Institut National des Sciences Appliquées de Toulouse
Mathematical and Modeling Engineering Department

October 2017 - January 2018

Fifth-year project

Identification and caracterisation of virome and bacteriome in calves suffering from infectious bronchopneumonia

Authors :

Gicu Stratan
Soizick Magon de La Giclais

Tutor :

Nathalie Villa

Abstract

Contents

Acknowledgment

Introduction

1 Context and description of the data set

1.1 Context

1.2 Data set description

Two text files are available for the study:

- `abundances`
- `pathogenes`

They are both under *Comma-Separated Values* (CSV) format. In this chapter will be seen what they contain and what type of transformations will be perfomed, in order to study them.

The `abundances` file:

This file contains microbiote data. We call microbiote the set of bateria and micro-organisms in a body. In our case, it is in calf bodies, especially in the nose and in the lungs. In this file, there is information about each bacteria found within samples.

At the begining, there is 54 columns and 406 rows. The first 9 columns contains the name of each bateria, its type, its family, its "blast taxonomy" and technical information that we will not use. We are interested in the name of each bacteria (an identifier) and the measurments made in each sample. Without the 9 first columns, we have 45 columns which correspond to the samples taken from calves.

We can easely see a first problem: the number of sample is odd (45), but on each calf was supposed to be taken two samples: one in the nose and the other in the lungs.

```
> table(id_abundances)
## 01 03 04 06 07 09 10 11 15 16 17 18 19 20 21 23 24 25 26 28 29 30 31
##  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  1  2  2

> table(condition)
## condition
```

```
##  EN LBA
##  22  23
```

The 29th calf is the one which is not paired, there is no sample taken from the nose. It will be removed from the study when a paired analyse will be made.

A second problem concern the list of 406 bacteria. This list is indeed not composed by unique names of bacteria. We could have used the blast identifier but the problem remain the same.

```
> sum(unique(names(which(table(df_abundances[ ,1]) > 1))))
## 273
```

Let us look the 5 first non-unique bateria identifier:

```
##  [1] "Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;
Corynebacteriaceae;Corynebacterium 1;Corynebacterium sp."
##  [2] "Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;
Corynebacteriaceae;Corynebacterium 1;unknown species"
##  [3] "Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;
Corynebacteriaceae;Corynebacterium;unknown species"
##  [4] "Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;
Dietziaceae;Dietzia;Dietzia sp."
##  [5] "Bacteria;Actinobacteria;Actinobacteria;Corynebacteriales;
Nocardiaceae;Rhodococcus;Rhodococcus sp."
```

First, let us keep only the specie name. It is the last one on each row. But sometimes the specie is unknow, as we can see above. In this case, the name that appear before will be kept, *e.g.* for the second row we kept `"Corynebacterium 1"` as identifier for the bacteria. Then, the replicates are merged by adding together each one of them, which lead us to a list of 270 unique bacteria.

Another problem concern only one of these bacteria: the name is not normal: the complete name is `"&"`, and we do not know what it is.

The `pathogenes` file:

The second file is composed of 9 columns and 46 rows. The first and the last columns stands for the identifier of each calf and its condition, which will be merged. Each of the seven variables is a virus. We have for each observation the presence or not of the virus. Unlike the previous file, there is no missing calf: there is 23 paired individuals.

In addition to the first file, this one will allow us to study links between virus and bacteria.

2 Methods

2.1 Normalisation

Let us have a look at the distribution of the first simple, with a log scale for the frequencies:

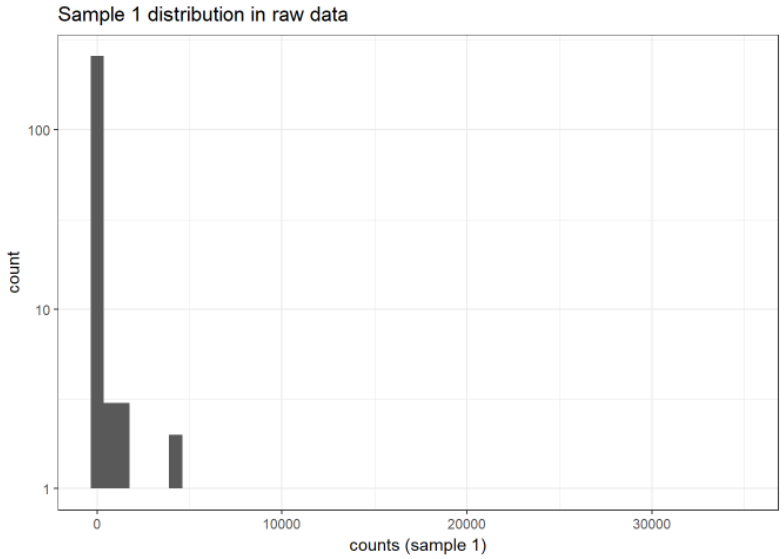

Figure 2.1: *Sample 1 distribution in raw data*

What can be noticed is that there is a lot of very low values, hence the symmetry of this distribution is completely skewed. In order to analyse this data set, we will try several transformations.

Log transformation

This transformation gives a better distrubution, but there is still a lot of very low values.

BOXPLOTS ?

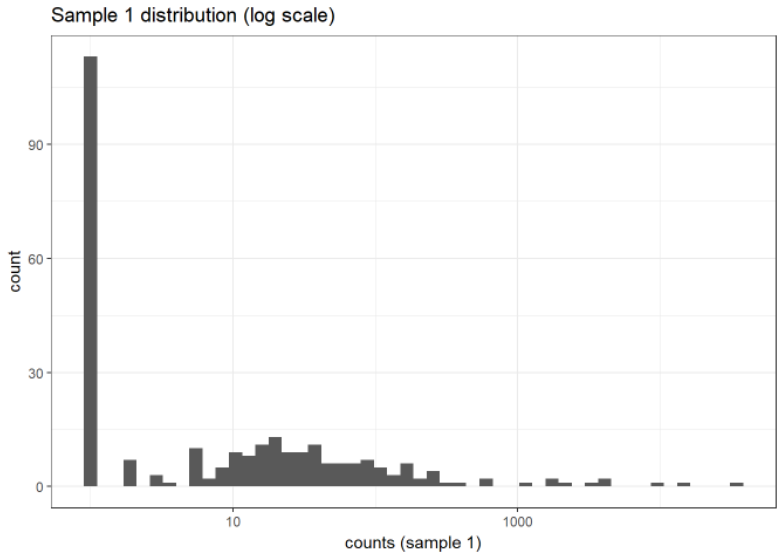

Figure 2.2: *Sample 1 distribution in log transformed data*

TSS transformation

It is commun in metagenomic datasets to perform TSS (*Total Sum Scaling*) before further normalization. TSS transformation computes relative abundances:

$$y_{ij} = \frac{n_{ij}}{\sum_{k=1}^{p} n_{ik}}$$

for n_{ij} the counts of species j in sample i, p the number of species and n the number of individuals.

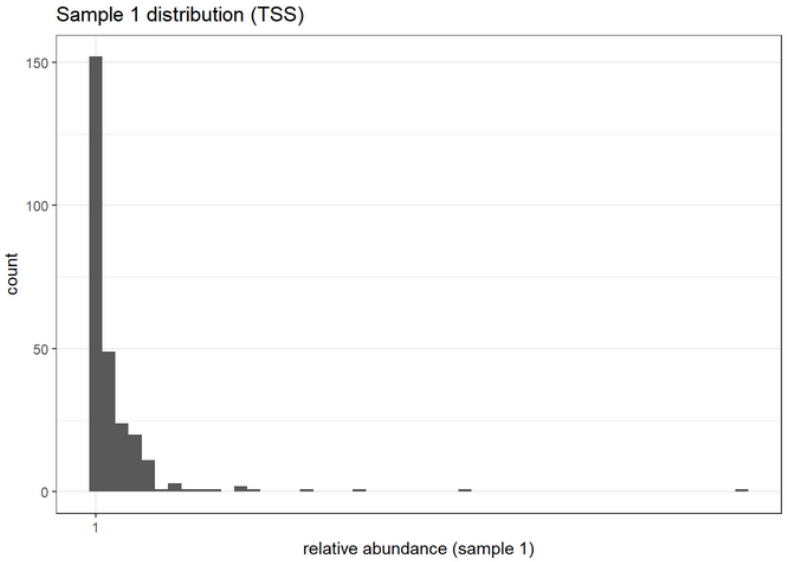

Figure 2.3: *Sample 1 distribution in TSS transformed data*

TSS+CLR transformation

CLR (*Centered Log Ratio*) transformation:

$$\tilde{y}_{ij} = \log \frac{y_{ij}}{\sqrt[p]{\prod_{k=1}^{p} y_{ik}}}.$$

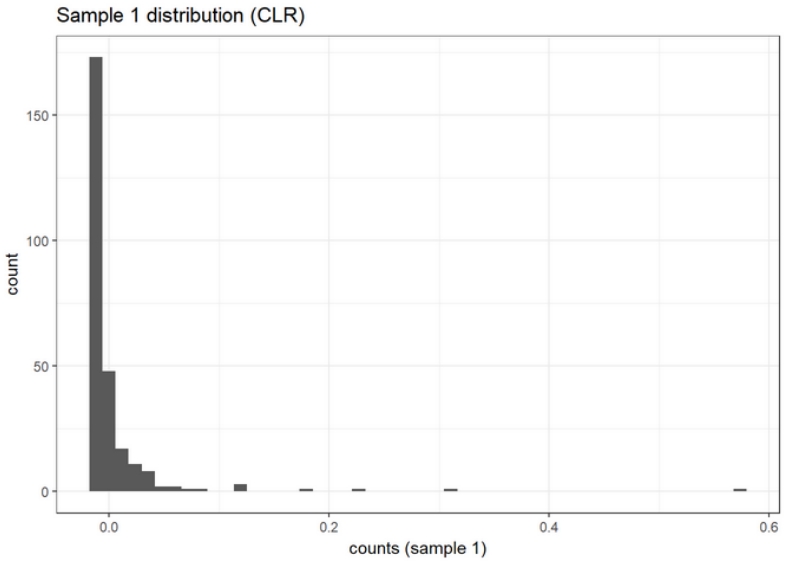

Figure 2.4: *Sample 1 distribution in TSS+CLR transformed data*

TSS+ILR transformation

ILR (*Isometric Log Ratio*) transformation

$$\tilde{\mathbf{Y}}' = \tilde{\mathbf{Y}} \times \mathbf{V}$$

for $\tilde{\mathbf{Y}}$ the matrix of CLR transformed data and a given matrix $\mathbf{V}$ with p rows and $p-1$ columns such that $\mathbf{V}\mathbf{V}^\top = \mathbb{I}_{p-1}$ and $\mathbf{V}^\top\mathbf{V} = \mathbb{I} + a\mathbf{1}$, a being any positive number and $\mathbf{1}$ a vector full of 1.

CSS transformation

CSS transformation, which is an adaptative extension for metagenomic data of the quantile normalisation used in microarray expression datasets. It is designed so as to account for technical differences between samples.

The less asymetric distribution seems to be the one obtained with the CLR transformation and the log-transformed CSS.

2.2 The mixOmics Library

2.3 Principal Component Analysis

2.4 PLS-DA

2.5 Random Forest

3 Results

Conclusion