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

Infectious bronchopneumonia represents, with diarrhea, 30% to 90% of calf breeding mortality but there is almost no information in biology litterature on these infections. The purpose of this project was to establish links between bacteria found from calf noses and those found in calf lungs, and also to explain the presence of a list of viruses using the abundance of these bacteria.

Hence, the data set under analysis gives information about infectious agents: some samples were taken from 23 calf breeding farms where the calves were suffering from infectious bronchopneumonia. We implemented statistical methods in order to analyse this data set, such as principal component analysis, sparse partial least squares discriminant analysis, partial least squares in regression mode and random forest. Then, healthy calf breeding data was compared with these results.

[...]

Contents

Acknowledgment

Introduction

1 Context and description of the data set

1.1 Context

[...]

1.2 Dataset description

Three text files are available for the study:

- `abundances`
- `pathogenes`
- `abundances_ctrl`

They are all 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 microbiota data. We call microbiota the set of bateria and micro-organisms in a body. In our case, it is in calf bodies, especially in the nasal cavities and in the lungs. In this file, there is information (counts) about each bacteria found within samples.

At the begining, there is 54 columns and 406 rows. A certain number of pretreatments have been performed in response to somes findings on the dataset:

- **First nine columns removed**. They contained 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;
- **Duplicates of bateria merged**. The list of bacteria names is not unique. We could have used the blast identifier but the problem remains the same. First, let us keep only the specie name. It is the last one on each row. But sometimes the specie is unknown, as we can see above. In this case, the name that appear before will be kept. Then, the replicates are merged by adding together each one of them, which lead us to a list of 270 unique bacteria.
- **Name of one bacteria corrected**. One name of the bacteria was "`&`". It is now replaced by the correct one.
- **29$^{\text{th}}$ column removed** (only for paired studies). The number of columns was odd (45), but on each calf was supposed to be taken two samples: one in the nose and the other in the lungs. It will be removed from the study when a paired analyse will be made.

These pretreatments lead us to a data frame with 270 rows and 45 columns. A part of this data frame can be seen in the Figure 1.1 below.

	10_EN	11_EN	15_EN	16_EN	17_EN	18_EN	19_EN	01_EN	20_EN	21_EN	23_EN	24_EN	25_EN	26_EN	28_E
&	39	0	6	15	7	9	26	16	39	8	0	0	112	178	
[Eubacterium] coprostanoligenes group	18	0	1	2	70	28	52	54	38	74	27	27	0	11	
[Ruminococcus] gauvreauii group	9	14	0	1	0	27	39	32	0	0	0	0	27	87	5
Acetobacteraceae bacterium SAP1007.2	15	0	0	1	3	0	6	9	0	0	0	0	0	0	
Acholeplasma laidlawii PG-8A	0	0	2	2	15	16	14	31	61	21	0	0	182	84	
Achromobacter sp.	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
Achromobacter spanius	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
Actinoalloteichus cyanogriseus	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
Aerococcaceae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Aerococcus sp.	25	0	2	0	0	4	0	0	96	102	10	5	30	44	
Aeromicrobium	0	0	2	2	0	0	0	2	0	0	1	0	0	0	
Agreia	0	0	0	0	0	0	0	0	0	0	0	1	7	0	
Agrobacterium tumefaciens	0	0	2	0	0	0	0	0	0	0	2	2	0	2	
AKAU3644	0	1	0	0	0	0	0	0	6	0	1	0	0	1	
Alcaligenes faecalis	1	44	0	0	0	2	16	8	23	38	5	14	10	21	
Alcaligenes sp.	88	67	0	0	23	13	19	17	24	23	43	43	67	60	
Alloprevotella	0	0	1	0	0	0	0	14	32	0	3	0	0	1	228
Alysiella crassa	0	5	0	1	0	0	0	0	2	0	1	0	0	0	
Anaerosporobacter	0	0	0	0	0	0	3	0	0	0	0	0	0	0	
Aquamicrobium	0	0	43	46	14	31	71	96	48	20	0	24	33	85	1
Arcobacter	12	21	1	2	0	0	0	0	79	109	11	0	11	59	2
Arcobacter cryaerophilus	0	0	7	3	9	9	35	0	22	9	0	0	0	9	1
Arenimonas	0	0	0	0	1	0	0	0	0	0	1	0	2	0	
Arthrobacter	183	154	4719	4920	6	25	188	261	456	531	9	112	273	398	2
Arthrobacter arilaitensis	0	0	8	2	30	7	19	6	12	0	0	0	0	11	

Figure 1.1: *`Abundances` data*

The `pathogenes` file

The second file is composed of 9 columns and 46 rows. This file contains the presence or not of seven viruses for each farm in the two condition. Some pretreatments have been performed:

- **Identifier and condition merged**. The first and the last columns stand for the identifier of each farm and the sampling condition. A unique identifier is needed, hence these two columns are merged.
- **Binary code for presence/absence of virus**. The presence of a virus was coded with a letter, and the absence with 0. In order to simplify the analysis, it will by replaced as 1 and 0 respectively.

Unlike the previous file, there is no missing calf: there is 23 paired individuals. A part of the data set is displayed in the Figure 1.2.

The viruses observed are the following:

- "Ct.RSV" for respiratory syncytial virus
- "Ct.PI.3" for parainfluenza virus
- "Ct.Coronavirus" for coronavirus
- "Ct.P.multocida" for pasteurella multocida virus
- "Ct.M.haemolytica" for mannheimia haemolytica virus
- "Ct.M.bovis" for mycobacterium bovis virus
- "Ct.H.somni" for histophilus somni virus

	Ct.RSV	Ct.PI.3	Ct.Coronavirus	Ct.P.multocida	Ct.M.haemolytica	Ct.M.bovis	Ct.H.somni
10_EN	0	0	1	1	0	0	1
11_EN	0	1	1	1	0	0	0
15_EN	0	0	1	1	1	0	1
16_EN	0	0	1	0	0	0	0
17_EN	0	0	0	1	0	0	0
18_EN	0	0	0	1	0	0	1
19_EN	0	0	1	1	0	0	0
01_EN	0	0	1	1	0	0	1
20_EN	0	0	1	1	0	0	1
21_EN	0	0	1	1	1	0	0
23_EN	0	0	1	1	0	1	0
24_EN	0	0	0	1	0	0	1
25_EN	0	0	0	1	1	1	1
26_EN	0	0	0	1	1	0	0
28_EN	0	0	0	1	1	0	0
30_EN	0	0	0	1	1	0	0
31_EN	0	0	1	1	0	0	0
03_EN	0	0	1	1	1	0	1
04_EN	1	1	1	1	0	0	0
06_EN	0	0	0	0	0	0	0
07_EN	0	0	0	1	1	0	0
09_EN	0	0	1	1	0	0	1
10_LBA	0	0	1	1	0	0	1
11_LBA	0	1	0	1	0	0	0
15_LBA	0	0	0	1	0	1	1
16_LBA	0	0	1	0	0	0	1

Figure 1.2: *Pathogenes data*

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

The `abundances_ctrl` file

This file has the same design as the `abundances` file; it will be treaded the same way. The first difference is that this file corresponds to the samples taken from healthy calves in 6 breeding farms, which gives us 12 samples.

The other difference concern the list of bacteria. The bacteria found in healthy individual are not the same as sick individuals. This list is shorter (293 rows) and the bacteria are not the same.

This list is still not composed by unique names of bacteria. It will be fixed by taking the last name of each row as explained before.

Now that this list is composed by unique bacteria names, let us see how much they are different from the `abundances` file in the Figure 1.3.

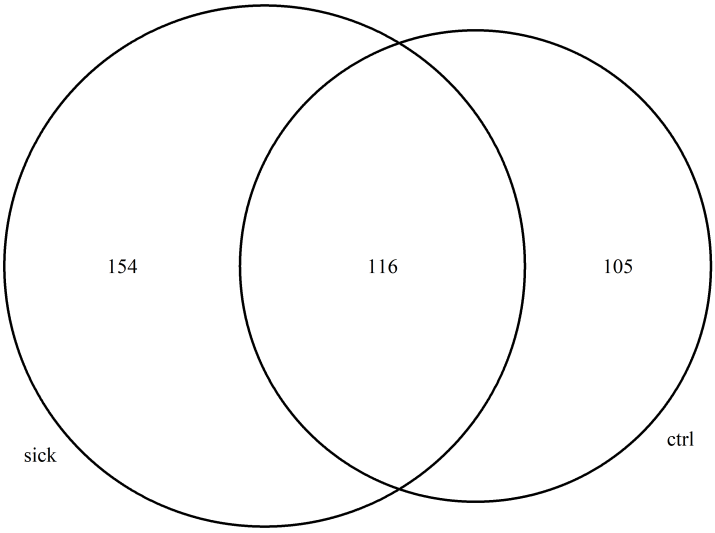

Figure 1.3: *Venn diagram comparing two lists of bacteria: in sick calves and in healthy calves*

2 Methods

2.1 Normalisation

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

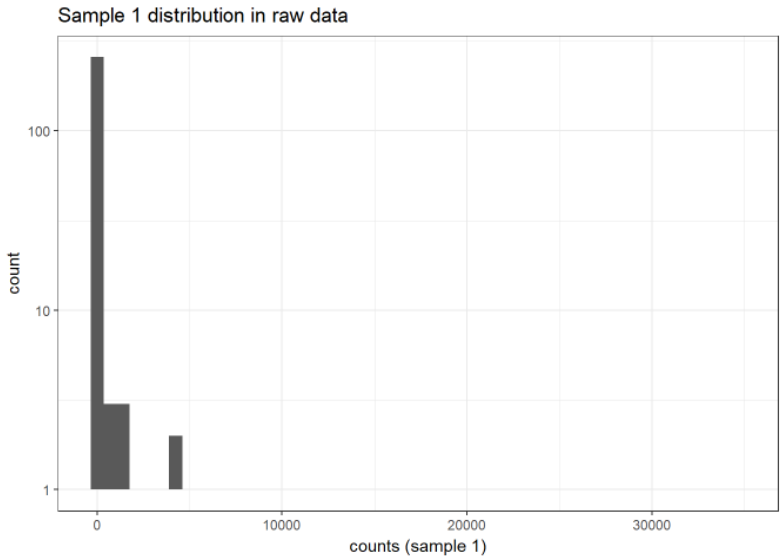

Figure 2.1: *Sample 1 distribution in raw data*

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

Log transformation

$$\tilde{y}_{ij} = \log(y_{ij} + 1)$$

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.

TSS+CLR transformation

CLR (*Centered Log Ratio*) transformation:

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

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

Principal Component Analysis (PCA) is a multidimensional descriptive method that allows to explore the links between variables and the similarities between individuals. The objective of the PCA is to identify the largest sources of variation and return to a reduced size space (for example 2 or 3) by deforming the less possible reality. In other words, PCA reduces the number of variables, in this way the information becomes less redundant.

From a mathematical point of view, the PCA corresponds to the approximation of a matrix (n, p) by a matrix of the same dimensions but of rank $q < p$. In fact, an orthogonal linear transformation is applied on the data to convert a set of observations of possibly correlated variables into uncorrelated principal components. q is often of small value 2, 3 and contributes to the construction of easily understandable graphs. The interpretation of these graphs helps to understand the structure of the analyzed data.

2.4 Partial least squares regression

Partial least squares regression (PLS) is a fast, efficient and optimal statistical method that is widely used to deal with situations with high multicollinearity and when big data are to be taken into account. Its use is recommended in the case where the number of variables p is much greater than the number of individuals n: $p \ll n$.

There are different versions of PLS regression depending on the objective:

- **PLS1**: A quantitative target variable Y is to be explained, model, predict by p quantitative explanatory variables X^j
- **PLS2**: Canonical version. Relates a set of q quantitative variables Y^k and a set of p quantitative variables X^j.
- **PLS2**: Regression version. Tries to explain, model a set of q variables Y^k by a set of p quantitative explanatory variables X^j.
- **PLS-DA**: Discriminant version. Special case of the previous case. The qualitative variable Y with q classes is replaced by q dummy variables of these classes.

During this project, only the PLS-regression mode and PLS-DA versions were used because they are adapted to our dataset and the objective pursued.

2.5 Random Forest

3 Results

3.1 Normalisation

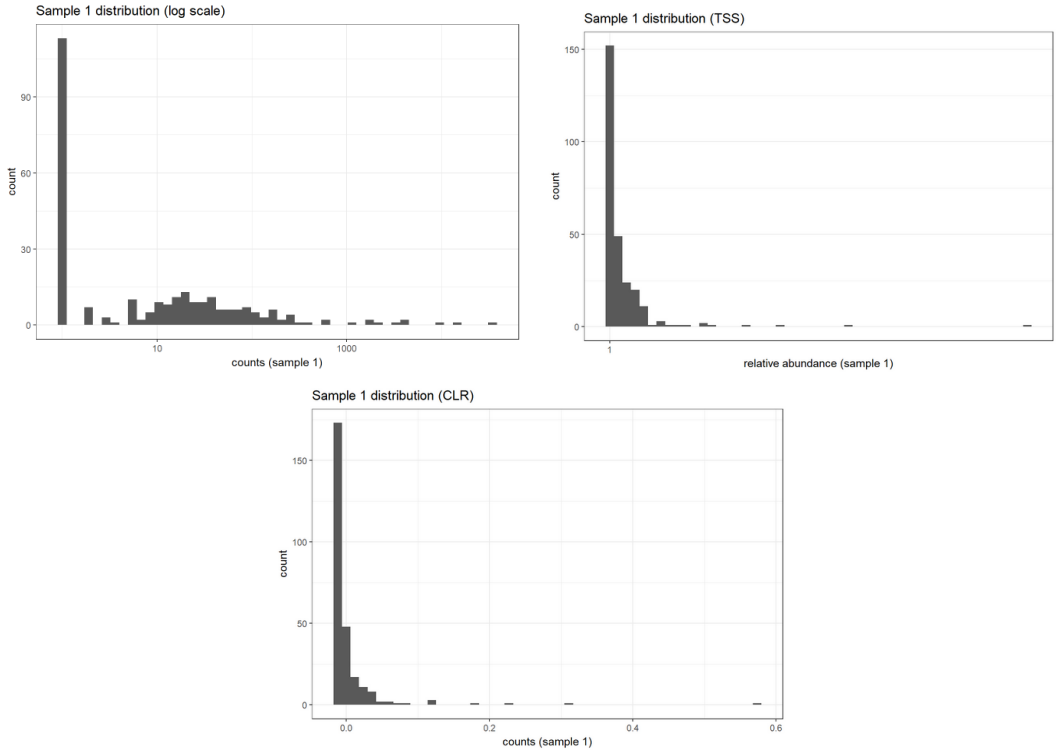

Figure 3.1: *Sample 1 distribution in log, TSS, CLR and ILR transformed data*

The log-transformation gives a better distribution, but there is still a lot of very low values.

3.2 Principal Component Analysis

It is important to have a clear idea of the data structure. A principal component analysis is adapted to this objective. In the following, we will analyze the results of the PCA on the Log and TSS + CLR transformations made on the observation matrix because the two others show very similar results.

Log transformation

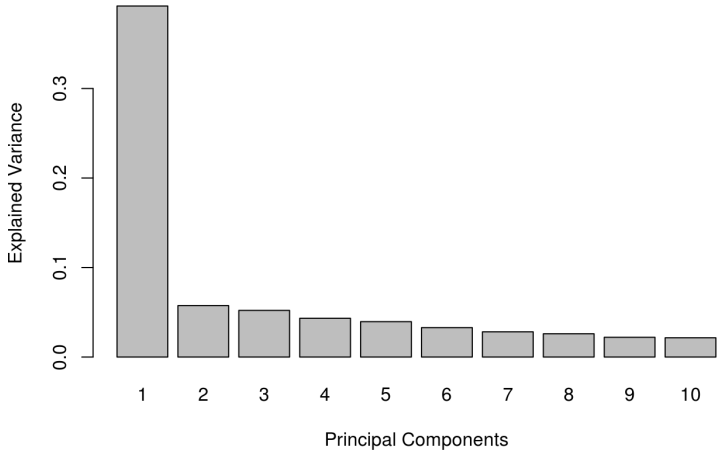

Figure 3.2: *Sample 1: The variance explained by the first 10 components*

This graph represents the decay of the eigenvalues associated to the principal components. It seems that the first component on 406 is sufficient to reconstitute the data. This is explained by the rapid decay of the first eigenvalue, which are not significant beyond that. The first axis, showing the projection of the data on the first principal component seems to illustrate its stability in the construction of the representation subspace.

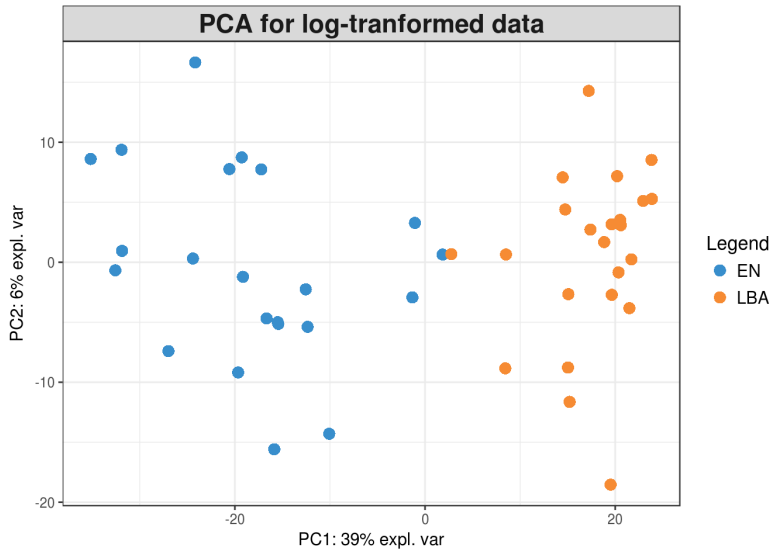

Figure 3.3: *Sample 1: Projection of individuals in the first factorial plane*

The projection of individuals in the first factorial plane shows that the separation of "EN" and "LBA" classes is marked. Also, it can be seen that the first axis associated to the first principal component explain 39% of variance and thus, contributes mainly to the reconstitution of the data.

Conclusion