Cell Cycle Time Inference from RNA sequences

Francesco Bollero Federica Padovano Mattia Mariantoni
Pau Mulet-Arabí Luca Rossi

Introduction

Advances in omics technologies make it possible to study cellular dynamics, providing accurate information on which genes encoded in our DNA are turned on or off in the continuously changing transcriptome. In particular, RNA-seq provides a powerful means to analyze molecular mechanisms underlying cell-state transitions, leading to an unprecedented opportunity to reveal latent biological processes. As a result, the reconstruction of cell development processes from RNA sequences has attracted much attention in recent years. Still, it remains a challenge due to the heterogeneous nature of the processes. The underlying idea in most methods proposed is that there is a biological process responsible for the main variations in the data. Then the goal is to infer the trajectory of that process in the gene expression space so that its effects can be removed. It allows the delineation of other cell subpopulations, which can be crucial to studying tumor evolution.

Objectives

The aim of the project is to make a pseudotemporal reconstruction of the cells processes from the gene expression profiles obtained by RNA sequences.

Goals

1. Improve phase estimation accuracy in already labelled data. [1]
2. Pseudo-time inference of the cell cycle process in unlabelled data. [2]

Methods

1. Autoencoder
2. Gaussian Process Latent Variable Model
3. Autoencoder with Residual connections

Autoencoder

Autoencoders are an unsupervised learning technique in which neural networks are used for the task of representation learning. Specifically, a neural network architecture will be designed so that we have a bottleneck in the network which forces a compressed knowledge representation of the original input.

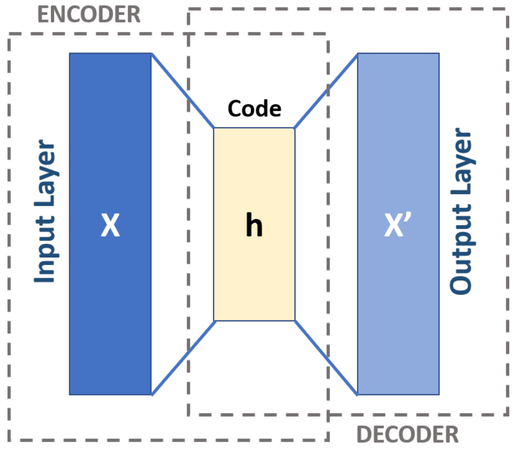

Figure 4: As shown in the image, the autoencoder is essentially composed of two distinct parts: the encoder part and the decoder part. Each of the two parts is a neural network with a varying number of layers and neurons.

Residual Autoencoder

Skip unit connections $\rightarrow \begin{cases} \text{avoid degradation} \\ \text{train deeper} \end{cases}$

Relu in the encoder $\rightarrow$ Avoid vanishing gradients

$sin(x)^2 + x$ in the decoder $\rightarrow$ no bad local minima

This changes lead to a more robust model with still excellent performance.

Graphical Results

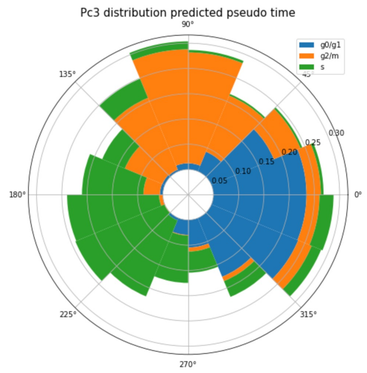

Figure 1: The histogram presents the distribution of predicted pseudo times for pc3 dataset using GPLVM. We coloured the distribution to make clear which are the real phases of the cells. GPLVM was the model that best fit the pc3 dataset.

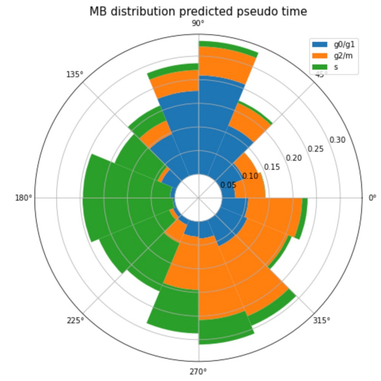

Figure 2: The histogram presents the distribution of predicted pseudo times for mb dataset using cyclum. We coloured the distribution to make clear which are the real phases of the cells. Cyclum was the model that best fit the mb dataset.

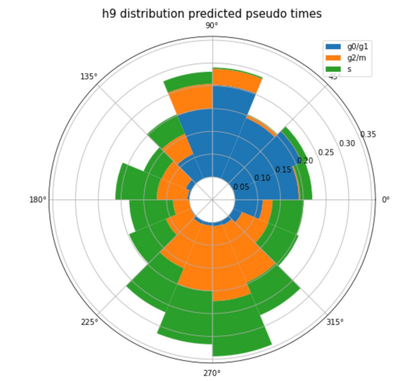

Figure 3: The histogram presents the distribution of predicted pseudo times for h9 dataset using residual autoencoder. We coloured the distribution to make clear which are the real phases of the cells. Residual autoencoder was the model that best fit the h9 dataset.

GPLVM

Gaussian Processes Latent Variable models are Bayesian nonparametric models for complex dimensionality reduction in high dimensional spaces:

Idea: Learn a distribution over functions of the form:

$$y = f(x) + \epsilon$$

where y would be the genes and x the phase.

1. Set a prior distribution
   (a) Latent: $p(X) = \prod_{n=1}^{N} \mathcal{N}(x_n, I)$
   (b) Functions: $p(f, X) = \mathcal{N}(\mu(X), K(X, X))$
2. Compute the posterior

$$p(y^* \mid \mathcal{D}) = \int p(y^* \mid x, f_x) p(f_x \mid x) p(x) dx df_x$$

(We perform this step with Variational Inference)

UMAP & Kernels

A crucial point in GPLVM is to set the priors properly.

Periodic kernel

In the function space, to favor the periodicity between the phase and the gene counter we use a periodic kernel.

$$k(x_a, x_b) = \sigma^2 \exp\left(-\frac{2}{\ell^2} \sin^2\left(\pi \frac{|x_a - x_b|}{p}\right)\right)$$

Uniform Manifold Projection

UMAP is a novel manifold learning technique for general non-linear dimensionality reduction. It is constructed from a theoretical framework based in Riemannian geometry and it allows to perform embeddings to non-Euclidean spaces. We used this model to set a prior over the latent space by performing a non-linear projection from the gene expressions to the unit circle.

References

[1] Andrew N. McDavid and Lucas Dennis. Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLoS Computational Biology*, 10, 2014.

[2] Henry E. Miller, Aparna Gorthi, Nicklas Bassani, Liesl A. Lawrence, Brian S. Iskra, and Alexander J. R. Bishop. Reconstruction of ewing sarcoma developmental context from mass-scale transcriptomics reveals characteristics of ewsr1-fli1 permissibility. *Cancers*, 12, 2020.

Conclusions

We reproduced the autoencoder in McDavid dataset[1] showing great accuracies and used this model to label the CHLA9 dataset. Additionally we tried to improve the raw autoencoder by implementing a deeper network with residual connections and changing the activation functions. These changes improved the robustness while maintaining the excellent results of the original model. Finally we proposed a different approach using a Gaussian Process Latent Variable model that led to an overall improvement in the accuracy.