Novartis Institutes for BioMedical Research

Student Project at Novartis Institutes for Biomedical Research

Sample classification via machine learning

The transcriptome of a cell is the sum of all its RNA molecules at a specific time. It is highly dynamic and changes over time, e.g. in response to external stimuli. At Novartis, we routinely use RNA-seq to analyze the transcriptomic landscape of diverse samples, from varying tissues and diseases, and measured under different conditions. Thousands of measurements have been performed over the past years. This vast cohort of data can be of invaluable use for future experiments, to ensure better quality control as well as identify cohorts suitable for sample comparison. Thus, we would like to exploit the information generated in the past to augment new measurements by A) classification of newly sequenced samples and B) by identifying similar sets of samples.

In project-part (A) we aim at classifying new data into different categories, depending e.g. on patient age, tissue, or disease. A classical approach is to use supervised machine learning methods, such as Random Forest or Linear Regression. In a first step, a suitable training set can be generated based on a subset of the existing data. This training set is used to select important features necessary for sample classification and to learn a classifier, which can then be applied on new samples. However, it remains to be shown how well these techniques work with our internal cohort and how fine-grained the classification can be performed (e.g. accurate prediction of particular regions within tissues). In addition, we would like to test deep learning approaches on the classification and compare both machine learning and deep learning prediction accuracy.

In project-part (B) we aim at identifying sets of existing similar samples for newly sequenced samples. A common approach to determine sample similarity is to compute a distance matrix based on the gene expression observed across samples and to select the nearest neighbors for the sample of interest. Based on unsupervised clustering of samples and dimension reduction plots, we can investigate the quality of the selected sample set, also with focus of identifying potential biological/technical batch effects, such as treatment bias or tissue bias.

This project requires the application and implementation of different machine learning techniques. Therefore, solid knowledge of either R or Python are required, and knowledge of machine learning methods are a strong plus. In addition, a basic understanding of the cellular machinery with a focus on gene expression is essential.

Start Date:	Early 2021
Duration:	Up to 8 months
Location:	Basel, Switzerland (Novartis Campus)
Contact:	stefan.reinker@novartis.com

NOVARTIS