Optimizing single-cell spatiotemporal delay variations to identify key features driving progression

Komlan Atitey and Benedict Anchang
NIH-NIEHS Biostatistics and Computational Biology Branch

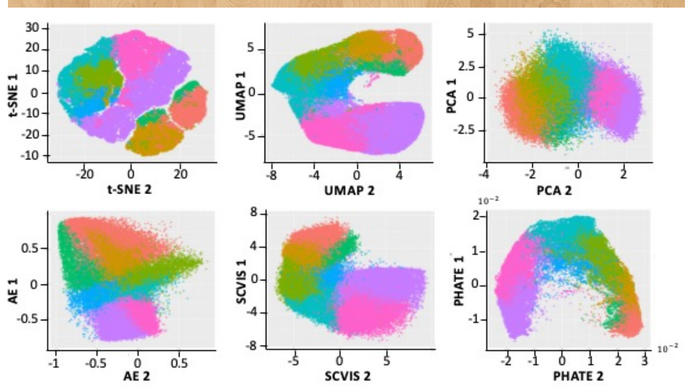
National Institute of Environmental Health Sciences

ISMB ECCB 2023 LYON FRANCE July 23-27 SCB — Lyon, France July 23-27, 2023

Abstract

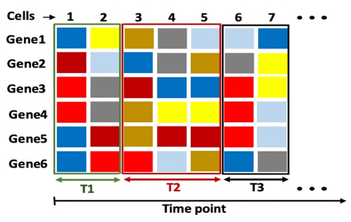
The high-dimensional nature of single-cell data poses challenges for visualization, analysis, and interpretation. Developing accurate and robust models for single-cell spatiotemporal trajectory analysis remains a significant challenge. In this study, we propose a computational framework called Time Order Structure Learning (TOSL) to quantify the spatiotemporal dependency of cells in a low-dimensional space resulting from temporal data reduction or imaging analysis. We utilize TOSL to evaluate the performance of six data reduction methods (DRMs) applied to visualize three dynamic biological processes: epithelial-mesenchymal transition (EMT), spermatogenesis, and stem cell reprogramming. The results obtained using TOSL indicate that none of the six DRMs demonstrate a significant preservation of the evolutionary dynamics of the various cell types. TOSL models the global delay time required to achieve the stationarity of cell state transitions. This modeling approach can be utilized to identify therapeutic targets that drive early deviations from normal development, leading to diseases such as cancer. In future research, our objectives are twofold: (1) identify genes and cells responsible for triggering the stationarity of state transitions during developmental biology and (2) develop an improved data reduction method that preserves global cellular spatiotemporal dynamics.

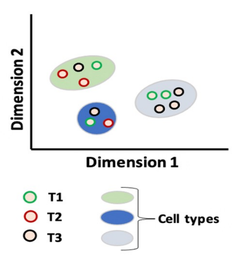
Dimensionality reduction methods



Example of dimension reduction of the CyTOF EMT dataset with six methods; tSNE, UMAP, PCA, Autoencoder, SCVIS, and PHATE presenting different projections with different scales from the lowest (PHATE) to the highest (t-SNE) which can confound visualization.

Features for good visualization





- ✓ Single-cell RNA-seq gene expression data collected at multiple time points (T1, T2, T3), is clustered and projected to a low dimensional space.
- ✓ The high-dimensional data would be clustered based on multiple time points at which data were collected, demonstrating the dynamic relationship among cells in the projected space.

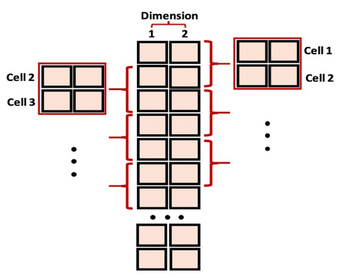
Computational framework

➢ **TOSL assesses the accuracy of interpreting the correlation and causation of a visualization technique. It applies multi-state Markov processes to evaluate the time dependency of cells in a temporal reduced space (2D).**

➢ **A TOSL uses Markov chain model with community structure to describe the possible states of individual cells during consecutive time steps in the projected space**

Convert non-square to square matrix

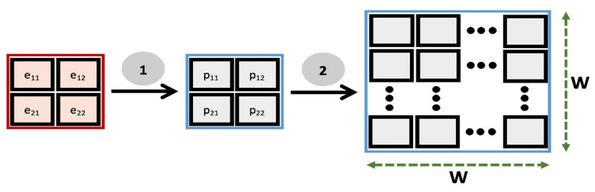
Capturing transition between time points



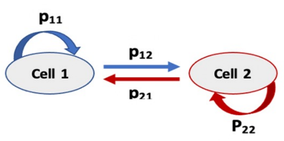
- ✓ A large projected non-square dataset is partitioning into consecutive 2x2 square matrices to capture transition between time points.

Markov process

Estimating cell state probabilities and computing expected transitions to generate a W×W transition matrix



- ✓ Markov chain estimation evaluates the dynamic evolution of cells based on a mixture of cells in different states with varying state compositions over time.
- ✓ The distribution of cell future states is independent on past states when the present state is known.

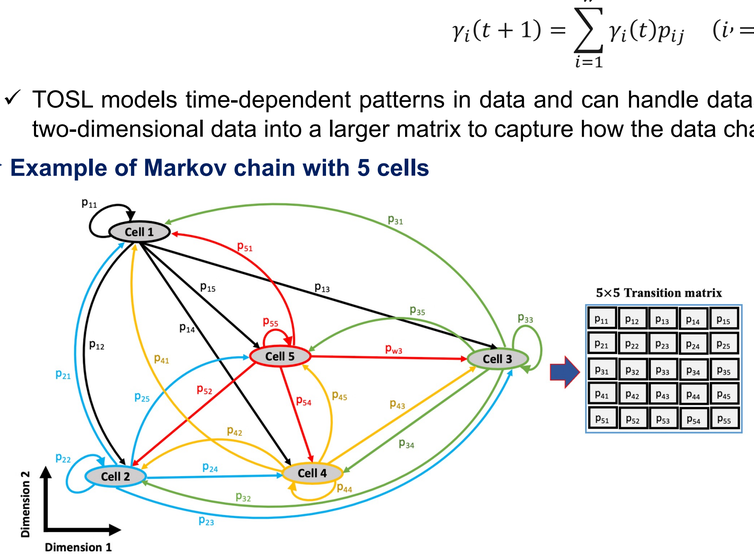


- ✓ The evolution of a random time-dependent process $x(t)$, assuming a probability of transforming from state *i* to state *j* at time *t+1* of p_{ij} and a probability of being in state *i* at time *t* of $\gamma_i(t)$, is captured by:

$$\gamma_i(t+1) = \sum_{i=1}^{w} \gamma_i(t) p_{ij} \quad (i = 1, 2 \cdots w)$$

- ✓ TOSL models time-dependent patterns in data and can handle datasets with any data point ordering. It expands two-dimensional data into a larger matrix to capture how the data changes over time.

➢ **Example of Markov chain with 5 cells**



- ✓ Example of a Markov chain with 5 cells at 5 different states with selected states transition. The state transition probability matrix of the Markov chain gives the probabilities of transitioning from one state to another in a single time unit.

Identifies temporal structure in $W \times W$

Time dependencies in the $W \times W$ square matrices

➢ **Compute the expected number of times a cell has visited other states in the Markov chain**

$$C_{i,j} = E\left\{\sum_{n=0}^{\infty} I\{W_n = j / W_0 = i\}\right\} = \sum_{n=0}^{\infty} P_{ij}^{n}$$

C_{ij} is the expected number of times (over all time steps) that the Markov chain visit state $j \in G$ for $G = \{1, \cdots, W\}$ given the initial state $i \in G$.

$C = (C_{ij})(C_{ij})$ is a $W \times W$ matrix

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1W} \\ \vdots & \ddots & \vdots \\ c_{u-1w} & \cdots & c_{u-1W} \end{pmatrix}$$

➢ **Weight each row of the matrix C by the proportion of associated label of cell types in the reduced data**

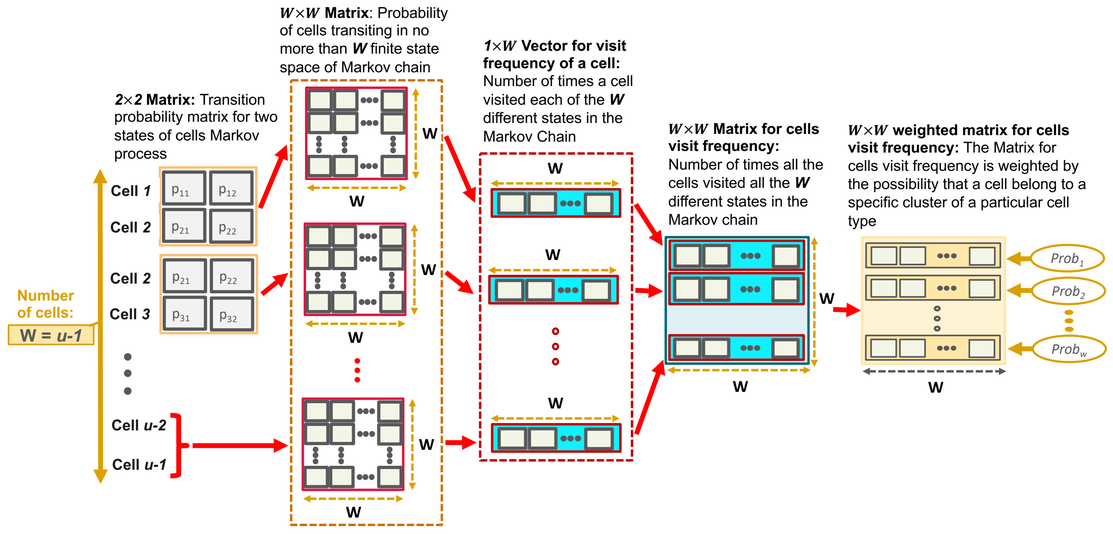
$$C = \frac{Num_k}{|C_{type}|} \times \left[\left(C_{k_{i,j}}\right)_{i=1}^{u-1}\right]_{j=1}^{W}$$

- ✓ $|C_{type}|$ refers to the total amount of cells in C_{type}.
- ✓ $C_{type} = (k_j)_{j=1}^{u-1}$ for $k \in C_{type}$, and K is the finite number of distinct cell types.
- ✓ $Num_k = |k_j|$, the total number of a particular cell type k in the label vector C_{type}.

➢ **Estimate the variability of data around the expectation**

$$V_j = \sum_{i=1}^{u-1} \frac{(C_{ij} - \mu_j)^2}{u - 2}$$

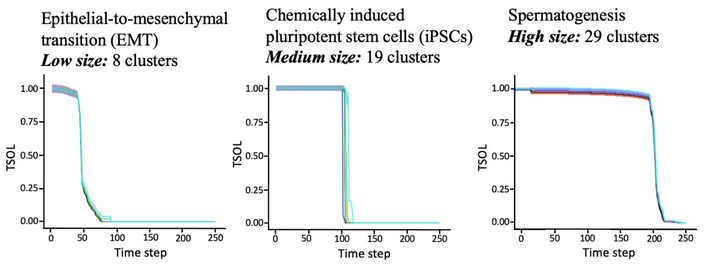
- ✓ The TOSL evaluates the stationary distribution of a Markov chain of W steps, where the variance converges to equilibrium. The sooner the variance (V_j) converge to equilibrium, the less time-dependent the outcomes of Dimensionality Reduction Methods (DRMs) are.



➢ **TOSL evaluates the stationary distribution of a Markov chain of W steps, where the variance converges to equilibrium**

$$TOSL = \frac{n_\theta}{N_{step}}$$

- ✓ n_θ is the number of steps in each interval τ which display a variance different to the equilibrium (the variance does not converge to equilibrium)
- ✓ N_{step} is the total number of steps in each interval τ



- ✓ TOSL evaluates the DRM performance based on data complexity measured by the dimension size and number of cell types.
- ✓ Changes in data complexity and cellular dynamics are associated with changes in DRM performance, with an increase in delay to stationarity observed with increased complexity.

	tsne	umap	pca	ae	scvis	phate
TOSL				✓		✓

	tsne	umap	pca	ae	scvis	phate
TOSL					✓	✓

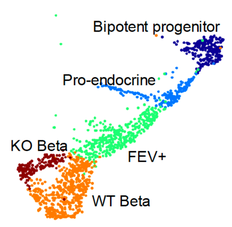
	tsne	umap	pca	ae	scvis	phate
TOSL						✓

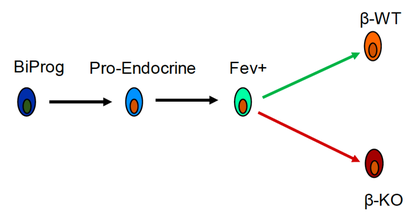
- ✓ The performance of all DRM seem to be associated with changes in the complexity as well as cellular dynamics. An increase in delay to stationarity can be observed with increase in data complexity.
- ✓ The observed delay to stationarity with increase in complexity warrants further investigations.
- ✓ For example, preliminary results suggests that the identification of the variations in delay time to stationarity allows for the identification of therapeutic targets that may be useful to study deviations from normal development resulting in diseases like cancer.

Enhance TOSL to understand mechanisms during β-cell development

- ✓ Diabetes is caused by insufficient insulin secretion due to **β**-cell loss(Type1), or insulin resistance (Type2)

Beta cell dysfunction and insulin resistance are inherently complex with their interrelation for triggering the pathogenesis of diabetes also somewhat undefined. Both pathogenic states induce hyperglycemia and therefore increase insulin demand. Beta cell dysfunction results from inadequate glucose sensing to stimulate insulin secretion therefore elevated glucose concentrations prevail.





- ✓ The restoration of normal functioning of **β**-cell might offer a cure
- ✓ Understanding mechanisms during **β**-cell development is key
- ✓ The goal is to identify key genes that are responsible for KO β-cells differentiating away from WT β-cells, as well as the branching point using scRNA-seq

Conclusion

- ✓ Optimizing the variability and integrating with feature selection approaches through TOSL could provide useful insights on the major timepoints and state transitions that drive the underlying biological process. Also, the significant lack of variability between current DRMs in the non-stationarity part of the TOSL Markov chain shows that current DRMs are not optimized to fully account for the complex underlying dynamic structure between different cell types during a given biological process.
- ✓ These motivate our current research in the direction to enhance the TOSL framework to identify genes and cells which trigger the stationarity of state transitions of cells during the developmental biology accounting for gene-gene interaction, and cell-cell communication. It would be useful to identify key genes that are responsible for KO β-cells differentiating away from WT β-cells.