Trajectory inference analysis

Paulo Czarnewski, ELIXIR-Sweden (NBIS)
Åsa Björklund, ELIXIR-Sweden (NBIS)
Why trajectory inference?

The workflow is dataset-specific:
- Research question
- Batches
- Experimental Conditions
- Sequencing method
- …
What is trajectory inference / pseudotime?

- Cells that differentiate display a **continuous spectrum** of states
  
  *Transcriptional program for activation and differentiation*

- Individual cells will differentiate in an **unsynchronized** manner
  
  *Each cell is a snapshot of differentiation time*

- **Pseudotime** – abstract unit of progress
  
  *Distance between a cell and the start of the trajectory*
Are you sure that you have a developmental trajectory?

Do you have intermediate states?

Do you believe that you have branching in your trajectory?

Be aware, any dataset can be forced into a trajectory without any biological meaning!

First make sure that gene set and dimensionality reduction captures what you expect.
FAST development of Trajectory Inference


Saelens et al (2019) *Nat Biotechnology*
Trajectory Inference Overview

ICA

Independent Component Analysis

A method for decomposing the data
Why ICA?

True biological signals

Receptor Signaling
Cell activation
Cell proliferation
Marker expression

ICA

What we see in the data

Gene A
Gene B
Gene C
Gene D
Gene E

Pseudotime
How does ICA work?

ICA assumes that:

1. The source signals are independent of each other.
2. The values in each source signal have non-Gaussian distributions.
ICA: summary

It is a **LINEAR** method of dimensionality reduction.

ICA is used to **estimate the sources** that compose the data.

The sources are assumed to be **independent** of each other

*This might not be true for single cell*

**Problems with ICA for single cell data:**

- Assumes that the data distribution is non-Gaussian
  *This might not be true for single cell*

- Each component has **equal importance**
  *Unlike PCA where they are sorted by variance*

- ICA cannot identify the actual number of source signals
Diffusion Maps

in brief
How Diffusion Maps work?

Diffusion maps is a non-linear dimensionality reduction algorithm.

The distance between points A and B is defined as the probability of going through the nodes using $K$ steps.

**#2 Steps (1|6):**

$P(1|2) \times P(2|6) = 0.2$

**#3 Steps (1|6):**

$P(1|4) \times P(4|5) \times P(5|6) + P(1|4) \times P(4|7) \times P(7|6)$

- de la Porte et al (2008)
- Coifman et al (2005) PNAS
How Diffusion Maps work?

To transform probabilities to distance, diffusion maps calculates the difference in probabilities to an intermediate point:

$$\text{diff\_dist}(A|B) = P(A|C) - P(C|B)$$

If the $P(A|C) \approx P(C|B)$, $\text{dist}(A|B)$ approaches 0, indicating that $A$ and $B$ are well connected via the intermediate point $C$.

Dimensionality reduction is done by eigenvalue decomposition (like PCA does). The dimensions should be selected by the contribution to each dimension (like PCA).
Diffusion Maps: summary

It is a **NON-LINEAR** method of dimensionality reduction.

The distances between points are measured as **probability** from going from one to another.

The data must present connectivity (transitional cells).
MST

Minimum spanning tree
What is a minimum spanning tree (MST)?

Given a set of points, how do we connect them so that the total sum of all distances is minimized?

Having more transitional cells improves the definition of the tree.

The weights can be the distance in the ICA space or a correlation between cells, etc.

By definition, a MST has no cycles.

So you cannot use MST to define cyclic trajectories (i.e. cell cycle).
Monocle ICA (v1)

Cells represented as points in expression space

- Reduce dimensionality
- Build MST on cells
- Label cells by type
- Order cells in pseudotime via MST

Differentially expressed genes by cell type
Differentially expressed genes across pseudotime
Gene expression clusters and trends

Graph Abstraction and principal tree learning

i.e. DDRTree and others
The limitation of MST

Trajectory construction using MST is highly dependent on single data points

Monocle DDRTree (v2)

Initial dimension reduction
PCA (LLE, DM, ...)

Repeat until both cell trajectory and cell positions are stable under the optimization function

Update map back to high-dimensional space

Guess initial cell trajectory on centroids

Update cell positions based on current trajectory

Distance from root defines pseudotime

Branches define fate decisions

Many methods derived from RGE idea

The methods differ on the dimensionality reduction used, the clustering method or the way the tree is constructed.

Monocle UMAP (v3)

**Preprocess**
- High dimension noisy scRNA-seq dataset
- Normalization + PCA

**Non-linear dimension reduction**
- tSNE
- UMAP

**Learn graph**
- Partition clusters of cells strongly connecting components
  - DDRTree (principal tree learning + dimension reduction)
  - SimplePPT (principal tree learning in the same dimension)
  - L1-graph (general principal graph learning in the same dimension)

**Clustering** (Louvain clustering)

**Steps**
- Pseudotime assignment
- Graph based differential analysis
RNA velocity

gene expression trajectory
How does RNA velocity work?

It uses the proportion spliced/unspliced reads to predict the future state of a cell

RNA velocity allows a biologically-driven identification of cell transcriptional trajectories:

Defines **start**, **ends** and **bifurcations**

The position of the **spliced** is represented by the **arrow-head**

How does RNA velocity work?

Cyclic trajectories are also captured

Differential expression
TradeSeq: differential expression

**Differential Expression Tests**

<table>
<thead>
<tr>
<th>Lineages</th>
<th>association Test</th>
<th>startVsEnd Test</th>
<th>diffEnd Test</th>
<th>pattern Test</th>
<th>earlyDE Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Wrap-up
Final Considerations

• In reality, distance in multidimensional space reflects difference in transcriptional landscape, not actual time.

• Necessary to have a **continuum** of states among your cells
  
  *Will not work well with 2 distinct clusters.*

• May work with **single time-point** if ongoing differentiation process
  
  *It is better to have multiple experimental time points.*

⚠️ Be aware, any dataset can be forced into a trajectory without any biological meaning!

⚠️ First make sure that gene set and dimensionality reduction captures what you expect.
Which method should I use?

http://guidelines.dynverse.org

Which method should I use?

### a) Method

<table>
<thead>
<tr>
<th>Method</th>
<th>Priors required</th>
<th>Wrapper type</th>
<th>Platform</th>
<th>Topology inference</th>
<th>Cycle</th>
<th>Linear</th>
<th>Bilateral</th>
<th>Multifurcation</th>
<th>Tree</th>
<th>Connected</th>
<th>Disconnected</th>
<th>Inferrable trajectory types</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAGA</td>
<td>×</td>
<td>Direct</td>
<td>Python</td>
<td>Free</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RaceID / StemID</td>
<td>Proj</td>
<td></td>
<td>R</td>
<td>Free</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLICER</td>
<td>×</td>
<td>Cell</td>
<td>R</td>
<td>Free</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### b) Summary

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Accuracy</th>
<th>Scalability</th>
<th>Stability</th>
<th>Usability</th>
<th>Aggregated scores per experiment</th>
</tr>
</thead>
</table>

### Tree methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Priors required</th>
<th>Wrapper type</th>
<th>Platform</th>
<th>Topology inference</th>
<th>Cycle</th>
<th>Linear</th>
<th>Bilateral</th>
<th>Multifurcation</th>
<th>Tree</th>
<th>Connected</th>
<th>Disconnected</th>
<th>Inferrable trajectory types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slingshot</td>
<td></td>
<td>Direct</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAGA Tree</td>
<td>×</td>
<td>Direct</td>
<td>Python</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MST</td>
<td>Proj</td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pCreode</td>
<td>Proj</td>
<td></td>
<td>Python</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCUBA</td>
<td>Clust</td>
<td></td>
<td>Python</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monocle DDRTree</td>
<td>Cell</td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monocle ICA</td>
<td>×</td>
<td>Cell</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cellTree maptpx</td>
<td>Cell</td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLICE</td>
<td>Direct</td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cellTree VEM</td>
<td>Cell</td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELIPIGraph</td>
<td>Direct</td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SineCell</td>
<td>Cell</td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URD</td>
<td>×</td>
<td>Direct</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellTrails</td>
<td>Cell</td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mpath</td>
<td>×</td>
<td>Cluster</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellRouter</td>
<td>×</td>
<td>Cell</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>