Silesian University of Technology

Faculty of Automatic Control, Electronics and Computer Science

Developing a system of automatic identification of cellular subpopulations in data from single-cell mass cytometry with the use of algorithms for grouping of high dimensional data

Doctoral Dissertation

by

Aleksandra Suwalska

Supervisor

Professor Joanna Polańska, PhD, DSc

2023

Gliwice, Poland

Abstract

The development of single-cell technologies like mass cytometry allowed for an in-depth understanding of cellular processes and the discovery of new cell subpopulations and their roles in the organism. The knowledge can help invent new therapies, drugs, and disease prevention methods. Mass cytometry enables the simultaneous measurement of dozens of markers used for cell identification, resulting in a matrix of expression values for each cell and feature. However, the high dimensionality of single-cell datasets makes the analysis a challenge.

Tuberculosis, an infectious disease caused by the bacteria *Mycobacterium Tuberculosis*, kills millions of people every year around the globe. The medications are expensive, especially when the germs resist one or more primary drugs. The outbreak of the COVID-19 pandemic has led to the loss of recent years' progress in combating the spread of Tuberculosis. Scientists are trying to reduce the occurrence of newly diagnosed cases to a minimum by working on new therapies and drugs using, among others, mass cytometry technology.

The dissertation thesis focuses on the analysis of high-dimensional mass cytometry datasets. Based on the publically available ones, the proposed analysis pipeline tries to solve problems occurring with the existing methods. The implemented algorithms are used to process the Tuberculosis dataset provided by partners from Stellenbosch University, RPA. The analysis includes a new technique that is fully automated and reproducible for pre-gating mass cytometry events. In addition, different methods for batch effect correction are compared in terms of removing the technical variance and influencing the cell-type identification results. The proposed machine learning technique of cell-type identification considers the existence of heterogeneity of cell groups during model evaluation which is a novel approach. Furthermore, introducing the expanded feature space and two-step clustering technique allows for obtaining well-defined and separated clusters.

The analyzes carried out indicate the potential of the introduced methods in the identification of cell types and the appropriate verification of their results.