Cancer Immunotherapy Grand Data Science Challenge - Challenge 3 Peer Review_v2

Reviewer: Samuel Chazy - Submission ID: 676611

Review 1:
Submission ID: 674631

The author didn't clearly differentiate between the statistic from the scoring function proposal. However, we can infer that the author utilized the same statistic as the one proposed in the challenge without bringing any novelty or change to it.

On the other hand, the author proposes an interesting scoring function based on the inverse of a weighted average of the knockout losses minus the loss of the unperturbed distribution. Introducing the variance in this case gives less weight to the variables with most uncertainties and account for the growth rate. Taking the exponential of the denominator and the numerator, however, is not well justified to avoid negative values in the scoring function. This formula can be further explored.

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

Review 2:
Submission ID: 676621

The authors propose an interesting and novel approach to the statistic part of this challenge. Computing KNN graphs across cells and then utilizing neighborhood enrichment analysis to identify the strength of the perturbations is a smart approach. Furthermore, the approach is detailed in implementation through the use of GLM fits using the MILO package. However, the authors propose to count the number of neighborhoods based on the spatial FDR with a threshold of 0.15. There is no further explanation as to how this number came about. A 5-dimensional vector N is obtained by counting the number of enriched neighborhoods in each cell state, whereby then the vector is normalized to account for data imbalance. Overall, the statistic function is well thought of.

On the other hand, the authors propose a scoring function that aligns the N vector with delta Q by taking the negative and positive enriched neighborhood count as -1 & +1 denoted as D. Using the L1 distance, the score is calculated with a normalized D vector and the delta Q. Although that this is a valid approach, but the sample size, the growth rate, or the uncertainty in the data were not considered.

Review 3:
Submission ID: 676651

The author proposes a statistic based on a desired vector Q with proportions defined as (0.6, 0.25, 0.05, 0.1, 0.00). However, we don't know how the author reached these desired proportions, and there is no explanation to support this strategy.

As for the scoring function, the approach is rather interesting. On one hand, the Mahalanobis distance is a good alternative to the standard vector distances such as the Euclidean distance because it accounts for correlations between the variables using a covariance matrix. The covariance matrix accounts for the uncertainties in the measurements as well as the growth rate. On the other hand, cosine similarity measures the similarity between the vectors while considering the magnitude and the direction of the vectors. The dot product of these two terms seems promising.

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

Review 4:
Submission ID: 676654

The author proposes a statistic that incorporates information about the distributional variety of gene expressions. The objective is to capture the differential cellular outcomes of the perturbations and improve the predictions. The use of variance to capture the difference in outcomes is interesting, however, the use of P*, the ideal distribution, is not clearly explained. Further explanation is required as to what the ideal distribution is.

As for the scoring function, it is a dot product of two terms. The first term is Vi, is the variance across the distribution, and therefore leads back to the missing explanation in the statistic. The second term utilizes partly the log of 2 * the dot product of Qi(j) divided by the sum of Qi and Qj. It is not clear whether this equation is interesting or not, because the logic and process behind it is not clearly explained. I suggest that the author defines all the variables inside the equation and the reasoning behind it.

The bonus part of this solution suggest that we choose the target Q based on Time-series experiments. However, and as the author indicated, that will reveal to be a difficult and costly exercise.

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

Review 5:
Submission ID: 676657

The authors propose a statistic that represents the change from unperturbed to perturbed cells obtained from screening P_o. However, the authors failed to mention that this formula ΔQ_{io} is the Euclidean distance formula in higher dimension. Thereafter, the authors state that the rate of growth G is reflected by the cycling state and assume that a low cycling state probability indicates a low growth rate. They based this assumption on a spearman correlation score of 0.47 and a p-value of less than 0.05. However, correlations and p-values are an indication of a relationship between the variables and not to be confused with causality. This score and p-value doesn't tell us that one value is driving the other.

The final proposed formula for the scoring function is rather sophisticated and it is not clear if it can produce better results than the actual formula. We need to understand the reasoning behind adding $\Delta Q_i + \Delta Q_{io} + G - s(Pi)$.

Finally for the proposal for Q, the authors give an example of an ideal desired Q which is [0.6, 0.1, 0, 0.3, 0]. It is not clear how they arrived at these proportions or if there is any scientific method to come up or support this approach.

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

Review 6:
Submission ID: 676673

The author proposes to utilize the Kullbeck-Leibler divergence between two Dirichlet distributions parameterized by the normalized compositions PQ and Q as a statistic. While the approach is valid, there is no sufficient development of the thought process or reasoning or even the equation itself.

Utilizing the standard errors of the above parameters estimate to propose gaussian approximations of the parameter posteriors, sample from those posteriors, and evaluate the distribution over KL divergences is a sound approach. However, this approach needs further development and explanation and implementation.