Cancer Immunotherapy Grand Data Science Challenge - Challenge 3 Peer Review

Reviewer: Samuel Chazy - Submission ID: 676611

Review 1:
Submission ID: 674189

The proposed solution entitled: Clonally Aware Measure of Effector Longevity or CAMEL is an interesting approach to the challenge. Optimizing the overall behavior of clones rather than cells makes sense. The author proposes to apply this by measuring the clonal capacity not as a function of time, but as a function of clone size. The theory is worth testing in my opinion. However, the proposed linear regression between clonal expansion time and terminally exhausted fraction is more of a baseline or a start. Further research should be explored to find the best way to solve or compare both variables.

Although the author has a novel approach for the challenge, he/she missed providing a clear statistic and scoring function as per the requirement of the challenge.

Review 2:
Submission ID: 676333

While going through the submitted files, I found it to be quite confusing as to where we should be reading and for what objective. The submission had 4 files: Statistic.txt, research.txt, Research_2.txt, and the goal of the scoring function.txt. As for the statistic part, the author proposes to use the Euclidean distance to measure the error for the gene expression distribution between the knock-out and the desired cells. The distance would be used to train a couple of ML algorithms along with cross validation to ensure generalization of the model to new data. While this approach is valid, I don't see novelty in the method.

However, calculating a weighted average with penalty to measure the difference in the distributions for the scoring function is a valid idea. The formula provided for the scoring function using the Kullback-Leibler divergence is a good proposition.

Note: There were no references / citations given in this proposal.

Review 3:
Submission ID: 676583

The statistic proposed by filtering the expression through their mean, variance across cell states, and variance across knocked-out cells is valid. However, the approach is not unique and needs experimentation to see if it will yield into the desired results.

Th proposed scoring function for perturbations is based on a numerator that takes the mean expression of the genes across the cell states, divided by a denominator that takes the variance of the gene's expression across the cell states. The function integrates a variance variable Vj that decides the cell state boundaries' confidence, which I find interesting and novel as an approach. In addition, it integrates a co-variance variable for all the states to favor the desired cell sate. Without testing the formula, it is difficult to judge its effectiveness, but it looks promising.

On the other hand, the author proposes a scoring function for cell states. It is not clear how the formula came about or what was it based on. This formula needs further explanation.

As for the proposal for Q, the author argues that the T cell states are currently independent of each other, only to refute this this idea later. However, it is clear from the beginning that the states are not independent from each other. Beyond this idea, the proposal for Q is not clear.

Review 4:
Submission ID: 676613

The author presented us with a detailed and well summarized steps of cancer definition, certain treatments, and cell behavior. However, I believe that the author is deviating from the challenge request. By proposing a different treatment, Cytokine Therapy, the author is venturing into an unknown field especially for us as data scientists. This is an approach that was not required in the challenge. We don't know whether favoring the effector and cycling cells instead of the progenitor cells would reach to a better result. Moreover, the author didn't propose a statistic nor a scoring function that were clearly defined as per the challenge request. This is not to state that Cytokine Therapy is ineffective, but rather to say that it is not the requirement of the challenge.

Review 5:
Submission ID: 676622

I couldn't find any document that has a write-up that solves challenge 3 requirement. The files provided were .py files for challenge 1 & 2. The Text documents were a recap of the 3 challenges requirements only.

Review 6:
Submission ID: 676646

The proposed statistic is valid and interesting. Using Principal Component Analysis to capture the variance in the gene expression, then to predict the cell type label, and then to predict cell type probability vector for each perturbed cell is a smart approach. After excluding cells with higher target gene expression than the mean of the unperturbed cells and averaging the probabilities of the remaining cells, the authors concatenate the average cell type probabilities with the total number of cells per perturbation. This statistic has a high potential.

The authors propose a scoring function with a weighted sum of the logarithm of the significance of the change induced in cell type numbers multiplied by 1 or -1. While this function is valid and promising, the authors didn't elaborate well on how to handle the weights, increasing or decreasing. Suggesting an increase in the weights based on previous observation is not enough. Moreover, the authors suggest normalizing the observations to a constant of 100 cells, so that the desired signal in the distribution is not lost. I find that flattening the observations to a constant number is limiting, especially when there isn't a method to determine this number.

Note: There were no references / citations given in this proposal.

Review 7:
Submission ID: 676672

The author proposes a statistic like the one provided in the challenge. Therefore, no novelty was introduced to this part of the challenge.

As for the scoring function, the author proposes the Euclidean distance which is the ratio between the ideal state proportion vector Q and the predicted Q-hat. While this method is a valid approach, it doesn't bring a unique or innovative approach to this challenge. Moreover, this approach doesn't take into consideration neither the growth factor nor the uncertainty factor.

Note: There were no references / citations given in this proposal.