Multistage Optimal Transport evaluated on scRNA-seq Human Bone Marrow data

Vilhelm Conradi & Hugo Nelsson

In order to study e.g. cancer development it is of interest to understand how genes are influencing different cellular processes. One of the most advanced and expensive methods of the present day used to collect gene information is single-cell RNA sequencing (scRNA-seq), which results in highly detailed snapshot data of individual gene counts for each cell. We apply the Multistage Optimal Transport method to human bone marrow data and can observe that for healthy cells, the model captures known biological relationships between cells in the process of hematopoiesis. We are then able to apply the method to datasets containing cells caused by Acute Myeloid Leukemia (AML) and can observe how gene expressions in the cells change when compared to healthy datasets.

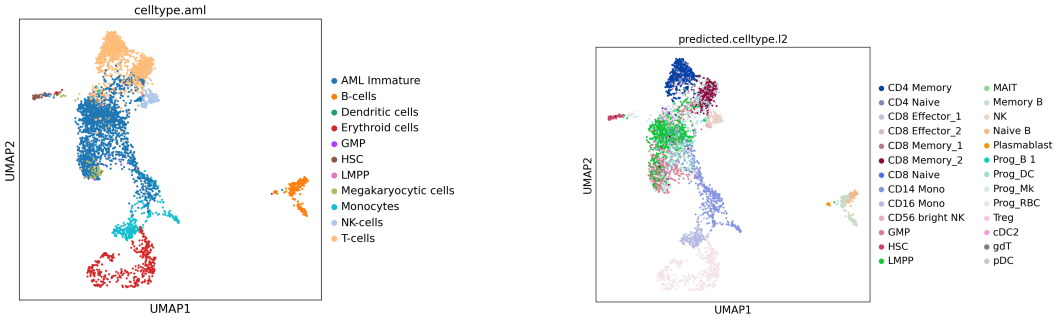
Acute Myeloid Leukemia (AML) is a type of blood cancer where white blood cells develop abnormally during the process of Hematopoiesis, the differentiation of blood cells. This causes uncontrolled crowding in the blood and bone marrow which could potentially be fatal. In order to better understand how to treat this disease, as well as other types of cancer, it is of great interest to understand how cells develop over time and to be able to detect abnormalities in gene information.

To study cells and their biological processes, advanced data collection methods such as single-cell RNA sequencing (scRNA-seq) allow us to look inside single cells in great detail and count the amounts of individual genes present in them. This type of data allows us to make calculations to discover the dynamics of the biological processes in the cells. Due to the nature of the data, we cannot follow specific cells and their development over time. We are therefore left with highly detailed snapshot datasets.

Another type of data that can be extracted from single cells is RNA-velocity, a high dimensional vector that points in the direction of some kind of later cell state. RNA-velocity is the time derivative of the gene expression state and it can be estimated through the proportion of spliced, mature mRNA, and unspliced pre-mRNA. From this information, a vector can be extracted that points in the direction of the next cell state.

In this thesis, we use the mathematical tools from Multistage Optimal Transport (MultistageOT) to describe cell differentiation by finding highly probable transitions between cell states in scRNA-seq snapshot samples of healthy blood marrow from humans and analyze

the gene expression of these trajectories over a notion of time that places the cells in order in the differentiation process, called pseudotime.



(a) aml-annotations. (b) l2-annotations.

Figure 1: UMAP representation of AML-positive sample with annotations corresponding to **(a)** healthy cell types and **(b)** AML-cell types.

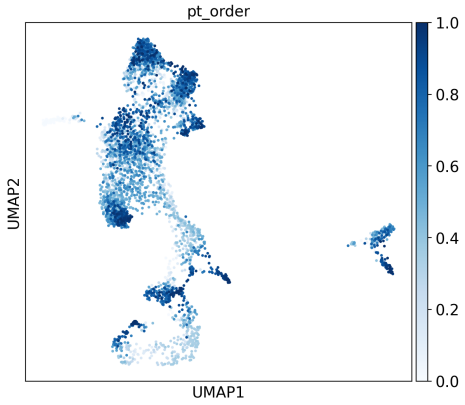
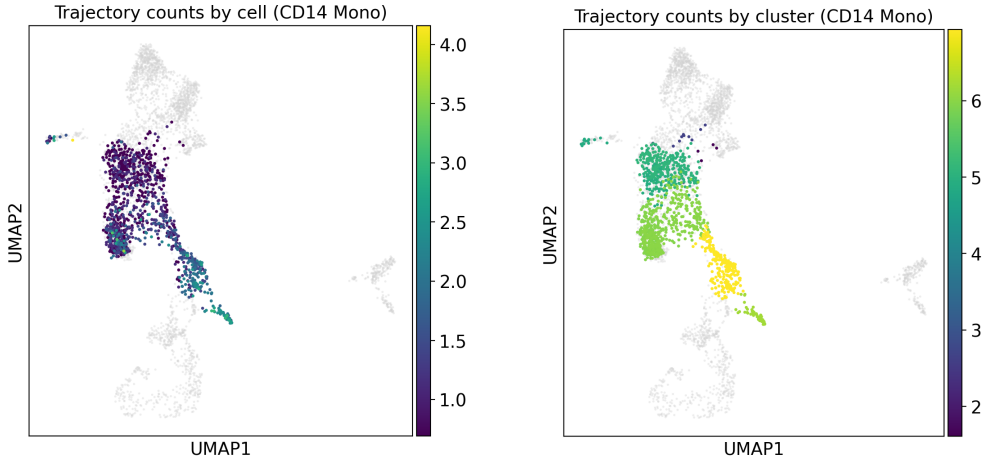


Figure 2: Pseudotime ordering of the AML dataset.

We find that the optimal transport algorithm captures the expected cell development in the datasets and using random walks on the computed cell-cell transition probability matrix we can capture subsets of cells that are highly probable of becoming a specified cell type, as can be seen in Figure 3. In these subsets of cells we can extract the averaged gene expressions in the cells and analyze the change in expression over pseudotime. This pseudotime ordering appears to also be consistent with biological understanding. Incorporating RNA-velocity, we see some improvement in predictions of the final fate of the cell, but the improvement was very slight. Finally, comparing with AML-data, an example shown in Figures 1 and 2 we find gene expressions that differs significantly from what we see in healthy data, suggesting that our method can be used in aiding the study of blood cancer. For the dataset presented here, an example is shown in Figure 4.



(a) Log-scaled visit density per cell.

(b) Log-scaled visit density per cluster.

Figure 3: UMAP representation of AML-positive sample with extracted trajectory subset for CD14$^+$ Monocyte lineage. **(a)** shows, for each cell, the logarithmized number of random walk visits for random walks which terminate in the Monocyte cluster. **(b)** shows the same metric, but averaged over subclusters in the AML dataset.

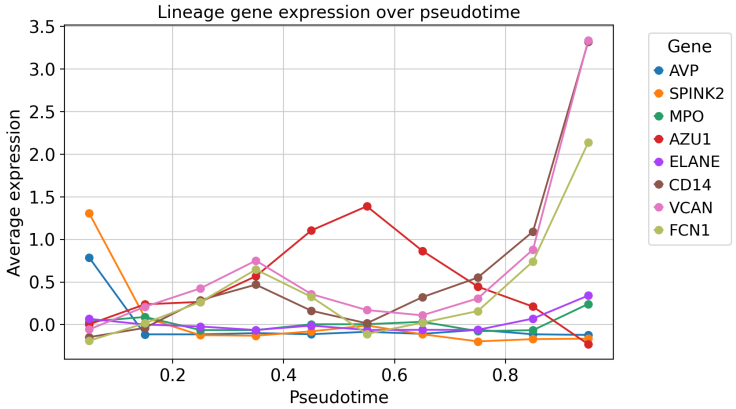


Figure 4: Some average gene expression levels for the AML dataset, plotted against the pseudotime generated by the MultistageOT algorithm.