The definition of pseudotime and its correlation with gene expression fluctuations in cell differentiation processes

Guilherme de Sena Brandine

October 3, 2016

Single cell RNA-Seq is a technique that allows individual profiling of the transcriptomes of tissues at an individual cell resolution, and is a potential tool to improve our understanding of many biological processes and the key genes that regulate them. One particular application of interest is the study of differentiation by sorting the sampled cells into their most likely developmental order, whence we can infer the temporal behavior of all sequenced genes and thus decide which ones are influenced by differentiation. Several methods have been developed to infer such ordering of cells. Monocle (4) was the first method to define the "pseudotemporal reconstruction" problem, which introduced the concept of pseudotime as a measure of biological development which is analogous to the elapsed time in the differentiation process, and attempts to find, for each gene, a smooth function that characterizes how its expression is influenced by this measure. After Monocle, many other methods, such as Wanderlust (1), TSCAN (3) and DPT (2) have been later developed as an attempt to improve robustness and speed performances for this problem, but they all use a similar standard procedure to define pseudotime: They first build a complete graph using a defined distance measure (Euclidean Distance for Monocle and TSCAN, Cosine distance for Wanderlust, Hellinger distance for DPT), after which they find the most likely reconstruction of their dataset timeline, and finally define the pseudotime as the relative distance of each cell to a user-defined root cell, which is later used to fit an expression $\times$ time function for each gene, either through generalized additive models or sliding windows. Pseudotime thus becomes a mathematical artifact used for curve fitting rather than a value with some biological meaning. In other words, it becomes the consequence of a combinatorial optimization algorithm rather than part of the problem definition. 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, including progenitors - which are intermediates of an ongoing differentiation process - and terminally differentiated cells - which have either exited cell cycle or yield phenotypically identical daughter cells upon division -. These relationships can be described by ordering the cells in such way that an earlier cell i differentiates into a later cell j through a subset of intermediate cells between i and j. In this context, every cell in the population is derived from a *root cell* and continuously divides into *terminal cells*. 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 root cell. 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 cell to the aforementioned root cell - assuming the latter is known -. More specifically, the pseudotime of a cell captured at a specific point in its life cycle is the ratio between the time necessary for a root cell to differentiate into that cell and reach the point in its life cycle at which it was measured and the time required for the root cell to fully differentiate into a terminal cell. As such, pseudotime is monotonically related to our naturally defined notion of time.

The differentiation of root cells into terminal cells is a highly coordinated process that continuously alters the expression of different genes throughout several cell divisions until the daughter cells reach an immutable specialized phenotype. The relative expression value of every gene g in a cell can hence be described by a function of a continuous argument value $t \to \mu_g(t) : [0,1] \to [0,1]$ that describes the fluctuation of the gene expression with the pseudotime. A safe assumption to make about μ_g is that it 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.

Expression values also vary in the life cycle of a cell due to diverse biological processes, such as transcription regulation, synchronized external signaling and mRNA degradation. Since current cell sampling technologies are capable only of measuring a static snapshot of a cell's transcriptome, each gene also has an inherent dispersion function $\sigma_g^2(t)$ that describes the biological variance that yields intracellular gene fluctuation. We also assume $\sigma_g^2(t)$ is piecewise continuous, but the discontinuity set does not need to be the same as that of μ_g.

References

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.

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

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

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.