Pseudotime Diffusion

Jacob D. Moss
Computer Lab
University of Cambridge
Cambridge, UK
jm2311@cam.ac.uk

Jeremy L. England
GSK
25 Basel St.
Petach, Israel

Pietro Lió
Computer Lab
University of Cambridge
Cambridge, UK

Abstract

Analysis of whole-genome sequencing data has been outpaced by the experimental techniques that generate those datasets. The computational challenges associated with these analyses typically make machine learning methods more suitable over more conventional methods like dimensionality reduction, which limit the information obtainable from a dataset. In this paper, we focus on the biophysical model of RNA velocity, which yields meaningful insights into the functional trajectories of individual cells. There are many downstream applications, such as the identification of key genes driving a disease pathway. We improve the dynamical model by relaxing unrealistic assumptions and using the resulting generative process to train a diffusion model to compute pseudotime. Our probabilistic model is able to quantify the uncertainty in its pseudotime predictions. Finally, we demonstrate the efficacy of our model on a series of benchmark tasks.

1 Introduction

Single-cell sequencing is a family of techniques which provide genetic information at the resolution of single cells. This enables understanding cellular function at a granular level, involving data types such as gene expression counts, structural accessibility of DNA regions, or raw nucleotide sequences. The downstream uses for such data are abundant, including the reading of DNA for cells which are difficult to culture in a lab as well as investigating the effect of mutations on cell differentiation trajectories [Jovic et al., 2022]. Single-cell experiments yield a collection of snapshots of cellular states, making the task of aligning the cells meaningfully crucial to extracting information from such datasets. This task is called pseudotime inference and is the primary focus of this paper.

Suppose that some technical accident has shuffled the frames of our favourite film. Using our understanding of mechanics and motion, we can piece together the action sequences, as well as group together frames depicting similar scenes. In other words, we can reconstruct a coherent order of the frames which will closely resemble the original film. Analogously in pseudotime inference, we have shuffled cells and seek to re-order them temporally. In this case, our hypothesis is that an expert with knowledge of cellular dynamics could piece together the temporal ordering by looking at each cell in the context of the population. In Figure 1, we can see that the cells follow a trajectory in terms of mRNA splicing dynamics. This trajectory is often modelled under the framework of RNA velocity [La Manno et al., 2018], which estimates a cell's pseudotime based on how the transcript counts varies in the population according to a set of biophysical differential equations.

While it is already possible to infer this pseudotime across benchmark datasets from the literature, current methods makes several unrealistic assumptions about the dynamics, such as quantile-derived steady states, shared splicing rates, and discrete transcription rates [Bergen et al., 2021]. In this work, we relax these assumptions in order to construct a generative process which we use to create our training dataset. Motivated by our hypothesis that temporal order is recoverable by a sufficiently

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

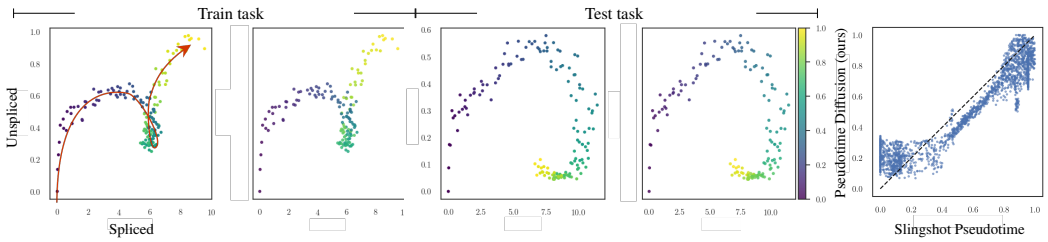

Figure 1: Training and test tasks displayed as paired phase plots, where the left and right plots are the ground truth and mean predictions respectively. The left plot shows a burst of transcription is followed by splicing and mRNA decay, a metastable state, and finally another burst of transcription. We have overlaid a red arrow to illustrate this trajectory.

trained expert, we then introduce a diffusion model capable of unshuffling cell pseudotime, and demonstrate successful results on benchmark datasets.

2 Preliminaries

Pseudotime & RNA Velocity Pseudotime is a latent variable corresponding to the temporal state of a cell which induces an ordering amongst a population of cells. RNA velocity is one method of determining a pseudotime by fitting the RNA splicing dynamics to a set of ordinary differential equations (ODEs). In this paper, we specifically define RNA velocity as the set of equations defined below for unspliced and spliced abundances, $u(t)$ and $s(t)$ respectively. For a single gene,

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} = \alpha - \beta u(t), \quad \frac{\mathrm{d}s_j(t)}{\mathrm{d}t} = \beta u(t) - \gamma s(t), \tag{1}$$

where α is the transcription rate, β is the splicing rate, and γ is the decay rate. Often, the splicing rate is occasionally assumed to be the same across all genes or split in two for both equations to account for amplification bias [Bergen et al., 2021].

Diffusion Models Diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020] are a class of generative models that transform samples from a simple noise distribution to a complex data distribution with a diffusion process. The forward process progressively adds noise to the observations, transforming the data until it matches a Gaussian prior distribution. The reverse process iteratively denoises to generate new data samples. Mathematically, this corresponds to

$$\overbrace{q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})}^{\text{forward}} \qquad \overbrace{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}(\mathbf{x}_t, t), \boldsymbol{\Sigma}(\mathbf{x}_t, t))}^{\text{reverse}} \tag{2}$$

where $\beta_1, ..., \beta_T$ is the variance schedule and $\sqrt{1-\beta_t}$ ensures that the variance remains constant at all timesteps. To avoid confusion, the diffusion index, t, is distinct from pseudotime. For a complete definition of these terms, we revert to the formulation in Ho et al. [2020]. The mean and covariance of the forward process posterior distribution, $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$, is typically modelled by a neural network.

Transformers The Transformer [Vaswani et al., 2017] consists of a stack of layers consisting of multi-head self-attention and a position-wise MLP, with residual connections and layer normalisation. For brevity, we omit the implementation details, as they have been covered extensively online. Self-attention is a mechanism which enables each element in the sequence to focus on different parts of the same sequence, thereby capturing dependencies regardless of their position. It is therefore capable of capturing relationships between elements of a set–important in our case where data is disordered.

3 Diffusion Models for Pseudotime

Data & Physical Bias Our approach does not encode biophysical equations directly in the model, but rather learns a more flexible representation of dynamics from simulated data. The first step is

therefore to generate a set of instances based on our RNA velocity equations. In reality, this dataset will have a different number of genes and cells to the experimentally-derived data at prediction-time. Our model will therefore have to deal with sets rather than matrices.

We first remove some unrealistic simplifying assumptions in the standard RNA velocity formulation [Bergen et al., 2021], in particular the discrete transcription rate and shared splicing rate for all genes. Instead, our simulations use continuous transcription rates as a function of time, and independent splicing rates across genes *apriori*, resulting in the following set of equations for gene j:

$$\frac{\mathrm{d}u_j(t)}{\mathrm{d}t} = \text{softplus}(f_j(t)) - \beta_j u_j(t), \quad \frac{\mathrm{d}s_j(t)}{\mathrm{d}t} = \beta_j u_j(t) - \gamma_j s_j(t), \tag{3}$$

where $f_j(t)$ is the continuous transcription rate as a function of time and to enforce positivity, $\text{softplus}(\mathbf{a}) = \log(1 + \exp(\mathbf{a}))$, since a transcription rate can never be negative. We assign a Gaussian Process prior with RBF kernel to $f_j(t)$, and sample the splicing and decay parameters from uniform distributions with ranges determined by looking at quantiles of the parameters determined in the pancreas, gastrulation, and dentate gyrus datasets [Bastidas-Ponce et al., 2019, Pijuan-Sala et al., 2019, Hochgerner et al., 2018]. Our simulated training dataset is then created using the Alfi framework [Moss et al., 2021] to compute the simulations.

Since our simulations use the same temporal space, cells in the same index share the same timepoint. A simple model would then be able to "cheat" by learning a fixed pseudotime position-wise along the cell axis. We therefore shuffle each simulation independently, encouraging the model to learn the reordering process itself. This led to significant generalisation improvements.

Diffusion Model Our aim is to recover the original temporal ordering of the generative process, which we achieve with a conditional diffusion model. With G genes and C cells, our training dataset consists of $\{\mathbf{x}_0, \mathbf{x}_c\}$, where $\mathbf{x}_0 \in \mathbb{R}^{G\times C}$ are our pseudotimes and $\mathbf{x}_c \in \mathbb{R}^{G\times C\times 2}$ are the unspliced and spliced transcript counts acting as our conditioning data. At prediction time, we have only conditioning information, $\mathbf{x}_c^*$, which is used to generate pseudotime samples. Our loss function is $L = D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$, where we learn the mean and covariance matrix of $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ with a Transformer as described below.

During training, the forward process consists of adding noise at randomly sampled diffusion timepoints using the reparameterisation trick. We use a fixed variance schedule as previous results show improved performance [Ho et al., 2020]. In the reverse process at time t, we concatenate a sinusoidal embedding of diffusion index, $\text{emb}(t)$, the conditioning data $\mathbf{x}_c$, and the noisy observation. This is followed by several layers of scaled dot-product self-attention. Multiple attention heads enable the model to consider multiple weightings between cells. While we did not encounter issues with scalability, for larger datasets with many thousands of cells, we recommend using a sparse attention mechanism if memory becomes a bottleneck. However, this has recently been brought down to linear in sequence length (number of cells) [Dao et al., 2022].

The use of a Transformer is motivated by order-invariance: the order of cells should not impact the model since our observations come as *sets* of cells. The Transformer, via its self-attention mechanism, satisfies this property. Moreover, we omit the positional embeddings typically used in NLP tasks, since we do not want positional indexing to be considered at all.

4 Results

We evaluate our approach by investigating the quantification of uncertainty, recapitulation of differentiation trajectories, and the correlation of our method with existing work. We use the pancreas endocrinogenesis dataset for benchmarking. In particular, we use the benchmark pancreas endocrinogenesis dataset from Bastidas-Ponce et al. [2019], which studies the development of pancreatic endocrine cells. We carry out `velocyto` [La Manno et al., 2018] preprocessing to extract two features: unspliced and spliced transcripts for each gene in every cell.

In Figure 2, we illustrate uncertainty in splicing dynamics by the size of the points. In boxes to either side, we show several samples of smaller regions. Observe in the intersection (Box 2) that there are two trajectories that could have taken place to yield such a splicing pattern. We also investigate whether the trajectories implied by our results match with those in the literature. As shown in Figure 3, the pseudotime suggests a progression from Ngn3, through pre-endocrine, terminating in beta and

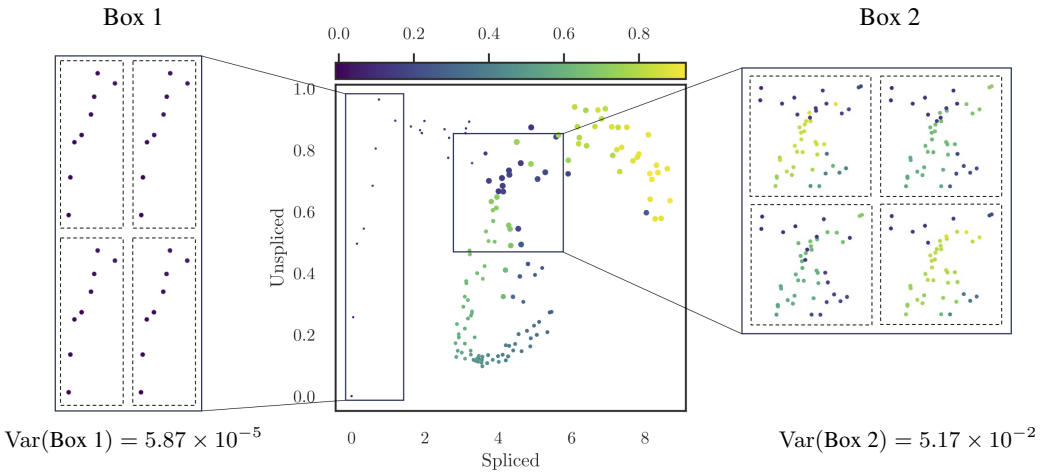

Figure 2: Quantification of uncertainty in the inferred pseudotimes. We generated 100 pseudotime samples from the diffusion model and plotted the mean in the scatter plot. The size of points indicates the sample variance. Boxes 1 and 2 show some of the low- and high-uncertainty samples respectively.

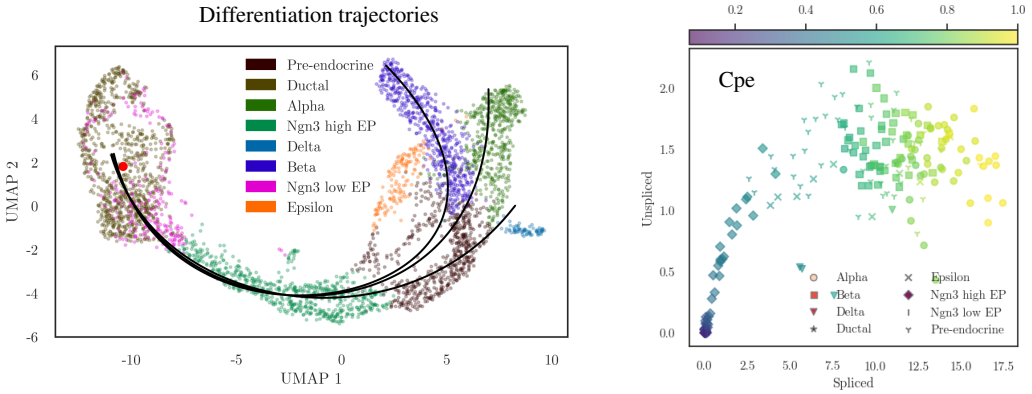

Figure 3: Demonstration of pseudotime correlation with results from the literature. Left: differentiation trajectories inferred by Slingshot on the UMAP embedding. Right: splicing dynamics phase plot showing pseudotime inferred by our model and cell types from the dataset metadata.

alpha cells. Some terminal cell types, for example beta cells, appear to be reached at an earlier pseudotime than alpha cells.

5 Conclusion

The contributions of this paper are two-fold: 1) we have introduced a generative process of creating RNA velocity instances; 2) we have constructed a diffusion model to recapitulate the temporal ordering of cells. Moreover, we have demonstrated the capabilities of our model to generalise beyond the training distribution to real-world experimental data. Our model can also quantify the uncertainty in pseudotime estimation.

Limitations The approach is only *weakly mechanistic*, with the splicing dynamics not being strictly imposed in the model. This may, however, lead to better performance on real datasets where the equations are too rigid. Correlations between genes are not taken into account, which could greatly improve inferences. Extending the framework to multiple lineages is an opportunity for further work.

Related Work Our model is not to be confused with diffusion pseudotime [Haghverdi et al., 2016], which estimates pseudotime via random walks through an embedding space. Slingshot [Street et al., 2018] identifies lineages using a minimum spanning tree, and fits principal curves for each through a low-dimensional embedding of the transcriptome. However, these do not use biophysical biases nor does it quantify uncertainty.

References

Aimée Bastidas-Ponce, Sophie Tritschler, Leander Dony, Katharina Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, Ingo Burtscher, Anika Böttcher, Fabian J Theis, et al. Comprehensive single cell mrna profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146(12):dev173849, 2019.

Volker Bergen, Ruslan A Soldatov, Peter V Kharchenko, and Fabian J Theis. Rna velocity—current challenges and future perspectives. *Molecular systems biology*, 17(8):e10282, 2021.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845–848, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hannah Hochgerner, Amit Zeisel, Peter Lönnerberg, and Sten Linnarsson. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell rna sequencing. *Nature neuroscience*, 21(2):290–299, 2018.

Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12(3):e694, 2022.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018.

Jacob D Moss, Felix L Opolka, Bianca Dumitrascu, and Pietro Lió. Approximate latent force model inference. *Science-Guided AI Symposium at AAAI 2021*, 2021.

Blanca Pijuan-Sala, Jonathan A Griffiths, Carolina Guibentif, Tom W Hiscock, Wajid Jawaid, Fernando J Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard CV Tyser, Debbie Lee Lian Ho, et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566 (7745):490–495, 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19:1–16, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.