The definition of pseudotime and its correlation with gene expression fluctuations in time-dependend biological processes

Guilherme de Sena Brandine

October 5, 2016

Many biological processes in complex systems are governed by successive transitions between cellular states that occur on the span of several cell division cycles. These transitions start from one or many progenitor cells with similar characteristic phenotypes (such as embryonic or hematopoietic stem cells) and are regulated by a tightly controlled set of steps that continuously alter the expression levels of many genes as cells divide, giving rise to intermediate phenotypes until cells fully differentiate into their terminal states with specialized roles. In an ideal scenario, in order to understand which genes have a role in these processes, we could observe the expression profile of all genes in a single progenitor phenotype as it differentiates and infer from these observations which genes are up and downregulated as time progresses. In reality, if we observed this temporal development starting from a different progenitor cell each time, we would see a high variability in the set of expression profiles that yield a particular intermediate phenotype, the rate at which cells divide and the number of divisions necessary to terminate the differentiation. Conversely, the expression profile of a cell uniquely characterizes a phenotype, so we can correlate the gene expression values to a relative measure of how far down the differention pathway a cell is. We call such measure the cell's *pseudotime*.

The notion of pseudotime has been previously introduced in literature. Trapnell et al (1) first defined the "pseudotemporal reconstruction" problem, which takes a set of expression profiles of cells captured in different time points and attempts to sort them from earlier to later down the differentiation path. With this ordering, it attempts to find the relation of gene expressions with pseudotime assuming there exists a smooth function $f : \mathbb{R}^D \to [0, 1]$ that associates them. This is problematic since only a small subset of $\mathbb{R}^D$ actually have expression values that make sense to associate to a pseudotime value in a particular process. The definition should thus be reversed, and instead map a particular value of pseudotime to a set of expression values. Many methods were developed afterwards under the same mindset (2)(3) (4) (5) (6) and either work on the same definition given by Monocle or treat pseudotime as a latent one dimension variable that describes the variability in of the expression profiles observed. These methods usually need to impose restrictions on the behavior of the observed variability, such as Gaussian noise (6) or that the pseudotime difference between two cells is directly proportional a (dis)similarity measure of the expression of both cells(2)(3) (4), which does not necessarily describe the full spectrum of the biological variance observed in biological processes. We attempt here to put the concept of pseudotime in a biological context and define how gene expression is influenced by it, as well as the assumptions we make

concerning functions that describe expression based on our pseudotime definition.

We begin by assuming that a population of cells is organized according to developmental relationships that can be described by ordering the cells in such way that the earlier phenotype associated to cell i differentiates into a later phenotype associated to cell j through a subset of cells with intermediate phenotypes between i and j. For instance, if the population of interest has a defined stem cell phenotype, and the stem cells give rise to the entire population, then that stem cell phenotype would be the progenitor and correspond to the earliest time point. In experiments of interest, all cells would be sampled simultaneously, making it difficult to claim that some cells correspond to earlier or later stages of differentiation. We thus define a biological measure of differentiation development, called *pseudotime* as a value that indicates the relative developmental distance of a phenotype to the aforementioned root. More specifically, the pseudotime of a phenotype is the ratio between the time necessary for the progenitor from which it arose to differentiate into that phenotype and the time required for the same progenitor to fully differentiate into a terminal phenotype. As such, pseudotime is monotonically related to our naturally defined notion of time, but is potentially subject to different scales in different differentiation pathways.

The differentiation of progenitors into terminal phenotypes is a highly coordinated process that continuously alters the expression of different genes throughout several cell divisions until the daughter cells reach a final specialized phenotype. The relative expression value of the genes in a cell can hence be described by a function of a continuous argument value $\mu(t) : [0, 1] \to [0, 1]^D$ that describes the fluctuation of the gene expression with the pseudotime. A safe assumption to make about μ is that each one of its coordinates is continuous except in a subset of $[0, 1]$ of measure zero, which, in the biological context, means that discontinuities in expression values happen - if ever - in discrete time points, and forbids nonsensical behaviors like Cantor functions. Biologically, $\mu(t)$ is the expected expression profile of a cell in pseudotime t.

To account for expression variations yielding the same phenotype, we also introduce variation function. $\delta : [0, 1] \to V$, where V is the vector subspace of random vectors in $[0, 1]^D$ that describes the biological variance that yields intracellular gene fluctuation. By our previous definition of μ, we require $\mathbb{E}[\delta(t)] = 0$. In this context, each observation of an expression profile of a cell in a pseudotime t can be modeled by an observation of the random variable $\mu(t) + \delta(t)$.

References

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

Sean C Bendall, Kara L Davis, El-ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Peer. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725, 2014.

Zhicheng Ji and Hongkai Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, page gkw430, 2016.

Laleh Haghverdi, Maren Buettner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *bioRxiv*, page 041384, 2016.

Tarmo Äijö, Vincent Butty, Zhi Chen, Verna Salo, Subhash Tripathi, Christopher B Burge, Riitta Lahesmaa, and Harri Lähdesmäki. Methods for time series analysis of rna-seq data with application to human th17 cell differentiation. *Bioinformatics*, 30(12):i113–i120, 2014.

Kieran Campbell and Christopher Yau. Bayesian gaussian process latent variable models for pseudotime inference in single-cell rna-seq data. *bioRxiv*, page 026872, 2015.