Lecture Title and Date

Single Cell Applications: Pseudotime Cell Trajectories (3/325)

Objectives of the Lecture

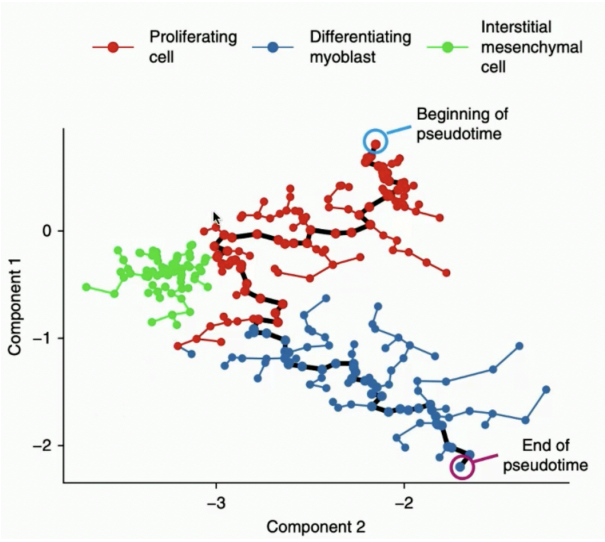
- Understand pseudotime and its contribution to understanding single cell dynamics
- Know the different methods of trajectory inference and the challenges of current computational inferences

Key Concepts and Definitions

Pseudotime: a computational concept that orders cells along trajectories that represent processes like cell differentiation
Cell differentiation: occurs through dynamic developmental processes
Trajectory inference (TI): a computational approach to reconstruct these cellular transitions from single-cell data

Main Content/Topics

Pseudotime (see figure below) in which each dot on the plot represents a single cell and its position along the axis represents the cell's developmental state. The red, blue, and green colors indicate the different states and the black path shows the inferred developmental trajectory. Pseudotime does not measure the real-time but the relative progress.



Monocle 2 is a popular method for TI with 5 steps

1. **Choose genes that define progress**

a. The cell should change along the differential path. Confirm using:

$$x_i = (g_1, g_2, ..., g_d)$$

b. Select high-variance genes based on differential expression patterns. There are different ways:
   i. Compute the variance across cells for each gene using:

$$\mathrm{Var}(g) = \frac{1}{n}\sum_{i=1}^{n}(X_{i,g} - \bar{X}_g)^2$$

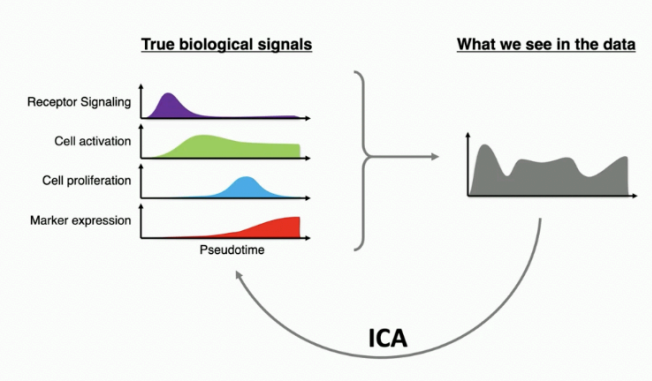
   ii. Use genes that have defined biologically relevant processes such as stem cell markers
   iii. Use statistical tests such as likelihood ratio or T-tests to compare gene expression between different cell groups. Select those with significant p-values

$$H_0 : \mu_1 = \mu_2, \quad H_A : \mu_1 \neq \mu_2$$

2. Reduce data dimensionality (ICA)

a. ICA method used in monocle 2 and is represented by the equation $x = As$ in which x is observed mixed signals, A is the mixing matrix, and s is the original independent source signals.

b. The goal of ICA is to achieve $y = Wx = WAs$ where y is the estimated independent components, and W is the unmixing matrix that is computed by the ICA equation

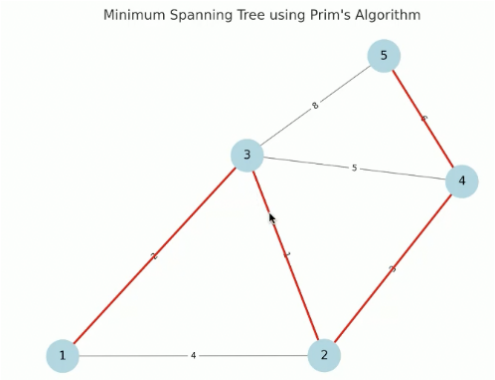


c. ICA is preferred over PCA
   i. PCA finds the direction of maximal variance, and ICA finds the direction of maximal independence. Our goal is to isolate the signals which PCA cannot reliably do

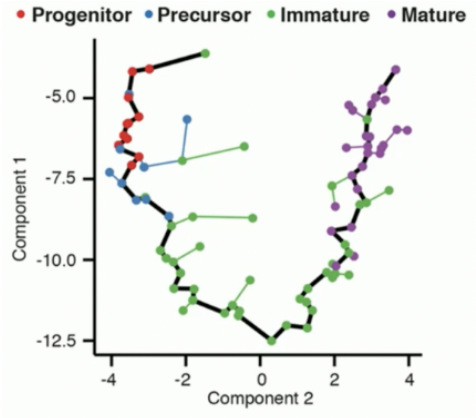
3. Construct minimum spanning trees (MST) on the cells

a. Cells are positioned along a dimensional space at this point and we want to connect these cells or vertices. MST has no cycles and minimizes the sum of edge weights. This is done by:
   i. Computing the pairwise distance in reduced-dimensional space

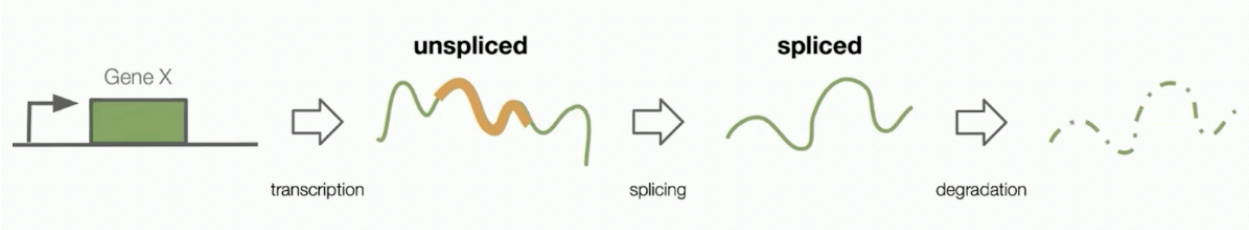
ii. Prim's algorithm. Start at an arbitrary node and select the shortest path to the next node. On and on.



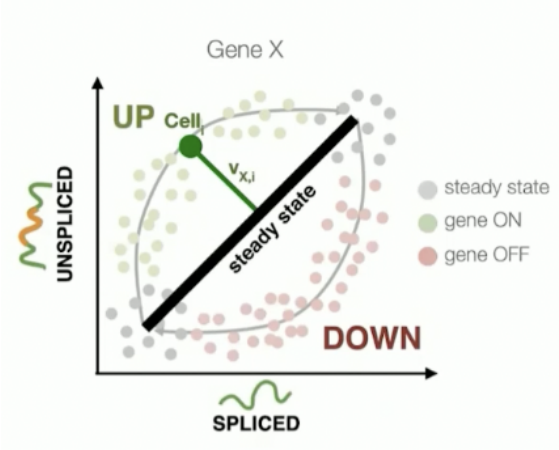
4. **Find the longest path through the MST**
    a. At this point, there is a fully connected tree but MST does not tell us which cells come first. The principle is that the longest continuous path is the best proxy for the differentiation timeline
        i. Cells that are far apart represent the early versus late-stage cells
5. **Order cells along the trajectory**
    a. The longest path we see can be considered the sequence of cells in pseudotime order
    b. Pseudotime value to each cell:
        i. Cells early in the path: undifferentiated
        ii. Cells later in the path: fully differentiated
    c. Ex: developmental trajectory of olfactory neurons in mice



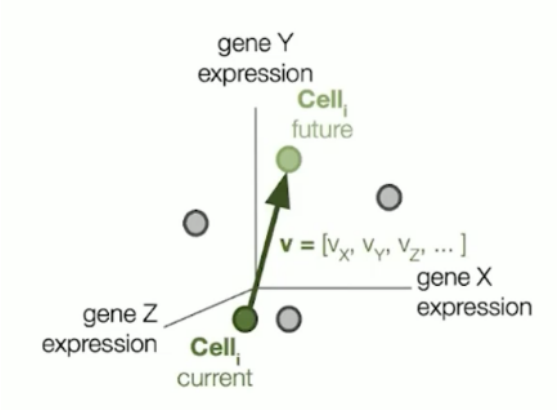
RNA velocity: another popular method in which spliced and unspliced RNA levels indicate changes in gene expression



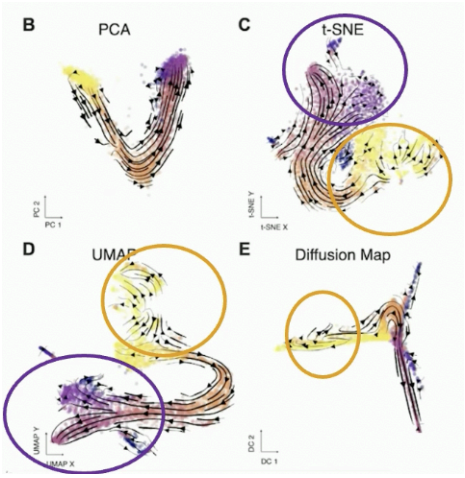
If there are high levels of unspliced, relative to the steady state, then we can infer that more transcription is happening and gene expression is increased.



By looking at the cell's overall velocity, you can predict the future state and direction of gene expression and trajectory



To visualize the velocity trends, you project onto existing 2D embeddings. PCA, t-SNE, UMAP, and diffusion maps are commonly used embedding maps that project different results and interpretations. It is challenging to determine which interpretation is most accurate.



1. Limitations of PCA– only captures linear trends and prioritizes global variance, ignoring local trajectories (RNA velocity is a local property and nonlinear)
2. Limitations of t-SNE– focuses on clustering rather than continuity. It is also stochastic and has no fixed geometric structure, also has no global structure
3. Limitations of UMAP– it is based on topological structures and points representing cells are connected if the distance between them is below a certain threshold. There is a manifold alignment that is distance based but this is not descriptive of directional processes or transcriptional dynamics
4. Limitations of diffusion map– this assumes Markovian diffusion which is a reversible transition but RNA velocity is irreversible. It constructs a global manifold structure which may smooth over local velocity variations

2D embedding using Veloviz (more accurate than the above four)

1. Obtain current and predicted future transcriptional states
2. Compute the composite distances for all cell pairs (novel!) how different gene expression of cell A is to cell B. Also accounts for the comparison between the predicted state A_p and neighbor cell B.

$$D_{A \to X} = -\cos(\theta_{AX}) * \frac{1}{\omega * d_{AX} + 1}$$

3. Identify k neighbors with a minimum composite distance
4. Prune edges with similarity thresholds
5. Visualize with graph based embedding

LIVE-seq is a more expensive but most accurate method of inferring single cell trajectories. Time lapse microscopy enables sequential extraction and profiling of cells. However, it is challenging to scale protocol to large numbers of cells

Zman-seq is another method (middle ground) in which cells are injected with fluorescent labels as a temporal barcode. Can build cellular trajectories with ground truth time-stamps.

Discussion/Comments

Pseudotime orders cells along a trajectory to represent differentiation. Monocle 2, a widely used TI method, involves five steps: selecting genes relevant to progression, reducing dimensionality via Independent Component Analysis (ICA), constructing a minimum spanning tree (MST), identifying the longest path through the MST to infer differentiation order, and assigning pseudotime values to cells.

RNA velocity provides another approach by analyzing spliced and unspliced RNA levels to predict future cellular states. Visualization of RNA velocity trends can be done using 2D embedding methods like PCA, t-SNE, UMAP, and diffusion maps, each with its limitations in capturing local and directional dynamics. Veloviz is noted as a more precise method for visualizing RNA velocity.

LIVE-seq and Zman-seq offer more biologically grounded alternatives, with LIVE-seq providing high accuracy through time-lapse microscopy but at high cost, while Zman-seq uses fluorescent barcodes for time-stamped trajectory reconstruction.

List all suggested reading here and please answer:

Are the readings for the class useful? If so, are the specific subsections useful or would change. If not, are there other references you could suggest? Please suggest one.

There are no suggested readings. Perhaps take a look at this reading below that is referenced in the slides and describes pseudotime and the monocle method

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology, 32(4), 381–386. https://doi.org/10.1038/nbt.2859

References ISL/ESL (if any)

n/a