Automatic Classification of Cellular Expression Data

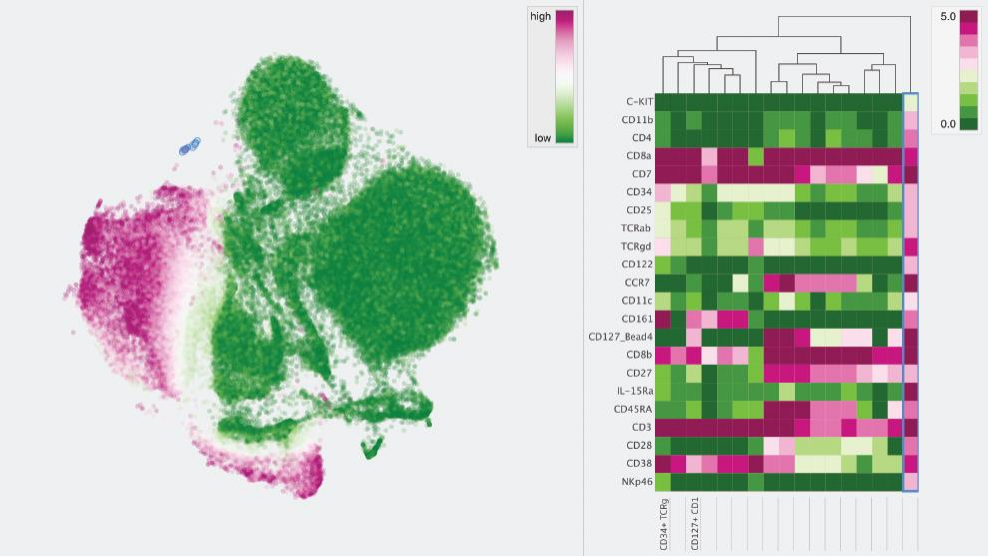


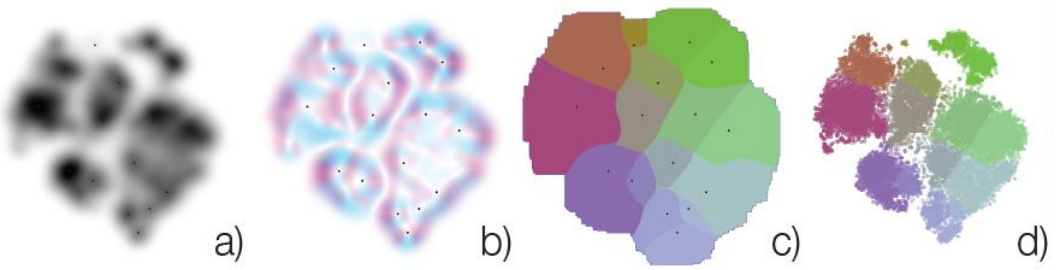
Figure 1 Dimensionality reduction-based visualization of a single cell dataset, acquired using Mass Cytometry (left) and a heatmap visualization of automatically generated clusters of the same dataset (right).

Abstract

Single cell analysis is a common tool in many biomedical areas. With novel tools it is possible to acquire datasets consisting of high dimensional expression data for millions of cells in a short time. The analysis of such data is a cumbersome task. A combination of dimensionality reduction and clustering the dimensionality reduced data emerged as the state of the art for classification of different cell types. In this project we base on these techniques and aim to improve the clustering of the dimensionality reduced data.

Context

Currently we employ tSNE for dimensionality reduction and mean shift clustering of the dimensionality reduced data. Mean shift clustering uses density computations (Figure 2a) to partition the visual space (Figure 2c) and finally group similar data points (Figure 2d). The density computation, is dependent on a smoothing kernel, which can be adjusted, but that is usually fixed for the complete dataset. This poses two problems, first the kernel needs to be defined (in our case by the user) and secondly, tSNE produces features of different scales, that cannot necessarily be captured with a fixed kernel.



Assignment

The goal of this project is to investigate ways to improve clustering of dimensionality reduced/embedded data. Different approaches are possible. Keeping classical mean shift clustering we would like to investigate ways to find optimal kernel parameters automatically in a data driven fashion. The next step could be an adaptive mean shift clustering that uses different kernels in different regions of the embedding. Finally, verification and comparison of the clustering results to clustering in the high-dimensional space is of interest.

Requirements

Good programming skills are an advantage. The results shall be implemented in the Cytosplore system which is based on C++/Qt/OpenGL/D3.js. Cuda/OpenCL experience optional.

Contact

For more information please get in contact with:
Thomas Höllt (T.Hollt-1@tudelft.nl) or Anna Vilanova (A.Vilanova@tudelft.nl)