Gene Network Inference from pseudotime ordered scRNAseq data

Faure L(1,2,3,4)[†] , Barillot E(1,2,3), Zinovyev A(1,2,3), Albergante L(1,2,3)

(1) Institut Curie, PSL Research University, F-75005 Paris, France
(2) INSERM, U900, F-75005 Paris, France
(3)MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France
(4) Institute of Technology and Innovation (ITI), PSL Research University, PSL Research University, F-75005 Paris, France

[†] Email: louis.faure@curie.fr

Motivation
Thanks to recent technological advances, single-cell gene expression measurements (scRNAseq) are now quite common for biological research. While many challenges remain on the the analysis of such data, scRNAseq datasets provide crucial information about the studied biological sample. For example, it is possible to identify cell types, their molecular state, and also marker genes associated with a dynamic process. If the sample contains cells displaying different level of commitment to a biological process, e.g. bone marrow progenitors differentiation into dendritic cells, it is even possible to perform pseudotime analysis to understand the molecular changes associated with the biological process under consideration and thus to explore different ways to control or even reverse such biological process. For this reason, we investigated how scRNAseq data and pseudotime reconstruction can be used to recover regulatory interactions, which can, in principle, be used to reconstruct the Gene Regulatory Network (GRN) of a cell.
Methods
To reconstruct pseudotime from scRNAseq data we used a versatile method that is developed in our group called ElPiGraph. ElPiGraph is a machine learning approach that is able to learn complex structure form the data and can be used, among other things, to reconstruct linear pseudotime (associated with differentiation), circular pseudotime (associated with periodic processes) and branching pseudotime (associated with differentiation into different populations). To better contextualise our results, we used a publicly available dataset focused on the differentiation of mouse conventional dendritic cells.
From the dataset, we selected genes identified as targets of a transcription factors (TF) from RegNetwork, an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse, containing experimental data. For the selected TF, we identified all its targets and also selected the same number of random genes to perform statistical validation.
GRN inference was performed by applying different methods based on different strategies, and hence requiring distinct inputs. GENIE3 and ARACNE do not use

pseudotime information, but only the expression matrix to infer a GRN. dynGENIE3 and SWING are both designed for bulk RNAseq data with several time points, but it is possible to process our scRNAseq data by fitting the gene expressions with pseudotime using lowess, leading to time-series with equally spaced time points. Finally, SCODE is using the pseudotime ordering as input. From the results of the different algorithms, causality scores (weight, mean importance,…) on the interaction between the TF and the randomly selected genes and the targets from the database were extracted and compared in order to assess if the two groups are distinguishable by the algorithm.

Results

Five TFs were selected in such a way to perform our analysis across TFs with both a large and small number of regulated genes. In most cases, no statistical differences can be detected between RegNetwork targets and the random ones. This can be explained on one hand by the small sample sizes (the number of regulated genes is generally small), and on the other hand by the fact that the concerned TFs are possibly having no significant impact to their target at that differentiation stage. In the case of the gene Fos, both SWING and GENIE3 managed to differentiate the two groups, with the group of targets from the RegNetwork showing a significantly higher causal score than the random group. Further exploration will be focused on exploring how the different algorithms behave on simulated scRNA-seq with known GRN. These synthetic datasets will allow us to explore to which extent data processing affects the algorithm used . In particular, pre-processing of scRNAseq data such as normalization, dropouts imputation, dimensionality reduction, and pseudotime algorithm ordering will be assessed to measured the impact on the resulting GRN.