Cancer Immunotherapy Data Science Grand Challenge 3
Scoring function

David S. Fischer

1 Motivation

First, we would like to take uncertainty in the empirical distribution estimates into account that derive from sample size ("uncertainty"). Second, we would like to positively account for the number of T cells in a tumor, thus favouring perturbations that do not just yield a high fraction of desired T cells out of all T cells, but a high total number of those desired T cell states ("abundance"). Third, the scoring function should take the unperturbed distribution of cells into account ("reference").

2 Method

Let Q be a desired distribution over five discrete cell states. The proposed score function s between the predicted perturbed distribution P_Q and Q as the Kullbeck-Leibler divergence between two Dirichlet distributions parameterized by the normalized compositions P_Q and Q. This divergence is available in analytical form and therefore easy to compute.

3 Uncertainty in the score function

Where perturbation samples are available, we can compute P_Q directly as a maximum likelihood estimator. Here, we can use the standard errors of the parameter estimates to propose gaussian approximations of the parameter posteriors, sample from those posteriors, and evaluate the distribution over KL divergences on these samples, thus tackling "uncertainty".