Cancer Immunotherapy Data Science Challenge – Submission to Challenge 3

Authors: Anna Hakobyan, Loan Vulliard.

Introduction

The T-cell composition has a major impact on the efficacy of immunotherapy response across a range of therapeutic strategies. A versatile framework to manipulate the composition, and the functional properties of a T-cell population is a promising new direction to enhance the available cancer treatments. This challenge encourages proposing a conceptual and implementable framework to select T-cell perturbations with a desired outcome. For this, one would need to predict the impact of a knockout and understand how these changes relate to the desired phenotypes.

Notation and definitions

We denote the gene expression distributions to be P_0 and P_i for unperturbed and perturbed cells for knockout i, respectively. The statistic s transforms the gene expression distributions P_i to an informative representation, from which the proportion of T cells in different states can be inferred. It is assumed that $s(P_i)$ can be approximated by $\hat{s}(i)$ for a given perturbation i. In challenges 1 and 2, s is set to be $s_{challenge}$ as defined below. This maps its input for a target i directly to the corresponding 5-level cell state proportion vector Q_i:

$$s_{challenge}(P_i) \to Q_i = (a_i, b_i, c_i, d_i, e_i) \tag{1}$$

Participants in the first challenge are asked to find a mapping $\hat{s}_{challenge}$ that, given a gene name, returns an expected cell state proportion vector.

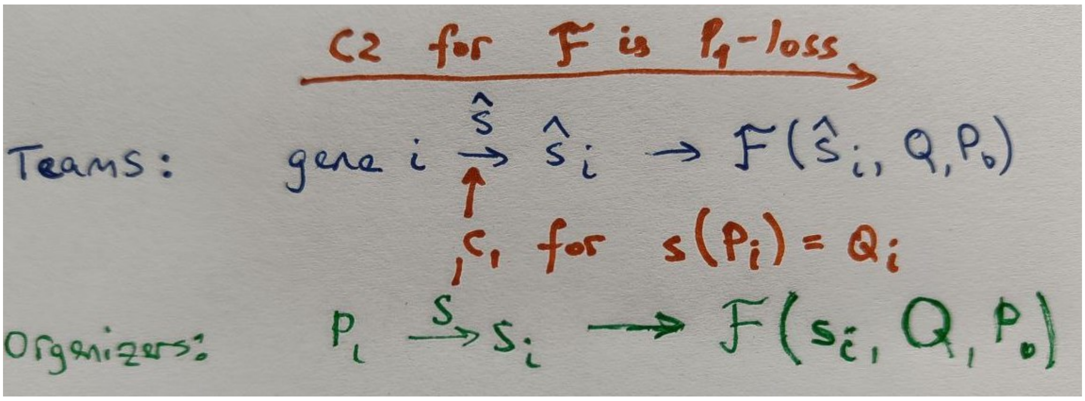

Figure 1: Global scheme of the framework, separating the prediction challenge performed by participants (in blue) and the validation task performed by the organizers (in green). Scopes of challenges 1 and 2 are shown in orange.

Given the desired cell state $\mathbf{Q}$, for an initial unperturbed condition P_0 and a knockout i, $\mathcal{F}(\hat{s}(i), \mathbf{Q}, P_0)$ returns a score which can be used to assess the agreement between predicted effect of perturbation i and desired state $\mathbf{Q}$. Following these notations, challenges 1 and 2 are using particular choices of the statistic s and the scoring function $\mathcal{F}$. In challenge 1, the participants were provided gene expression distributions for perturbed and unperturbed cells and a statistic s defined in (1). The task was then to learn a predictive model $\hat{s}(i)$, which approximates $s(P_i)$ for a perturbation i. The task of challenge 2 was to apply the learned predictive model $\hat{s}(i)$ over a range of genes, score the resulting $\hat{Q}_i$'s with a predefined scoring function $\mathcal{F}$ (*e.g.* l_1-loss with a fixed desired cell distribution $\mathbf{Q}$ in challenge 2A) and provide an ordered list of perturbations optimizing the score. The complete framework and how the different components are defined in challenges one and two are represented on Figure 1.

Goals

We aim to provide a statistic s and a scoring function $\mathcal{F}$ which, taken together, should offer a framework to identify the best knockouts for a chosen purpose. More precisely, this framework should be able to adapt to any desired cell composition $\mathbf{Q}$, and to take into account the initial transcriptomic state P_0 of a biological system of interest, making the approach transferable to other cell lines or patients. This is an important step towards having an *in silico* method to design the steps needed to engineer lymphocytes for immunotherapies.

Our proposition

Here, we propose a framework adaptable to diverse biological systems, by modeling changes in composition rather than absolute values. Then, we make the assumption that it is easier to predict whether a perturbation will have an effect or not rather than the effect itself, to reduce the search space to interesting candidates. We also provide a way to account for the compositional structure of cell state fractions.

Summary statistic

We propose that s maps P_i to the changes ΔQ_i that the knockout will induce in the cell population. Moreover, we also include a binary factor x_i that represents the ability of knockouts to induce significant changes in cell composition. Predicting x_i only involves predicting the norm of Q_i and not the five corresponding values and is therefore an easier computational task. The resulting score is defined in Equation 2.

$$s : P_i \rightarrow (\Delta a_i, \Delta b_i, \Delta c_i, \Delta d_i, \Delta e_i, x_i) \tag{2}$$

In practice, this information can be obtained explicitly in the case of the training set. The vector $Q_0 = (a_0, b_0, c_0, d_0, e_0)$ can be inferred using the method previously used for this challenge and subtracted from the vectors $(a_i, b_i, c_i, d_i, e_i)$ for each knockout. By learning from the examples in the training set, the corresponding statistic s can be approximated by $\hat{s}$ to be applied to any gene, as shown in Figure 2. Moreover, we suggest using probabilities of belonging to each class instead of the discrete values in Q_i when learning $\hat{s}$, to match the probabilistic nature of lineage commitment.

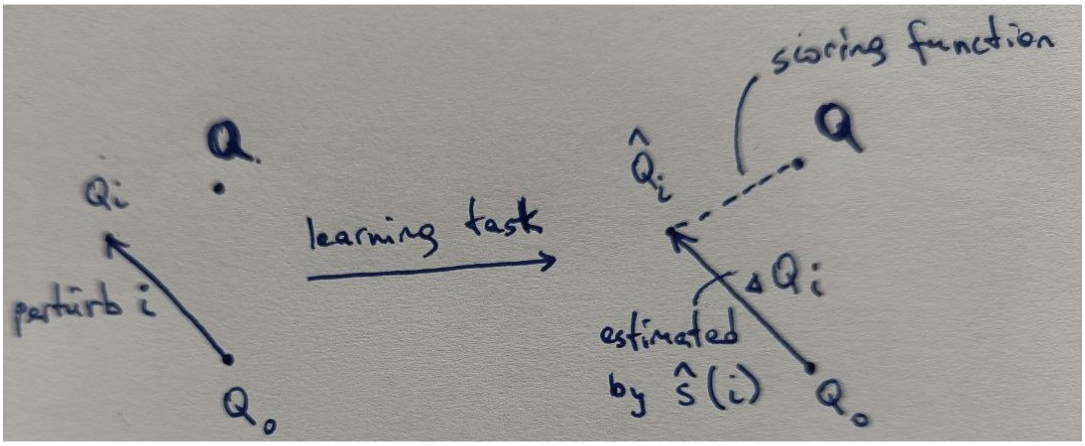

Figure 2: An illustration of the proposed statistic and the scoring function in the space of cell states.

The term x_i can be inferred by quantifying if the norm of the corresponding ΔQ_i is close to zero (in which case $x_i = 0$) or not (in which case $x_i = 1$). For this, we suggest defining a threshold based on the top 20 percentile of

ΔQ values for individual unperturbed cells. If there are only a few of such cells and many knockouts can be assumed to induce minor changes in composition Q_i, the threshold can also be defined across all cells in the training set.

In the case where we predict $x_i = 0$, the changes caused by knockout i will be small so the knockout won't be a good choice to reach another state $\mathbf{Q}$, and the estimation of Q_i can be skipped entirely.

Scoring function

To rank the ability of knockouts to reach a composition $\mathbf{Q}$ based on the predicted change vector $\hat{s}(i)$ and the unperturbed state P_0, we propose to compute $\mathcal{F}(\hat{s}(i), \mathbf{Q}, P_0)$ using the following approach:

- If $x_i = 0$, set $\hat{Q}_i = Q_0$ as the knockout is not expected to induce significant changes compared to the unperturbed state.
- Else, we compute the unperturbed cell composition Q_0 described above. Then, from $\hat{s}(i)$ we derive $\Delta Q_i = (\Delta a_i, \Delta b_i, \Delta c_i, \Delta d_i, \Delta e_i)$ and compute $\hat{Q}_i = \Delta Q_i + Q_0$.
- Next, we need to compare $\hat{Q}_i$ and $\mathbf{Q}$. Both $\mathbf{Q}$ and $\hat{Q}_i$ are compositional, *i.e.* the sum of all their components is equal to 1. A limitation of using the Euclidean distance D_E or similar metrics on such data is that we disregard the important difference between a complete absence of a cell state and its presence at a relatively low but functional level (Ricotta, 2021). To avoid this pitfall, represented in Figure 3, we first normalize each vector as detailed in Equation 3 (Orlóci, 2013).

$$T_C : (a_i, b_i, c_i, d_i, e_i) \to \left(\frac{a_i}{z}, \frac{b_i}{z}, \frac{c_i}{z}, \frac{d_i}{z}, \frac{e_i}{z}\right) \; (3)$$
$$\text{with } z = \sqrt{a_i^2 + b_i^2 + c_i^2 + d_i^2 + e_i^2}$$

- Finally we use the Euclidean distance on the normalized vectors. This is equivalent to transforming the data to a hyperspherical space and taking the length of the chord between both points, which is larger when the proportion vectors do not share all states (Orlóci, 2013).

$$\mathcal{F}(\mathbf{Q}, \hat{Q}_i) = D_C(\mathbf{Q}, \hat{Q}_i) = D_E(T_C(\mathbf{Q}), T_C(\hat{Q}_i)) \tag{4}$$

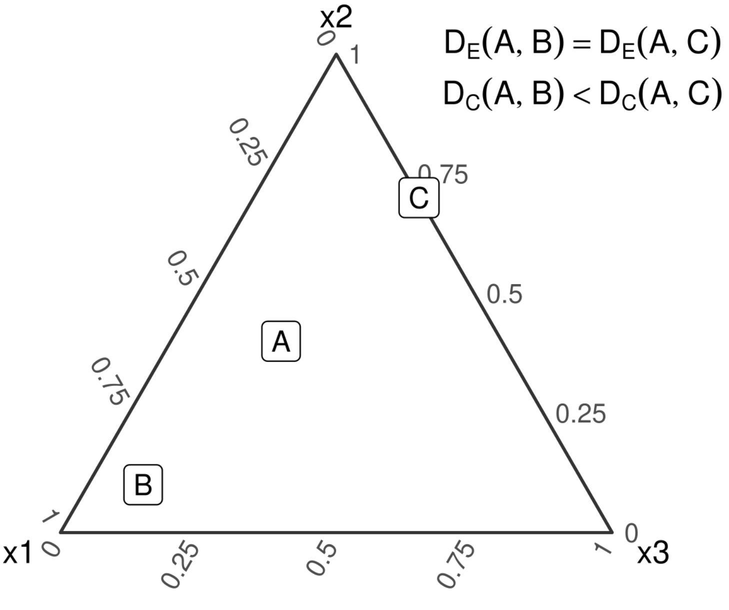

Figure 3: A ternary diagram demonstrating the inadequacy of Euclidean distance for compositional data. For points $A = (0.4, 0.4, 0.2)$, $B = (0.8, 0.1, 0.1)$, $C = (0, 0.7, 0.3)$, the Euclidean distance between A and B is equal to the distance between A and C, while point C is entirely lacking the first component. The proposed distance metric D_C penalizes differential compositions, hence is a more appropriate metric for Q.

Discussion

The framework we suggest builds upon the metrics used in parts one and two of the challenge. This aims to improve the biological ground on which the predictions rely, while being easily adaptable to the challenge dataset. First, the proposed statistic reflects not the cell state composition but their relative change. This allows to estimate the effective composition from a variety of initial conditions and potentially to predict the effect of combined perturbations. Second, the proposed variable x_i offers to bypass the prediction of irrelevant knockout effects, which reduces noise and improves performance. The addition of x_i further allows to integrate prior knowledge about genes and pathways that may be relevant to cell fate. Third, the scoring function considers the compositional nature of $\mathbf{Q}$, providing a biologically relevant scoring, where absence of a cell population is penalized heavier than a simple reduction.

Multiple options exist to achieve therapeutic goals, in addition to defining values of $\mathbf{Q}$. We encourage further research in identifying data-driven clusters of immune cells that differ between healthy tissue and tumors to constitute new treatment targets. Such catalogs have already been proposed for instance in the case of breast cancer (Jackson et al., 2020) and colon cancer (Hartmann et al., 2021). This shows that not only the T cell states but also their location and metabolic activity are needed for an efficient immune response. Further information about the immunogenicity of tumors should also be integrated in predictive models (Rooney et al., 2015). T cells also do not act in isolation but in a carefully orchestrated way together with other cells (Shilts et al., 2022). In conclusion, it will be important in the future to include more information about the T cells cytolitic activity and about the immune response at large, adding to the predefined set of T cell states.

References

Hartmann, F. J., Mrdjen, D., McCaffrey, E., Glass, D. R., Greenwald, N. F., Bharadwaj, A., ... & Bendall, S. C. (2021). Single-cell metabolic profiling of human cytotoxic T cells. Nature biotechnology, 39(2), 186-197.

Jackson, H. W., Fischer, J. R., Zanotelli, V. R., Ali, H. R., Mechera, R., Soysal, S. D., ... & Bodenmiller, B. (2020). The single-cell pathology landscape of breast cancer. Nature, 578(7796), 615-620.

Orlóci, L. (2013). Multivariate analysis in vegetation research. Springer.

Ricotta, C. (2021). From the Euclidean distance to compositional dissimilarity: What is gained and what is lost. Acta Oecologica, 111, 103732.

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., & Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell, 160(1-2), 48-61.

Shilts, J., Severin, Y., Galaway, F., Müller-Sienerth, N., Chong, Z. S., Pritchard, S., ... & Wright, G. J. (2022). A physical wiring diagram for the human immune system. Nature, 608(7922), 397-404.