Improve tumor specificity of CAR T cell therapy by selecting suitable combinations of antigens

Overview

CAR T cell therapy is a break-through technology in cancer therapy. In short, this method is based on engineering T cells with custom **C**himeric **A**ntigen **R**eceptors (CARs) that bind to antigens present on the surface of tumor cells. Since any given target antigen on the surface of a tumor cell is unavoidably also present on healthy tissue in the body, current CAR T cell therapy can lead to severe, sometimes even life-threatening side effects due to the destruction of this healthy tissue. The aim of the present project is therefore to identify antigen combinations that can better discriminate tumor cells from healthy tissue cells.

Bounding effects

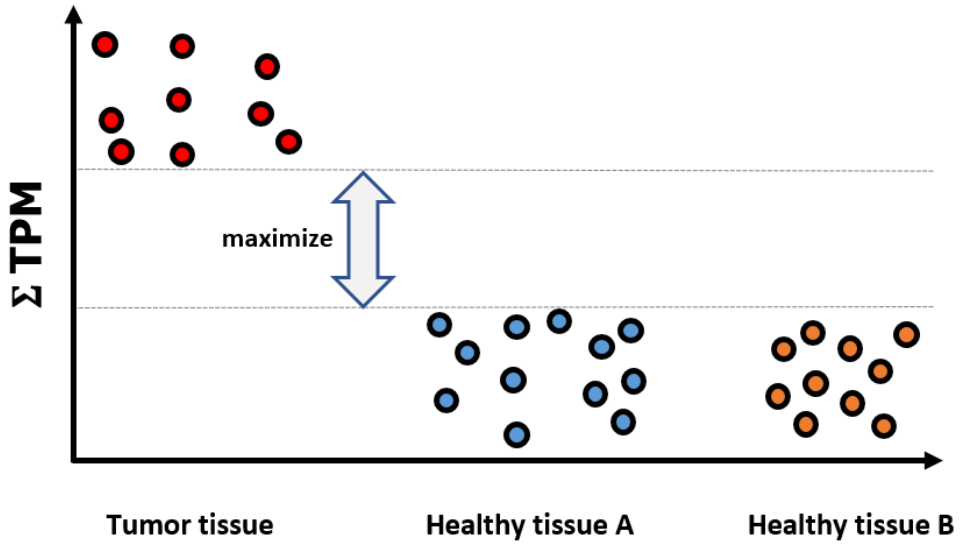
Activation of T cells after binding to its target antigens via the CAR has a sigmoid threshold behavior, which is determined by the number of antigen molecules on the surface of the target cells and the affinity of the interaction between the CAR and the antigen.

Dataset description

The provided dataset contains approx. 17,000 samples of healthy tissue from 53 different locations in the body and about 200 samples of neuroblastoma, one of the most frequent solid types of childhood cancer. For each tissue sample, the Transcripts per Million (TPM) value of about 3.3 thousand genes have been determined. TPM is a normalized unit that measures the amount of mRNA transcripts of a specific gene. It therefore indicates the level of surface expression of a specific antigen. (However, it has to be kept in mind that the level of mRNA transcripts is only a rough approximation for protein expression, because the rate of protein translation from mRNA can strongly vary between different genes.)

Objective function

The goal is to find a set of 5, 10 or 20 genes that maximizes the distance between tumor tissue samples and healthy tissue samples measured by the summed TPM values of the selected genes. The sum of TPM values of the selected genes should be high in all tumor tissue samples (in order to provide a therapy suited for all neuroblastoma patients without exceptions), while being at the same time low for all different healthy tissue types. That is, the CAR T cells - upon interaction with the identified combination of antigens (i.e. expressed genes) - should be activated by the tumor cells but not by any type of healthy cells. This objective is illustrated in the following visualization:



The distance between the tumor samples and the different healthy tissue types should be measured with different cluster linkage methods (single linkage, average linkage, 1st quartile linkage etc.). The most suitable metric will likely be a mix between single linkage (considering the distance between the two nearest observations) and average linkage (considering the distance between the average observations).

The medical experts know that there is no single gene with high TPM values for all tumor samples while at the same time yielding low values for all healthy tissue types. Combining different genes to fulfill the objective function is therefore crucial. The goal is not to find the single best solution, but rather to find several suitable solutions, which can be further analyzed by medical experts.

Why brute force is not feasible

While it may be obvious in the first place to just check different gene combinations for their fulfillment of the objective function, running some rough calculations quickly reveals that this is not feasible. Selecting 5 genes out of 3.3 thousand results in 3 x 10^15 different combinations to check. Even if one check takes only 1ms, checking all combinations would still take 95 thousand years of computation time.

Method 1: machine learning feature selection techniques

If the different tissue samples are viewed as observations, where the TPM values of the different genes are viewed as features of the observation and the tissue type as class assignment, the problem can be solved as feature selection in a classification problem. Classical machine learning approaches, such as Random Forest or Support Vector Machine, could be applied. Genes of the features that show a high importance in solving the classification problem are promising candidates for the stated problem.

One downside of this approach is that class separation in this feature space is a slightly different objective than maximizing the distance of summed TPM values between tumor and non-tumor tissue samples. In theory it could happen that a set of genes separate both classes well in the high dimensional space but yield no difference in summed TPM values between tumor and non-tumor samples. However, in practice this method should still lead to a good approximation of the main objective function.

Another conceivable way would be to utilize clever feature engineering to overcome these circumstances.

Method 2: heuristic optimization with Genetic algorithm (GA)

Despite the fact that various metaheuristics would be suitable for this problem, genetic algorithms are especially convenient here, since they allow to exploit expert knowledge by including promising genes in the initial population. In general, metaheuristics are suitable to find near optimal solutions in a computationally demanding problem. One disadvantages of solving this problem with metaheuristics is that they can lead to feasible but still long computation times.

Method 3: single gene analysis as pre-step

A third way to tackle this problem is to first process each gene on their own and omit unpromising genes. For instance, after calculating the cluster distance between tumor samples and each healthy tissue type for each gene, all genes that do not have a certain distance for at least 80% of tissue types could be omitted. The remaining genes could that be further processed either as initial solutions for heuristic optimization or if now feasible for solving the problem brute force.

While this method is fast and straight forward, it can easily happen that good solutions are omitted in single gene analysis such that no suitable solution at all is found. However, omitting only a small amount of very unpromising genes, could greatly increase the computation time of Method 1 or 2 without harming the outcome too much.