Trajectory Detection and Analysis for Single Cell Data

John Valainis, Sam Kimmey, Sean Bendall

Abstract

Single cell tools such as mass cytotometry and scRNA-seq are capable of providing extensive high dimensional data sets on dynamic cell processes such as development, cell cycle, activation, and disease progression. Cytoskel is a collection of algorithms for the discovery, visualizaion, and analysis of the stochastic (often branching) trajectories characterizing such dynamical systems. We present trajectory based methods for gating, visualization, and analysis of such high dimensional single cell data in the context of data sets for HSCs. The trajectories enable a denoised, clear, intuitive visualization of the data.

This is a very preliminary write up of the basics of cytoskel, a trajectory inference algorithm. We first describe the steps of the algorithm and the reasoning behind them.

Algorithm

Figure 1 illustrates the basic steps of the cytoskel algorithm.

It begins with appropriately cleaned and transformed single cells data , Figure 1A, which consists of N cells with each cell having p features to be used for calculating distances between cells. Each data point may also have other features or coordinates which do not affect the nearest neighbor calculations but will participate in the later averaging step.

Cytoskel then constructs a k nearest neighbor graph (k-NN graph) with edges connecting cells labeled by distances - dissimilarities between data points. This step captures the local and global structure of the cell cloud. Distances are usuallly the Euclidean distances between cells in the feature space. Here we have a simple 2D feature space with graph edges shown in green, Figure 1B.

A candidate for a trajectory connecting two cells would be the shortest path in the k-NN graph connecting the cells. However, such a shortest path may involve spurious edges (also known as short circuits). The true path between two cells could be a large set of short edges, but a shorter path might be found which utilized a small set of long edges giving an incorrect trajectory. Trajectory inference algorithms have various ways of dealing with this issue. The Cytoskel approach is based on the idea that the edges most likely to not be spurious are the shortest edges. The minimum spanning tree (MST) that cytoskel constructs

is the subgraph of the k-NN graph which connects all the cells and has the smallest possible total edge length. Cytoskel then only allows trajectories to be constructed from the edges which are in the MST. For our illustrative data set, the MST is shown in blue in Figure 1D

Branch or trajectory construction then proceeds by first finding the two cells which are furthest apart as measured along the edges of the MST and constructing the graph path joining them. The next path segment is set of cells and edges from this first existing path to the cell furthest from the existing path. The process is then repeated a specified number of times. The result is a trajectory tree graph linking some subset of the data cells. Figure 1D shows the result of two branchings. Note that 2 branchings create 4 branch ends and two branching points. We note that the branching step is of very low computational cost. When a cytoskel project is created the k-NN graph and the MST are stored. The project can be re-opened later and the branching step rerun with a different number of branches, usually in a few seconds.

The final computational step is averaging using the MST. The data cells are duplicated, and the coordinates of each duplicated cell are replaced by an average over near neighbors out to some distance in the MST. This process is iterated a number of times. We use MST neighbors so that the averaging is over cells most directly related to a given cell.

The averaged cells are referred to as pseudo-cells. The pseudo-cells which are part of the found trajectories are also referred to as trajectory cells. For mass cytometry the main purpose of the averaging is to smooth the trajectories. In the case of sc-RNAseq data or related technologies this averaging is similar to imputation as done in MAGIC [10] and kNNsmoothing [11]. Figure 1E shows the averaged cell cloud and the the MST edges connecting the cells, while Figure 1F shows the averaged trajectory in the context of the original cell cloud.

Note that the averaging will construct imputed cells and smoothed trajectories in the orginal data space. The construction of the graphs could be based, for example, on PCA coordinates or UMAP coordinates. However, we have typically used a subset of the data coordinates to construct the graphs.

The trajectories are represented as linked pseudo-cells. No assumption is made in the algorithm about the starting point of the trajectories. After trajectory construction the user can specify the starting cell, which need not be a branch end. Given the starting cell, pseudotime can be assigned to the trajectory pseudo-cells. Original data cells are associated with their closest pseudo-cell.

Visualization

Two dimensional plots of the pseudo-cell trajectories can be constructed using metric multidimensional scaling (MDS) based on distances between trajectory cells in the original p dimensional feature space. This method creates a two-dimensional layout of the trajectory data points while trying to preserve as closely as possible the p dimensional distances between the data points. The cells in the plot are then colored by feature values of interest.

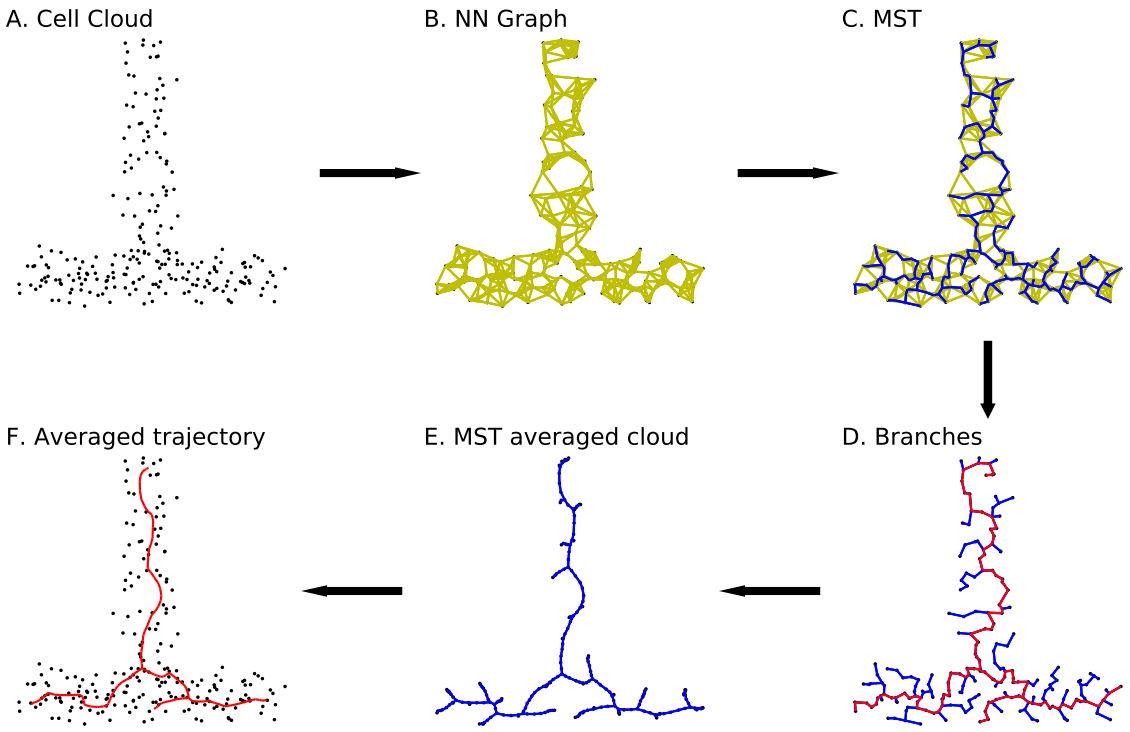

Figure 1: Steps of the basic cytoskel algorithm

This method clearly shows relationships between the trajectory branches and progression of features along the trajectories.

It is also possible to use the trajectory layouts to initialize UMAP. We call this trajectory initialized UMAP, tiUMAP.

3D PCA visualizations and subway plots are also available in an interactive form.

Comparisons

Examples and Benchmarks

References

[1] Bendall, S.C. , Davis, K.L., Amir, E.D., Tadmor M. D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er D. (2014) Single Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. Cell 157:714-725

[2] Bendall, S. C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell 157, 714–725 (2014).

[3] Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 20, 59 (2019).

[4] Cornea N.D., Silver D., Min P. (2006), Curve-Skeleton Properties, Applications and Algorithms, IEEE Transactions on Visualization and Computer Graphics, Jun 2006

[5] Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat. Biotechnol. 34, 637–645 (2016).

[6] Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods 14, 979–982 (2017).

[7] Cannoodt, R. et al. SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. Preprint at bioRxiv https://doi.org/10.1101/079509 (2016).

[8] Moon, K. R. et al. Visualizing structure and transistions in high-dimensional biological data. Nat. Biotechnol. 37, 1482–1492 (2019).

[9] Hou et al. A systematic evaluation of single-cell RNA sequencing methods. Genome Biology (2020) 21:218

[10] Van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, et al. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(3):716–29.

[11] Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for high-throughput single-cell RNA-seq data. bioRxiv. 2017;217737.

[12] Behbehani, G. K. Cell Cycle Analysis by Mass Cytometry. in Cellular Quiescence: Methods and Protocols (ed. Lacorazza, H. D.) 105–124 (Springer New York, 2018).

[13] Kimmey, S. C., Borges, L., Baskar, R. & Bendall, S. C. Parallel analysis of tri-molecular biosynthesis with cell identity and function in single cells. Nat. Commun. 10, 1185 (2019)

[14] Mørup, M.,& Hansen, L. K. (2012). Archetypal analysis for machine learning and data mining. Neurocomputing, 80, 54–63.

[15] Korem Y, Szekely P, Hart Y, Sheftel H, Hausser J,Mayo A, Rothenberg ME, Kalisky T, Alon U Geometry of the Gene Expression Space of Individual Cells Plos Computational Biology , (2015)

[16] Hart Y., Sheftel H., Hausser J., Szekely P., Ben-Moshe N. B., Korem Y., Tendler A., Mayo A. E. & Alon U. Inferring Biological Tasks Using Pareto Analysis of High-Dimensional Data Nature Methods , (2015)

[17] Kimmey, S.C. et al. Single Cell Dissection of Human Lineage Formation: Assessing

Single-Cell Multiplex Protein Regulator Expression and Biosynthesis Dynamics in Differentiating Pluripotentent Stem Cells. 2021.