Extracting Progressions in Biological Data by Identifying Branches in Potential of Heat-diffusion for Affinity-based Transition Embedding

Ngan Vu – Advised by Smita Krishnaswamy

January 2019

1 Background

1.1 Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE)

As high dimensional biological data becomes more easily available, there is a pressing need for visualization tools that reveal the structure and emergent patterns of data in an intuitive form. Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE) [3] is a visualization method that captures both local and global nonlinear structure in data by an information-geometry distance between data points. The advantage of PHATE over many other visualization methods is its ability to reveal branching structures that commonly exist in differentiation systems. For example, PHATE can faithfully visualize the underlying trajectories in a newly generated scRNA-seq dataset of human germ layer differentiation. Here, PHATE reveals a dynamic picture of the main developmental branches in unparalleled detail.

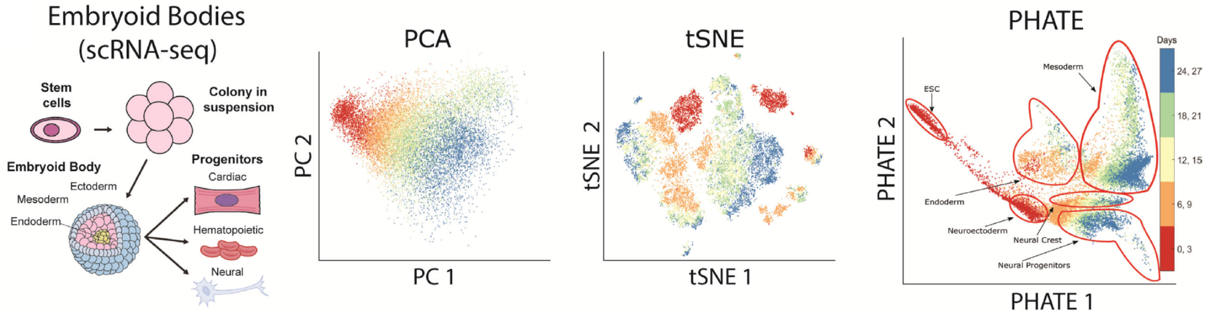

Figure 1: PHATE is better at recovering underlying trajectories in scRNA-seq data compared to other dimensionality reduction methods.

1.2 Extracting information from PHATE

Given a PHATE embedding, specifically one with trajectories, it would be informational to look at the local transitions, progressions, branches or splits in progressions, and end states of progressions. The information on progressions provided by PHATE offers many potential applications. For example, when PHATE is used on patient data, one branch could correspond to how patients of a certain disease progress over time. In this case, it is important to extract data points on a specific branch of interest for further study. Identifying branches also provides an annotation to aid the user in interpreting the PHATE visual.

2 Project Plan

This project focuses on building a tool to extract branch information from a PHATE embedding. Specifically, it will detect branch points (where one trajectory is split into multiple), endpoints (where one trajectory ends), as well as branches (connections between branch points or between a branch point and an end point).

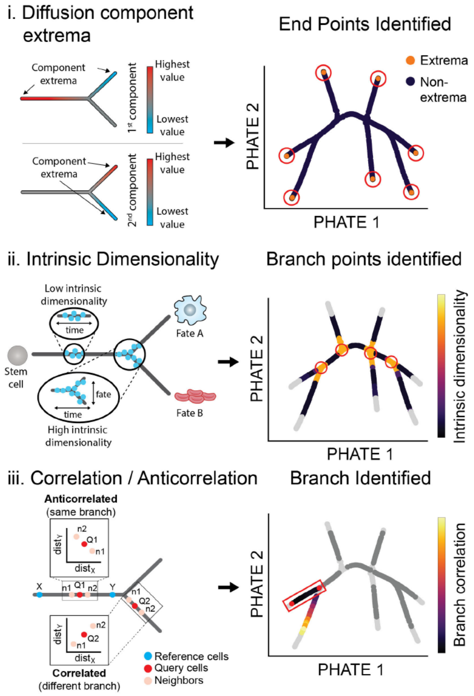

Figure 2: Branch identification using high dimensional PHATE coordinates.

2.1 Identifying Branch Points with Local Intrinsic Dimensionality

Intrinsic dimension of a signal describes how many variables are needed to represent the signal. Branch points lie at the intersections of progressions and therefore they have higher local intrinsic dimensionality than other points: there are multiple directions going to and from these branch points, unlike points that lie on a branch and only have one direction back and forth.

To estimate local intrinsic dimension of the data points, I will implement the algorithm presented in [1]. Then, I will choose data points with high intrinsic dimensionality as branch points.

2.2 Identifying End Points with Centrality

We also care about the end of branches. These points do not have high intrinsic dimensionlity, but we can still detect them using centrality. If we remove points that are not end points, the structure of the graph would change a lot as it becomes disconnected and its overall connectivity reduced. However, removing end points from the graph would not break the graph into multiple part. We say that these end points have low centrality, estimated using the eigenvector centrality measure.

2.3 Identifying Branches

After identifying branch points and endpoints, the remaining points can be assigned to branches between two branch points or between a branch point and endpoint. I will use an approach based on the branch point detection method in [2] that compares the correlation and anticorrelation of neighborhood distances.

3 Deliverables

The deliverables for this project include all final code for the branch detection feature of PHATE, as well as a final project report with results on both artificial data (e.g. an artificial tree) and real data (e.g. scRNA-seq data).

References

[1] Kevin M. Carter and Alfred O. Hero III. Variance reduction with neighborhood smoothing for local intrinsic dimension estimation. 2008.

[2] Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. 2016.

[3] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel Burkhardt, William Chen, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, Natalia B Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing transitions and structure for biological data exploration. 2018.