GENIE3 – Gene network inference with ensemble of trees

About GENIE3:

GENIE3 is a model-free method that infers gene regulatory networks in the form of a weighted adjacency matrix from steady-state expression data. This method exploits variable importance scores derived from Random forests to identify the regulators of each target gene. GENIE3 uses a Random Forests approach to predict the strength of putative regulatory links between a target gene and the expression pattern of input genes (i.e. transcription factors). GENIE3 provide a ranking of the regulatory links from the most confident to the least confident. To evaluate such a ranking independently of the choice of a specific threshold, use the precision-recall (PR) curve and the area under this curve (AUPR).

GENIE3 decomposes the prediction of a regulatory network between *p* genes into *p* different regression problems. In each of the regression problems, the expression pattern of one of the genes (target gene) is predicted from the expression patterns of all the other genes (input genes), using tree-based ensemble methods Random Forests or Extra-Trees. The importance of an input gene in the prediction of the target gene expression pattern is taken as an indication of a putative regulatory link. Putative regulatory links are then aggregated over all genes to provide a ranking of interactions from which the whole network is reconstructed. Targets is the subset of genes to which potential regulators will be calculated.

The targeted networks are directed graphs with p nodes, where each node represents a gene, and an edge directed from one gene *i* to another gene *j* indicates that gene *i* directly regulates the expression of gene *j*. One interesting feature is GENIE3 can predict directed networks, while methods based on mutual information or correlation are only able to predict undirected networks.

About the dataset:

This dataset was obtained from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81252. The cell types measured in the dataset are human mature hepatocytes (hHEP). The dataset is from an scRNA-seq experiment on induced pluripotent stem cells (iPSCs) in two-dimensional culture differentiating to hepatocyte-like cells. The dataset contains 425 scRNA-seq measurements from multiple time points: days 0 (iPSCs),6,8,14 and 21 (mature hepatocyte-like). Each row corresponds to a sample/experiment and each column corresponds to a gene. The data is also known as a gene expression profile, where gene expression means the amount of a gene expressed in given conditions. After subset is performed on the dataset to increase speed, it contains 100 genes from 50 samples.

Algorithm contains 3 steps:

1) Initially, it creates an ensemble of regression tress for each gene in the network. It predicts the behavior of each gene in the dataset. In this step, it breaks huge data into small subsets. Based on the threshold value of expressed genes it creates group. Random forest is used to split data and construct regression trees.

2) Possible regulators are ranked from each regression trees. It can be done by ranking influence of every other gene when tree is created. A score for a potential regulator gene is calculated by summing all the important scores from the nodes where gene was selected for splitting. Afterwards, based on scores ranking it is determined what genes are important for regulating gene.

3) Ranks the inferred edges overall. In this step, important scores are generated for each tree in each forest, which gives list of potential regulators for each tree.

Findings on Parameters:

1. When using the extra-trees method, if the number of chosen candidate regulators K is increased by 1 and keep nTrees constant, more weights in the weighted adjacency matrix will be greater than that in the original weighted adjacency matrix.
2. When using the extra-trees method, if the number of chosen trees per ensemble nTrees is increased by 10 and keep K constant, more weights in the weight adjacency matrix will be greater than that in the original weighted adjacency matrix.
3. When using the random-forests method, if the number of chosen candidate regulators K is increased by 1 and keep nTrees constant, less weights in the weighted adjacency matrix will be greater than that in the original weighted adjacency matrix.
4. When using the random-forests method, if the number of chosen trees per ensemble nTrees is increased by 10 and keep K constant, less weights in the weight adjacency matrix will be greater than that in the original weighted adjacency matrix.

Things to Note:

1. The R package GENIE3 is not included in the R repository. To download the package, follow the code in the R file.
2. Since the GENIE3() function takes as input argument a gene expression matrix where the rows are the genes and the columns are the samples/conditions, the dataset used in the associated R file had to be transposed to meet this criterion.
3. Have to subset the dataset before performing GENIE3 as the method takes a long time to run on large datasets.

References:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2946910/