GRNBoost2 Documentation

GRNBoost2 is a fast GRN inference algorithm using stochastic Gradient Boosting Machine regression with early-stopping regularization, the Arboreto flagship algorithm. Like GENIE3, GRNBoost2 trains a regression model to select the most important regulators for each gene in the dataset. GRNBoost2 achieves its efficiency by using stochastic gradient boosting machine regression with early stopping regularization to infer the network.

This class of GRN inference algorithms is defined by a series of steps, one for each target gene in the dataset, where the most important candidates from a set of regulators are determined from a regression model to predict a target gene's expression profile.

GRNBoost2 adopts the GRN inference strategy exemplified by GENIE3, where for each gene in the dataset, the most important features are selected from a trained regression model and emitted as candidate regulators for the target gene. All putative regulatory links are compiled into one dataset, representing the inferred regulatory network.

Input:

Single-cell RNA-seq expression matrix (the GSE81252.lineagesubsetted.txt file), where each column corresponds to a feature (gene) and each row corresponds to an observation; a list of gene names corresponding to the columns of the expression matrix, and a list of the transcription factors (the GSE81252.lineagesubsetted.txt file).

Output:

The inferred gene regulatory links, in the form of a Pandas data frame ["TF", "target", "importance"], saved as the output_updated.tsv file.

About Arboreto

A component in pySCENIC: a lightning-fast python implementation of the SCENIC pipeline (Single-cell Regulatory Network Inference and Clustering) which enables biologists to infer transcription factors, gene regulatory networks and cell types from single-cell RNA-seq data.

This pipeline has 3 steps:

1. First transcription factors (TFs) and their target genes, together defining a regulon (the regulatory network that connects a TF with its target genes), are derived using gene inference methods which solely rely on correlations between expression of genes across cells. The Arboreto package is used for this step.

2. These regulons are refined by pruning targets that do not have an enrichment for a corresponding motif of the TF effectively separating direct from indirect targets based on the presence of cis-regulatory footprints.
3. Finally, the original cells are differentiated and clustered on the activity of these discovered regulons.

Things to Note:

1. The final network execution takes about 20 minutes to execute in python.
2. If you obtain a dataset such as from single-cell genomics, the rows represent genes and the columns represent features. In this case you will have to transpose the dataset in order to obtain the final regulatory network.
3. No efficient R file exists for the GRNBoost2 inference method.
4. Both GENIE3 and GRNBoost2 are based on stochastic machine learning techniques, which use a random number generator internally to perform random sub-sampling of observations and features when building decision trees. To stabilize the output, can optionally add a seed value that is used to initialize the random number generator used by the machine learning algorithms.

References:

https://readthedocs.org/projects/arboreto/downloads/pdf/latest/