Prioritizing Cellular Perturbation Experiments using Distributional and Time-Evolving Criteria

Cancer Immunotherapy Grand Data Science Challenge - Challenge 3

The RNAvengers

Abstract

A major obstacle to developing robust immunotherapies is the difficulty of identifying genes whose manipulation (say, a knock-out by CRISPR) would have a beneficial effect for a particular individual's cancer. This is complicated by the difficulty in generalizing predictions across data from multiple individuals, gene manipulation targets, and tumor heterogenity. We propose that an effective method for prioritizing genes for research should: account for the aleatory and epistemic uncertainty associated with the model predicting the outcomes; account for the inter-cellular variation in outcomes (i.e: not all cells respond the same to a given perturbation); and model the time-evolving distribution of T-cell types. We propose some simple metrics that approach these goals.

1 Introduction

To borrow the notation of the challenge text:

> 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.

Our objective is to propose a simple summary statistic $s(\cdot)$ that can be applied to a predicted distribution to provide useful information for a scoring function, which we will call U, for "utility".

2 Summary statistic s

We propose that, in addition to predicted cell state distributions, $\hat{Q}$, s incorporate information about the distributional variety of gene expressions produced under the perturbation (and predicted by the model). This identifies models that fail to capture the differential cellular outcomes of the perturbation, which could improve predictions and ultimately help decrease the cost of immunotherapy drug discovery and treatment.

Thus we propose, $s(\hat{P}_i) =< Q(\hat{P}_i), \frac{V[\hat{P}_i]}{V[P^*]} >$, where $V[\cdot]$ is the variance across the distribution of the expression vectors, and P^* is an ideal distribution of T-cell transcripts (such as in a successful immune response, or resampled based on a desired Q^*).

3 Scoring Function U

The L1 score (used by Challenge 1) fails to differentiate failure modes with wildly different proportional differences. For example, with target distribution $< .95, 0, 0, 0, .05 >$, the vectors $< .9, .05, 0, 0, .05 >$ and $< .95, .05, 0, 0, 0 >$ are scored equivalently, despite functionally very different distributions. We propose instead the scoring function:

$$U(Q^*, s(\hat{P}_i) =< \hat{Q}_i, V_i >) = V_i \cdot \left(\sum_{j=1}^{5} |\log \frac{2 * \hat{Q}_{i(j)}}{\hat{Q}_{i(j)} + Q^*_{(j)}}| \right)^{-1}$$

4 Bonus: Choosing target Q

The ideal Q depends on the time at which it is measured, as the distribution of T-cell types naturally evolves over time. Time-series experiments detailing the single cell profile T-cells across multiple time points could provide an interesting distribution the shifting landscape of T-cell types. Single cell sequencing of the tumor cells concurrently could shed light on the changing tumor dynamics over time, and the subsequent response of T-cells states. Computational modeling of changes in the T-cell transcripts over a tumor lifetime could provide insight into T-cell state differentiation within the context of the changing tumor micro-environment and suggest temporally-informed interventions.

However, time-series experiments can be costly and difficult, and much information can still be extracted with single time-point experiments. For example, perturbations that activate proliferative pathways for anti-cancer T-cell types could be an useful metric to quantify across experimental and computational perturbations.

5 Conclusion

In conclusion, we have proposed summary statistic s that models distributional properties of the predicted gene transcripts, a scoring function U that favors the probably-most-beneficial genes, and we have also suggested further metrics for evaluating the target distribution Q.