STA426 treekoR project

Anouk Petitpierre and Tudor Jumuga

Version January 2, 2022

Contents

1 Introduction **5**
- 1.1 Clustering . . . 5
- 1.2 Hierarchy . . . 5
- 1.3 Proportions . . . 5
- 1.4 Significance Testing . . . 5
- 1.5 Visualisation . . . 5

2 Methods **7**

3 Results **8**

4 Discussion **9**

A Appendix **10**

References **11**

Chapter 1

Introduction

The development of high-throughput single-cell technologies provide bioinformaticians with a large amount of high-dimensional cytometry data.

Usually, these data are analysed using either Manual Gating or Hierarchical Clustering. Manual Gating clusters cells into discrete populations based on shared marker expression, Hierarchical Clustering stratifies cell subsets without a predetermined hypothesis.

Both methods have their advantages, but unfortunately, they have their disadvantages, too: Manual Gating is time-consuming and potentially biased, as the markers chosen for the hierarchical clustering are based on expert opinion. Hierarchical clustering is more time-efficient, but it dismisses cell hierarchy. treekoR is a novel framework developed to solve these disadvantages. It uses an automated hierarchical clustering algorithm to make the analysis more time-efficient and neutral, while still taking into account cell hierarchy. The treekoR framework consists of five steps:

1.1 Clustering

1.2 Hierarchy

1.3 Proportions

1.4 Significance Testing

1.5 Visualisation

- how does treekoR solve these problems? uses automated hierarchical clustering algorithm to make it more time-efficient and neutral, calculates both

- what is the workflow of treekoR? (1) clustering (2) constructing hierarchy (3) computing proportions (parent and total) (4) Significance Testing (5) Visualisation

- what is the goal of this project? reproduce an analysis on a raw high-dimensional cytometry dataset using the framework of treekoR, also specify which dataset and provide a description of the dataset compare how well treekoR is able to determine cellular relationship with disease

using Find an appropriate measure for the comparison that is not "balanced accuracy". Provide explanation of our choice here, as well.

Chapter 2

Methods

We're using R version blabliblu and Rstudio version bliblabla and BiocManager version 3.14 and we're using the packages treekoR, SingleCellExperiment etc.

part I: reproduction of treekoR analysis on Age Chronic (?) data set - reading in, transforming fsc to sce - arcsinh transformation with cofactor 5 of count data - clustering of data using flowSom-based function cluster() from the CATALYST package - construct cell hierarchy, once using "hopach" and once using "average linkage" algorithm, with function given in treekoR package, provide short explanation of how these these two hierarchical clustering algorithms work - compute parent as well as total proportions for both hierarchical trees - significance testing for both proportions - visualisation for both hierarchies

part II: Benchmarking - use provided datasets by the author(s) of treekoR - use their code (provide description of what the code does here) - on top of balanced accuracy, compute our chosen measure and explain how it is computed

Chapter 3

Results

part I: reproduction of treekoR analysis - provide hierarchical trees - provide treekoR heatmap plots in the end - report significant findings / observations

part II: Benchmarking - provide Benchmarking plots, once using balanced accuracy as a measure of comparison, once using our "own" - report findings / observations

Chapter 4

Discussion

part I: - why does our analysis look different from theirs? - what has remained unclear? - what could be explored further?

part II: - why does our analysis look different than theirs? - are there differences when using other measure instead of balanced accuracy? why? what does that mean? - what remains unclear? - what could be explored further?

Appendix A

Appendix

- code - where the data can be found

References