size to get an accurate estimate of P_i. This may not be necessary since we are identifying optimal perturbations with respect to the desired 5-dimensional Q. So we do screening of the genes with the method proposed in the first two challenges.

As a brief review, for the original gene expression matrix, we calculated the variances and correlation coefficients of the columns. For two highly correlated columns, we only keep the one with a larger variance. Note that, here, by "highly correlated", we mean the two columns with a correlation coefficient which is larger than the 80% quantile of all the correlation coefficients. And we end up with 8,334 columns in the expression matrix. In other word, in terms of correlation and variability, we believe that the remaining 8,334 genes are sufficient to provide the full gene expression information.

Let P'_i denote the filtered gene expression vector, i.e., $P'_i = (p_{i1}, p_{i2}, \ldots, p_{i,8334})$, where $i = 0,1,2,\ldots$ The more different P'_i from P'_0, the greater the change in gene expression of a cell after knocking out gene i. We then calculate the l_1-loss of P'_i and P'_0, and would like to use it to reflect the degree of change in the expression of remaining genes of the cell in which gene i was knocked out.

We also take into account the sample sizes of various perturbations because knockouts with larger sample sizes are more robust against contingency and thus should be given more weight. And a sample size of 30 is usually considered large enough in statistics. Let n_i be the number of cells obtained by knocking out gene i. We define the weight as follows:

[1] R_i can be obtained with the model constructed in Challenge 1 and 2. The details of prediction will not be mentioned here.

$$w_i^P = \begin{cases} \frac{n_i}{30} & 1 \le n_i < 30 \\ 1 & n_i \ge 30 \end{cases} \quad (1)$$

Then the sub-statistic s_i^P used to measure the gene expression of a perturbation of knocking out i is given by

$$s_i^P = w_i^P \cdot \left(\sum_{k=1}^{8334} |p_{ik} - p_{0k}|\right) \quad (2)$$

2. Measurement of state distribution - s_i^R

A good perturbation should lead to the predicted cell proportion vector similar to the given Q as much as possible. We hope to use a method similar to the objective function used in part (b) of Challenge 2 to evaluate a knockout.

We first check the cell proportion vector $R_i = (r_{1i}, \dots, r_{5i})$, where $r_{1i}, \dots, r_{5i}$ represents the proportion of progenitor, effector, terminal exhausted, cycling and other state respectively. If $\sum_{t=1}^{5} r_{ti} = 1$, we continue following procedures. The state "other" is relatively less useful and won't be taken into consideration. Suppose $Q = (q_1, \dots, q_5)$, We then calculate the loss of R_i and Q as follows, which denotes the "distance" between R_i and the desired distribution.

$$l_i = \frac{r_{1i}}{q_1} + \frac{r_{2i}}{q_2} - \frac{r_{3i}}{q_3} + \frac{r_{4i}}{q_4} \quad (3)$$

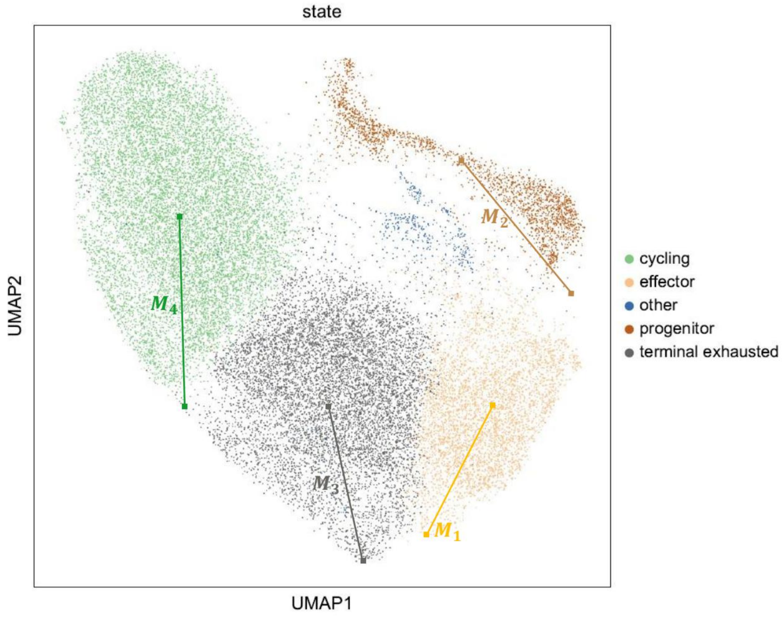

Figure 1. UMAP of Original Gene Expression Data

Recall that the UMAP shown in Challenge 1 colors each cell population within the projected space by its cell state distribution. A simple idea is that perhaps cells lying close to the boundaries might be assigned less confidence than those lying more towards the center of a state population. Inspired by this, we define another weight w_i^R. To be specific, let $cell_t^0$ be the centered cell for the t^{th} state on UMAP, where $t = 1, 2, 3\ 4, 5$ represents the progenitor, effector, terminal exhausted, cycling and other state respectively. Suppose f_t[2] is a function related to the state t. Given R_i, f_t will return the distance between the corresponding cell and $cell_t^0$. Let M_t be the farthest distance of a cell in a certain state from the center of that state, i.e. $M_t = \max_{cell(R)\in\{the\ t^{th}\ state\}} f_t(R)$, where $cell(R)$ is the cell whose state distribution is given by R. Suppose $\mathcal{J}(R_i) = \max_j r_{ji}$ is an "indicator" mapping from R_i to the most possible state. We define the weight as follows:

$$w_i^R = 1 - \frac{f_{\mathcal{J}(R_i)}(R_i)}{M_{\mathcal{J}(R_i)}} \tag{4}$$

Then the sub-statistic s_i^R used to measure the state distribution of a knockout is given by

$$s_i^R = \mathbb{I}\{\sum_{t=1}^{5} r_{ti} = 1\}\cdot w_i^R l_i \tag{5}$$

3. Measurement of a Given Perturbation - s

Based on the results above, s_i^P and s_i^R will evaluate whether a given perturbation is desirable from two different perspectives, i.e. gene expression and state distribution. Now we need to find a way to combine s_i^P and s_i^R as the final scoring statistic s in a way such that both sub-statistic can be involved and used as adjustments for each other. Here, we consider marking the knockout genes as points in two-dimensional space.

The construction of s_i^P and s_i^R makes sure that both sub-statistic only take positive values, and a larger value indicates a better perturbation. If the value of the horizontal coordinate of a point is large, i.e. s_i^R is large, then knocking out gene i tends to induce quite different gene expression from the control group(no perturbation). Similarly, if s_i^R is large, knocking out gene i may bring quite ideal state distributions. For a certain knockout gene i, when P_i and R_i are given, we compute the s_i^P and s_i^R, then mark the point (s_i^P,s_i^R) on a plane, as show in Figure 2.

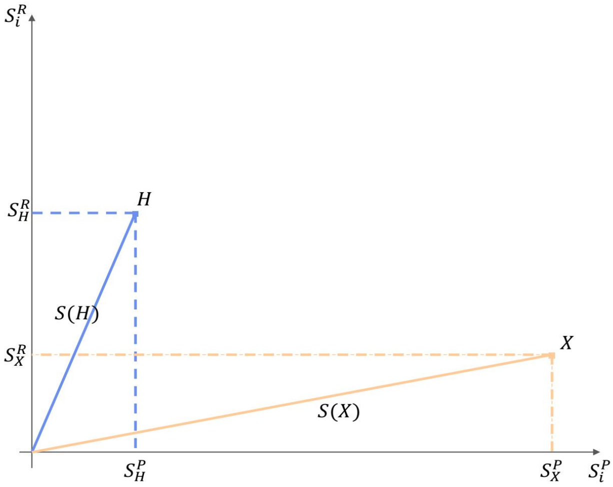

Figure 2. Example of Proposed Scoring Function

[2] The detailed definition of the function is not given here. Please refer to the computing process of generating the UMAP in the code sample provided in Challenge 1.

Note that here we just use s_i^P and s_i^R as coordinators to mark the positions of points. There is no necessary logic relation between the two, such as the commonly used dependent and independent variables(i.e., "y" and "x"). As shown in Figure 2, X performs better than H in gene expression because X has a larger s_X^P, but H outperforms X in state distribution with a larger s_H^R. In order to give comprehensive evaluations, we decide to use the distance from a point to the origin as the final statistic, that is

$$s(i) = \sqrt{\left(s_i^P\right)^2 + \left(s_i^R\right)^2} \tag{6}$$

In this way, s_i^P and s_i^R could make up for one another if the perturbation performs pretty good in one aspect while relatively worse in the other.

4. Conclusion

Generally speaking, we propose the scoring statistic $s(\cdot)$ from two aspects, gene expression and state distribution, resulting in two sub-statistic s_i^P and s_i^R; then we manage to combine the two in a reasonable and comprehensive way, bringing the final statistic $s(\cdot)$. Given a knocked out gene i, corresponding gene expression vector P_i and state proportion vector R_i in perturbed cells as well as the desired proportion vector Q, the evaluation score for i is given by

$$\begin{aligned} s(i, P_i, R_i, Q) \ &= \ \sqrt{\left(s_i^P\right)^2 + \left(s_i^R\right)^2} \\ = \ &\sqrt{\left(w_i^P \cdot \left(\sum_{k=1}^{8334} |p_{ik} - p_{0k}|\right)\right)^2 + \left(\mathbb{I}\{\sum_{t=1}^{5} r_{ti} = 1\} \cdot w_i^R l_i\right)^2} \end{aligned} \tag{7}$$

where w_i^P and w_i^R are defined by (1) and (4) respectively. And a higher score $s(\cdot)$ means that the corresponding perturbation could bring more desirable results.