Stem Cell Reports, Volume *18*

Supplemental Information

Single-cell multi-omics and lineage tracing to dissect cell fate decision-making

Laleh Haghverdi and Leif S. Ludwig

Single-cell multi-omics and lineage tracing to dissect cell fate decision making
Supplemental Notes

Laleh Haghverdi[1], Leif Ludwig[1,2]

[1]Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany
[2]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany

Note S1: The diffusion-drift model of cell differentiation and its relation with Optimal Transport

Consider the probability density $p(s,t)$ of cells occupying state coordinate s at a time point t. The change in this density over time can be modelled by the diffusion-drift (also known as Fokker-Planck) equation including three terms corresponding to stochasticity (diffusion), the potential energy landscape $U(s)$, and birth/death components of these dynamics. When denoting the diffusion coefficient as $D(s)$ (assuming that D and U are time independent), and population birth with birth/death rate as $S(s,t)$, the Fokker-Planck equation reads:

$$\frac{\partial}{\partial t}p(s,t) = \nabla \cdot \Big(\nabla D(s)p(s,t) + p(s,t)\nabla U(s) + \nabla S(s,t)p(s,t) \Big) \tag{1}$$

Note that in case of $S(s,t) = 0$, we would have conservation of mass such that $\int p(s,t)ds = 1$ for any t. Otherwise, this integral can be less or greater than one, depending on the sum of birth/death events over the space.

In discreet space, the probability density distribution at time t can further be denoted as a vector $P_{(t)}$ of length N, where N is the number of the considered discreet cell states. The discreet version of equation 1 reads:

$$\Delta P_{(t)} = -P_{(t)}\Lambda(L^\alpha + W) \tag{2}$$
$$P_{(t)} = P_{(t-1)}(I - \Lambda(L^\alpha + W)) \tag{3}$$

where Λ is an $N * N$ diagonal matrix with the birth/death rates at each cell state, L the $N * N$ Laplacian matrix (see for example (Haghverdi, 2016)), W the $N * N$ drift matrix (similar to the energy gradients $\nabla U(s)$ in Equation 1) and I presents the identity matrix. $\alpha \geq 1$ specifies the relative strength of diffusion and drift terms, similarly to the role of diffusion coefficient D in the continuous space formulation (for simplicity let us assume constant coefficients over the discreet data points as well as over time). We can define $\Pi = (I - \Lambda(L^\alpha + W))$ as the differentiation propagation operator which maps $P(t-1)$ to $P(t)$. After t time steps, we get:

$$P_{(t_1+t)} = P_{(t_1)}\Pi^t \tag{4}$$

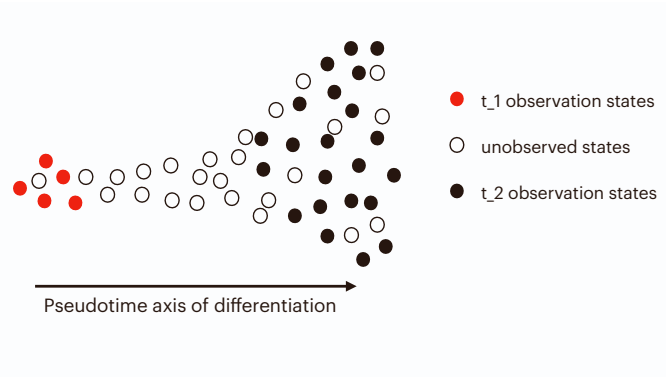

Figure S 1

Consider the N discreet cell states in the phase (e.g., transcription) space as shown in Figure S 1, which includes the observed cell states at two different time points. A realisation from the probability density at time t_1 is observed as $P_{(t_1)}$ with N_1 cells, and a sample $Q_{(t_2)}$ with N_2 cells at a later time point $t_2 = t_1 + t$. A number of N_h intermediate cell states are unobserved (hidden states), such that $N_1 + N_2 + N_h = N$. For the rest of this note, we will drop the time specifications of $P_{(t_1)}$ and $Q_{(t_2)}$ simply referring to them as P and Q. For a given propagation matrix Π the likelihood for such an observation set is given by:

$$L = P\Pi^t Q \tag{5}$$

$$= \sum_{i,k \in 1:N} P_{1i} \quad (\Pi^t)_{ik} \quad Q_{k1} \tag{6}$$

$P = \frac{1}{N_1}(1, 1, ..., 0, 0, ..., 0, 0, ...)$ and $Q = \frac{1}{N_2}(0, 0, ..., 0, 0, ..., 1, 1, ...)$ are both vectors of length N, with nonzero values $(= 1)$ only at the observed cell state positions at t_1 and t_2 respectively. Therefore, the only non-zero terms Equation 6 come from:

$$L = \sum_{i \in 1:N_1, k \in N-N_2:N} P_{1i} \quad (\Pi^t)_{ik} \quad Q_{k1} \tag{7}$$

$$= \sum_{i \in 1:N_1, k \in N-N_2:N} P_{1i} \quad \left[(I - \Lambda(L^\alpha + W))^t\right]_{ik} \quad Q_{k1} \tag{8}$$

$$= \sum_{i \in 1:N_1, j \in 1:N_2} \mathbb{P}_i \quad \hat{\pi}_{ij} \quad \mathbb{Q}_j \tag{9}$$

, where we have redefined the $1 : N_1$ compartment of P as a new vector $\mathbb{P}$ and the $(N - N_2) : N$ compartment of Q as $\mathbb{Q}$. the respective compartment of matrix Π^t is also denoted by a new $N_1 * N_2$ matrix $\hat{\pi}$. This implies that, the maximum-likelihood(ML) solution for the compartment of matrix $(\Pi^t)_{i,k}$ with $i \in \{1 : N_1\}$ and $k \in \{N - N_2 : N\}$ should be the same as the $\hat{\pi}$ matrix we seek to optimise in the Optimal Transport formalism (see Note S3).

Here we only describe the general form by which an ML solution for the diffusion-drift model would translate to the Optimal Transport optimisation scheme. How exactly maximisation of log-likelihood of the above function corresponds to each term in OT (see Note S3) has been researched recently (Léonard, 2013; Fournier and Perthame, 2019), but the precise details and conditions of it (e.g., weak topology requirement for the drift operator such that energy consumption of the transportation can be assumed proportional to the Euclidean distance between the data points) are out of the scope of this note. Interestingly, whereas assuming a model with an unknown number of unobserved intermediate states may seem overwhelming, there is a mathematical workaround for it known as "path integral";

one can assume a very large number of intermediate unobserved states ($N_h \to \infty$), but then account for paths of different length, i.e, summing transition probabilities over all possible paths of length t, but also summing the probabilities over different path lengths ($t = 1 : \infty$). Such an integrated probability of transition from a t_1 cell state to a t_2 cell state, turns out to be convergent and tractable (e.g., see (Haghverdi et al., 2016)).

(Schiebinger et al., 2019) also point out the connection between diffusion-drift and optimal transport frameworks and earlier works related to it (Cuturi, 2013; Léonard, 2013).

Note S2: Diffusion-drift's relation with cell state velocities

We can rewrite equation 1 as:

$$\frac{\partial}{\partial t}p(s,t) = \nabla \cdot \vec{J}(s,t) \tag{10}$$

$$\begin{aligned} J(s,t) &= \nabla D(s)p(s,t) + p(s,t)\nabla U(s) + \nabla S(s,t)p(s,t) \\ &= \vec{V}(s)p(s,t) + \nabla S(s,t)p(s,t) \end{aligned} \tag{11}$$

$\vec{J}(s,t)$ can be interpreted as the flux of cells. That is, the time-derivative of the density $p(s,t)$ is given by the divergence of the flux; how much the number of cells changes in a volume around s in time δt is equal to the number of cells that enter the volume minus the number of cells that exit it in δt (note that probability density is the number of cells per volume, $p(s,t) = \frac{\delta n(s,t)}{\delta \mathrm{VOL}}$). In absence of birth/death events, the mass flow in/out to the volume is given by $\delta p = \int_A \vec{J}\delta t = \int_A \frac{\delta n(s,t)}{\vec{A}\delta t}\delta t = \int_A \frac{\delta n(s,t)\vec{V}(s)}{\vec{A}\cdot\vec{V}(s)\delta t}\delta t = \int_A p(s,t)\vec{V}(s)\delta t$, where $\vec{A}$ denotes the normal vectors of the surface of the volume and $\vec{V}(s)$ the velocity vector field at position s. Therefore, by excluding birth/death events we have used the $\vec{J'} = \vec{V}(s)p(s,t)$ relation in equation 11, from which we conclude that cell state velocities are given by the sum of the diffusion (noise) and drift (directed force) terms of the Fokker-Planck equation:

$$\vec{V}(s)p(s,t) = \nabla D(s)p(s,t) + p(s,t)\nabla U(s) \tag{12}$$

Equation 12 is also known as the "Langevin equation" in the statistical physics literature for Brownian motion.

Note S3: The Optimal Transport model of cell differentiation

Here, we include the entropic regularised and unbalanced formulation of OT according to (Schiebinger et al., 2019). To compute the Optimal Transport map between the data points P at time t_1 and Q at time t_2, OT sets the following optimisation problem:

$$\begin{aligned} \hat{\pi}_{ij} = \arg\min_{\pi} \Bigg(& \sum_{i\in 1:N1, j\in 1:N2} c(s_i,s_j)\pi_{ij} - \epsilon \sum_{i\in 1:N1, j\in 1:N2} \pi_{ij}\log\pi_{ij} \\ & + \beta_1 \mathrm{KL}\Big(\sum_{i\in 1:N_1}\pi_{ij}||\mathbb{Q}_j\Big) + \beta_2 \mathrm{KL}\Big(\sum_{j\in 1:N_2}\pi_{ij}||\mathbb{P}_i\Big) \Bigg) \end{aligned} \tag{13}$$

$$\begin{aligned} = \arg\min_{\pi} \Bigg(& \sum_{i\in 1:N1, j\in 1:N2} c(s_i,s_j)\pi_{ij} - \epsilon \sum_{i\in 1:N1, j\in 1:N2} \pi_{ij}\log\pi_{ij} \\ & + \beta_1 \mathrm{KL}\Big(\mu_j||\mathbb{Q}_j\Big) + \beta_2 \mathrm{KL}\Big(\lambda_i||\mathbb{P}_i\Big) \Bigg) \end{aligned} \tag{14}$$

s_i and s_j determine the position of cells $i \in \{1 : N_1\}$ and $j \in \{1 : N_2\}$ from observation time points t_1 and t_2, and $\mathbb{P}$ and $\mathbb{Q}$ present the N_1 and N_2 dimensional normalised state vectors at the corresponding time points, similarly to the notation used in Note S1. $c(s_i, s_j)$ is the Euclidean distance between cell i and j in the phase space and constitutes the energy consuming term of the transportation, similar to drift in Note S1. ϵ determines the level of randomness (i.e., entropy) in the mapping between the two observations, similar to diffusion. When using the OT model, the parameters $\epsilon, \beta_1, \beta_2$ need to be specified by the user. In the last line, $\mu_j = \sum_{i=1}^{N_1} \pi_{ij}$ and $\lambda_i = \sum_{j=1}^{N_2} \pi_{ij}$ indicate the "inferred" birth/death rate for the corresponding cell states.

Too see display a form of the above regularized optimization problem of OT that more closely relates to a log-likelihood maximisation scheme of the diffusion-drift operator (see the likelihood function in equations 7-9), we expand the Kullback–Leibler divergence (KL) term as $\mathrm{KL}(\mu_j||\mathbb{Q}_j) = \sum_{j=1}^{N_2} \mu_j(\log(\mu_j) - \log(\mathbb{Q}_j))$ and use the relation $\log(\mathbb{Q}_{\mathbb{j}}) = \log(\frac{1}{N_2})$ for all $j \in 1 : N_2$ (similarly for $\mathrm{KL}(\lambda_i||\mathbb{P}_i)$):

$$\hat{\pi}_{ij} = \arg\min_{\pi} \Bigg(\sum_{i \in 1:N1, j \in 1:N2} c(s_i, s_j)\pi_{ij} - \epsilon \sum_{i \in 1:N1, j \in 1:N2} \pi_{ij} \log \pi_{ij} + \beta_1 \sum_{j \in 1:N_2} \mu_j(\log(\mu_j) + \log(N_2)) + \beta_2 \sum_{i \in 1:N_1} \lambda_i(\log(\lambda_i) + \log(N_1)) \Bigg) \tag{15}$$

Using the OT formalism as such, one tries to identify the $\hat{\pi}$ which best describes the observed data $\mathbb{P}$ and $\mathbb{Q}$ generally without knowing the true values for the underlying (hidden) parameters of the dynamics including the true birth/death rate at the position of each cell, the actual time steps t by which the two observations are apart and the relative magnitude of randomness to the directed (deterministic) component of cell differentiation.

References

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Fournier, N. and Perthame, B. (2019). Monge-kantorovich distance for pdes: the coupling method. *arXiv preprint arXiv:1903.11349*.

Haghverdi, L. (2016). *Geometric diffusions for reconstruction of cell differentiation dynamics*. PhD thesis, Doctoral Thesis, Technische Universität München.

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845–848.

Léonard, C. (2013). A survey of the schroedinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.