\*\*\* Cancer Immunotherapy Data Science Grand Challenge - Challenge 3 \*\*\*

Topcoder ID: SA799

challenge Goal: to make T cells, the fighter cells of our immune system, better at killing cancer cells.

Developing Cancer Immunotherapy Data science based on Machine learning, As Cancer Disease growing Rapidly nowadays, hope machine learning really helps to make feasible causes to develop the cure.

While scientists have tested some genetic modifications, or “perturbations,” to T cells in the lab, there are too many possible perturbations — and combinations of perturbations — to solve this problem experimentally. Our goal in this competition is to bring machine learning to this problem to help identify which perturbations could make cancer immunotherapy more effective.

T cell dysfunction is a hallmark of many cancers, but the basis for T cell dysfunction and the mechanisms by which antibody blockade of the inhibitory receptor PD-1 (anti-PD-1) reinvigorates T cells are not fully understood. Here we show that such therapy acts on a specific subpopulation of exhausted CD8+ tumor-infiltrating lymphocytes (TILs).

Dysfunctional CD8+ TILs possess canonical epigenetic and transcriptional features of exhaustion that mirror those seen in chronic viral infection. Exhausted CD8+ TILs include a subpopulation of 'progenitor exhausted' cells that retain polyfunctionality, persist long term and differentiate into 'terminally exhausted' TILs. Consequently, progenitor exhausted CD8+ TILs are better able to control tumor growth than are terminally exhausted T cells.

Progenitor exhausted TILs can respond to anti-PD-1 therapy, but terminally exhausted TILs cannot. Patients with melanoma who have a higher percentage of progenitor exhausted cells experience a longer duration of response to checkpoint-blockade therapy. Thus, approaches to expand the population of progenitor exhausted CD8+ T cells might be an important component of improving the response to checkpoint blockade.

While the problem of how to score different perturbations is critical for any down-stream task, there has been little work on identifying strong metrics, and simple metrics like the ones described in Challenge 2 (A) and (B) are currently being used (for more details about (A) and (B), refer to Challenge Part Details section in Challenge 2).

Let "Po" denote the empirical gene expression distribution of the unperturbed cells, i.e., "Po" is a distribution in 15,077-dimensional space. Similarly, let "Pi" 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. As an optional task, you are invited to submit your proposal for how to choose Q for cancer immunotherapy.

From challenge-2, part-A the objective is obtained for maximum proportion of progenitor cells over 15077 genes along with cycling\_constraint considering 5% of cycling cell state proportions.

# Scoring function

From challenge 2, the scoring function is

Po = Po-Pi

in which Po = (1,0,0,0,0) which is ideal cell state proportion vector which increases the maximum percentage of progenitor cells in the tumor.

Chimeric antigen receptor (CAR) T cells mediate anti-tumour effects in a small subset of patients with cancer, but dysfunction due to T cell exhaustion is an important barrier to progress. To investigate the biology of exhaustion in human T cells expressing CAR receptors, we used a model system with a tonically signalling CAR, which induces hallmark features of exhaustion.

For experimental design purposes, the statistic s(⋅) should be predictable for unseen perturbations. Also, keep in mind the trade-off between estimating a higher-dimensional and more informative statistic and the requirements in terms of sample size.

For example, using the identity map as a statistic would require predicting the full gene expression distribution obtained by knocking out gene i, which would require a large sample size to get an accurate estimate and may not be necessary for identifying optimal perturbations with respect to the desired 5-dimensional cell state proportion vector Q.

# Statistics function

Statistic function is the ratio of cell state proportion vector of perturbation cells to predicted cell state proportions of corresponding gene i.

Let the statistics function to be

S(X) = | X | / | (Y - Y1) |

where the function X-intercept denotes slope of curve

X = Q (be predicted proportion of 5 cell state proportion vectors of 15077 genes based on adataX)

i.e., adataX = adata[0:15077, :5].X.toarray()

Y = Po (ideal cell state proportion vector)

Y1 = Pi (gene expression distribution vector)

where Y - Y1 is the scoring function from challenge 2 where to increase the proportion of progenitor cells.

In part-B the objective is obtained for ratio of predicted cell state proportions to cell state proportions of pertubation cells in adata.X along with the cycling\_constraint considering 5% of cycling cell state proportions in part(A), the objective has positive and negative infinity values obtained that signifies no route cause, taking inverse ratio of cell state proportions which gives rise to non-infinity objective values, i.e., 0 instead of +inf or -inf or undefined values.

where these zero gene values to be removed from the no. of cells in tumor where uncertainty may lead to clotting those cells in tumor, and be removed which can increase the number of cells fighting the cancer.