Pairseek: Identifying prognostic pairwise relationships among bacterial species in microbiome studies

Sean Devlin and Irina Ostrovnaya

March 8, 2021

Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center
devlins@mskcc.org

Contents

1 Overview

This document presents an overview of the `PairSeek` algorithm. The goal of the method is to find pairs of bacteria that have differential dichotomized relationship as specified below between cases and controls.

Let X_{ik} be the abundance of i-th bacterium, $i = \{1, \ldots, p\}$ in the k-th subject, $k = \{1, \ldots, n\}$. For amplicon sequencing, X_{ik} would be the non-normalized number of reads matching the specific bacterial sequence. Suppose we want to test if the relationship between two bacterial abundances is dependent on some binary disease state Y_k, e.g. cancer vs control.

We define binary indicator variables $Z_k^{ij}(c) = 1(X_{ik} \leq cX_{jk})$ which take a value of 1 if X_{jk} is at least $c-$fold smaller than X_{ik}. We are looking for pairs that have $X_{ik} \leq cX_{jk}$ equal to 1, for example, in most cases and equal to 0 for most controls. If c is set to 1, the resulting dichotomized variables will only be useful for comparing bacteria with abundances on the same scale.

For intuitive explanation imagine prevalences of two bacteria plotted against each other. Using the proposed algorithm we will be able to detect pairs of bacteria that tend to have prevalences on the opposite sides of regression line specified by slope c. We will utilize resampling as in stability selection framework to both get more stable measure of pair's association with cohort and to get the slopes c. We will split subjects into two groups randomly: on one group for each pair of bacteria we will estimate optimal slopes c that

separate cases and controls, and then we will fit LASSO (least absolute shrinkage and selection operator) on the opposite set of patients with all possible pairs dichotomized at the estimated slopes and random penalty parameter. After repeating these steps say 1000 of times we will calculate how often each pair of bacteria was selected among these LASSO runs. This quantity, dominance score, can be used to rank pair's association with cohort.

2 Data example

We will illustrate how the method works based on oral cancer dataset from (Bornigen et al., 2017).

References

Daniela Bornigen, Boyu Ren, Robert Pickard, Jingfeng Li, Enver Ozer, Erica M. Hartmann, Weihong Xiao, Timothy Tickle, Jennifer Rider, Dirk Gevers, Eric A. Franzosa, Mary Ellen Davey, Maura L. Gillison, and Curtis Huttenhower. Alterations in oral bacterial communities are associated with risk factors for oral and oropharyngeal cancer. *Scientific Reports*, 7, January 2017. doi: 10.1038/s41598-017-17795-z. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5732161/.

```
> library(PairSeek)
> dataClinical <- read.delim(
+   "https://europepmc.org/articles/PMC5732161/bin/41598_2017_17795_MOESM4_ESM.txt",
+   header=T)
> y <-   dataClinical$casectrl
> table(y)

y
  0   1
242 121

> dataOTU<-read.delim(
+   "https://europepmc.org/articles/PMC5732161/bin/41598_2017_17795_MOESM2_ESM.txt",
+   header=T)
> testH <- strsplit(as.character(dataOTU$taxonomy), "__", fixed = FALSE)
> genuslist <- NULL
> for(i in 1:length(testH))  genuslist <- c(genuslist, testH[[i]][7])
> uniquegenus <- unique(genuslist)
> uniquegenus <- uniquegenus[uniquegenus != ";s"]
> OTUmatrix <- NULL
> for(i in 1:length(uniquegenus)) OTUmatrix <-
+   cbind(OTUmatrix,
+         apply(dataOTU[which(genuslist == uniquegenus[i]) ,2:364], 2, sum))
> colnames(OTUmatrix) <- uniquegenus
```

```
> rownames(OTUmatrix) <- colnames(dataOTU)[2:364]
> OTUmatrix <- as.matrix(OTUmatrix[ order(rownames(OTUmatrix)),])
> OTUmatrix <- OTUmatrix[,!grepl("\\[",colnames(OTUmatrix) )]
> OTUmatrix <- OTUmatrix[,apply(OTUmatrix != 0,2, mean) > 0.10]
>
```

The dataset contains 121 samples from oral cancer patients and 242 samples from healthy controls. After aggregating bacterial abundance from 2770 operational taxonomic units (OTUs) to the genus taxonomic level and filtering out genera observed in fewer than 10% of the samples, a total of 61 genera were included in the analysis.

Rows of data correspond to samples, and columns to bacterial species.

```
> dim(OTUmatrix)

[1] 363  61

> table(y)

y
  0   1
242 121

> table(rownames(OTUmatrix)  == dataClinical[,1])

TRUE
 363

> set.seed(100)
> dom.scores<-PairSeekBinary(y,OTUmatrix,LassoIterations=100)

Percent computed: 10.. 20.. 30.. 40.. 50.. 60.. 70.. 80.. 90.. 100..

>
```

Below are the details of the session information:

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Mojave 10.14.6

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
```

```
attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] glmnet_4.0-2 Matrix_1.3-0 Rcpp_1.0.5   PairSeek_1.0

loaded via a namespace (and not attached):
 [1] compiler_4.0.2   tools_4.0.2      survival_3.2-7   splines_4.0.2
 [5] codetools_0.2-18 grid_4.0.2       iterators_1.0.13 foreach_1.5.1
 [9] shape_1.4.5      lattice_0.20-41
```