Progress Summarization

Jiyuan Jiao

Feb 2025

1 Background

The aim of this work is to adapt the approach from 'Universal Prediction of Cell-Cycle Position Using Transfer Learning' (***tricycle***) [1] to the setting of sc-DNA methylation data, leveraging periodic methylation patterns instead of gene expression for cell stage predictions.

The ***tricycle*** method leverages principal component analysis of cell-cycle genes and transfer learning by projecting new data into a pre-defined reference embedding based on a proliferative dataset. This approach enables robust and scalable cell-cycle inference, overcoming dataset-specific biases and technical variations.

In ***tricycle***, the reference embedding for Tricycle is primarily built using single-cell RNA-seq data where cells are actively cycling, so that cell cycle is the dominant source of variation and ensuring a clear periodic pattern in gene expression. In our work, the sc-DNA methylation data is at the CpG level, represented as a binary variable indicating whether each CpG site is methylated. We aimed to identify a dimension reduction method that transforms the binary data into a latent space with a circular pattern so that helps with building up a good reference embedding.

It is hard to have a real-world CpG-level sc-DNA methylation data. Thus we integrated a dataset containing CpG site positions with a dataset of gene positions to determine the number of CpG sites located within each gene. Then we simulated 1000 cells with periodic methylation proportion for CpGs inside gene region. We constructed a distance metric for comparing two binary vectors (two cells), and conduct principal coordinate analysis on generate distance matrix obtained from comparing each pair of binary vectors (cells).

2 Setup

Let Y be a matrix, where $y_{i,j}$ represents methylation proportion of cell $j \leq M$ and row represents genes $i \leq N$, such that $j \leq M$ and $i \leq N$.

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1M} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2M} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \vdots & y_{iM} \\ \vdots & \vdots & \ldots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{Nj} & \cdots & y_{NM} \end{bmatrix}$$

Let the $x_{i,j}$ be pseudo-time range from $(0, 2\pi)$ represents unknown cell-cycle position for gene j at cell i, A represents amplitude and L_j represents gene-specific location of peak methylation proportion. Now assume each gene having 1000 time points (1000 cells). The simulated methylation proportion for gene i at time cell j (time j) is given by the following equation:

$$y_{i,j}(x_{i,j}) = A \times \cos(x_{i,j} - L_j) + \epsilon_{i,j} + 0.5$$

, where $\epsilon_{i,j} \sim N(0, 0.01)$ is noise ($\forall i$ and a, j and b, $\epsilon_{i,j}$ independent of $\epsilon_{a,b}$) and 0.5 is added to make sure that simulated methylation proportion falls between [0,1]

Note that in reality methylation proportion is impossible to exceed 1 or fall below 0. Thus we force any value exceeding 1 to 1, and any value below 0 to 0. Here is the plot of some genes of the simulated data result with some cells marked out.

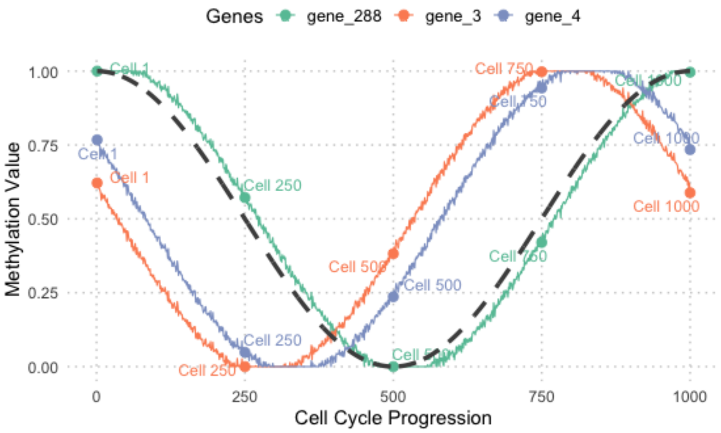

Figure 1: Plot methylation proportion of simulated data

Then we expand gene-level methylation proportion data to CpG level methylation data E, where $e_{i,j} = I($ whether CpG_i of some gene in $cell_j$ is methylated $)$. Previously we have integrated a dataset containing CpG site positions with a dataset of gene positions to determine the number of CpG sites located within each gene. Thus for each $y_{i,j}$ in the Y, it is going to be expanded to a vector: $\mathbf{y_{i,j}} = (e^i_{k,j}, e^i_{k+1,j}, \ldots, e^i_{s,j})^T$, where s is number of CpG sites inside $gene_i$. Note that for each $\mathbf{y_{i,j}}$, we are going to make first k CpG sites to be methylated so that

$$\frac{\sum_{k=1}^{s} I(\text{CpG}_k \text{of gene}_i \text{cell}_j \text{is methylated})}{s} \approx y_{i,j}$$

For any $i, j \leq M$, we define dissimilarity between a pair of cells by $d_{i,j} = 1 - |cor_{tetra}(e_i, e_j)|$, where cor_{tetra} is tetrachoric correlation between two binary vectors. After computing any pairs of cells in E, we can have a matrix symmetric $D_{M \times M}$. Then we conduct principal coordinate analysis (Multi-Dimensional Scaling) on D. The squared dissimilarity matrix $D^{(2)}$ is D where each entry is squared:

$$D^{(2)}_{ij} = D^2_{ij}.$$

We then apply double centering to obtain the Gram matrix B:

$$B = -\frac{1}{2} H D^{(2)} H,$$

, where H is the centering matrix defined as: $H = I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T$. I_m is the $m \times m$ identity matrix and $\mathbf{1}$ is an m-dimensional column vector of ones. Since B is symmetric, we perform eigen decomposition:

$$B = U \Lambda U^T,$$

where U is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues. The principal coordinate representation is obtained by:

$$\Omega = U \Lambda^{\frac{1}{2}}.$$

This provides the m-by-k matrix Ω, where the rows correspond to the coordinates of the objects in the derived Euclidean space. The columns of X correspond to principal coordinates, with the first few coordinates capturing the most variance in the data. The dimensionality k is typically chosen based on the largest positive eigenvalues of B. Here $K = 2$ is choosen since periodic functions can be decomposed into orthonormal functions cos and sin, and principal components of that gives two eigenvectors $(cos(x), sin(x))$ that contributes most to variance in the D.

3 Results

Here we are going to provide plot E's PCA embedding space with color gradient on some gene-specific methylation data from Y or cell index order.

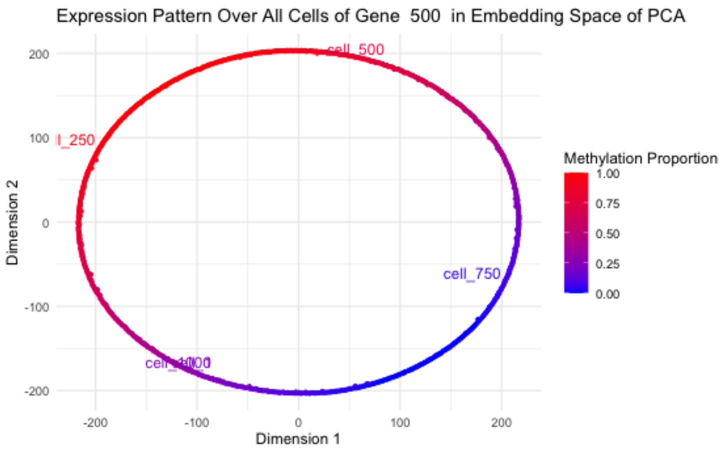

Figure 2: PCA Embedding of E Colored by Gene 500 Methylation Proportion from Y

Please check the R Markdown file (lines 290–335) for the dynamic plot of the PCA embedding space, where genes vary, with a color gradient representing gene-specific methylation proportions across different genes.

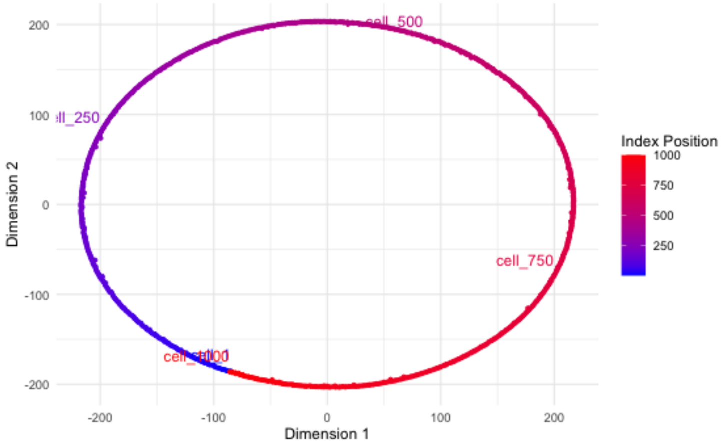

Figure 3: PCA Embedding of E Colored by Cell Indices Order from Y

Next, we present the PCoA (Multi-Dimensional Scaling) embedding space Ω of D, with a color gradient representing either gene-specific methylation data from Y or cell index order. For the PCoA plots, cells 1 to 500 form a perfect circle, while cells 500 to 1000 form another perfect circle. Therefore, for each plot displaying the color gradient based on gene-specific methylation proportion or cell index, we will provide separate plots for cells 1 to 500 and cells 500 to

1000.

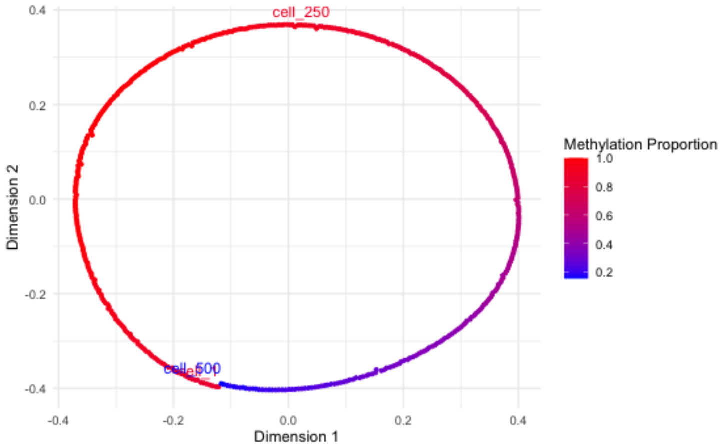

Figure 4: PCoA Embedding of D Colored by Gene 500 Methylation Proportion from Y for cell 1 to 500

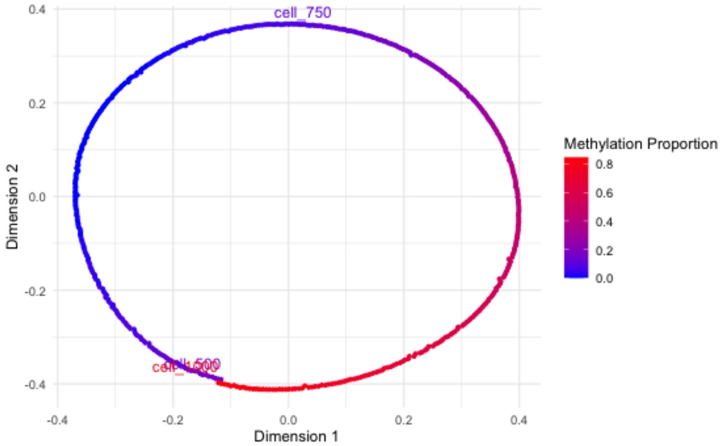

Figure 5: PCoA Embedding of D Colored by Gene 500 Methylation Proportion from Y for cell 500 to 1000

Please check the R Markdown file (lines 436–480) for the dynamic plot of the PCoA embedding space, where genes vary, with a color gradient representing gene-specific methylation proportions across different genes.

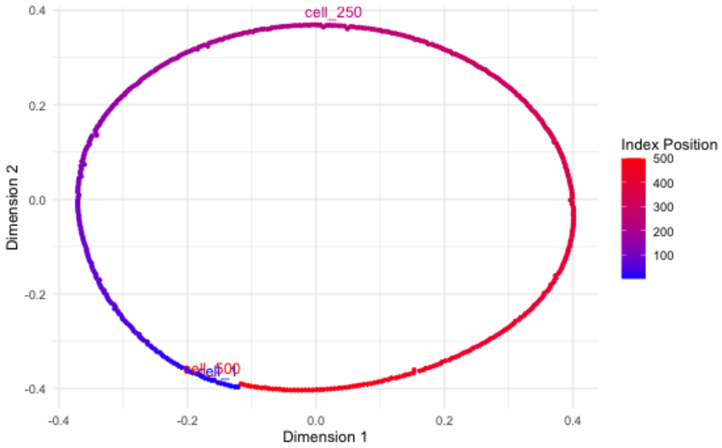

Figure 6: PCoA Embedding of D Colored by Cell Indices Order from Y for cell 1 to 500

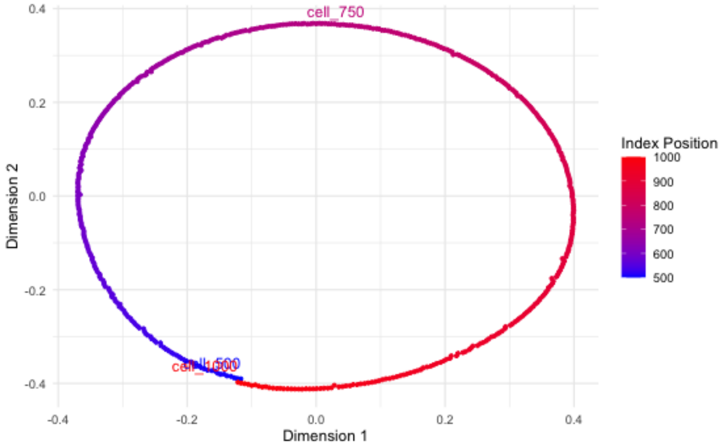

Figure 7: PCoA Embedding of D Colored by Cell Indices Order from Y for cell 500 to 1000

References

[1] Shijie C. Zheng et al. "Universal prediction of cell-cycle position using transfer learning". In: *Genome Biology* 23.1 (2022), p. 41. DOI: `10.1186/s13059-021-02581-y`. URL: `https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02581-y`.