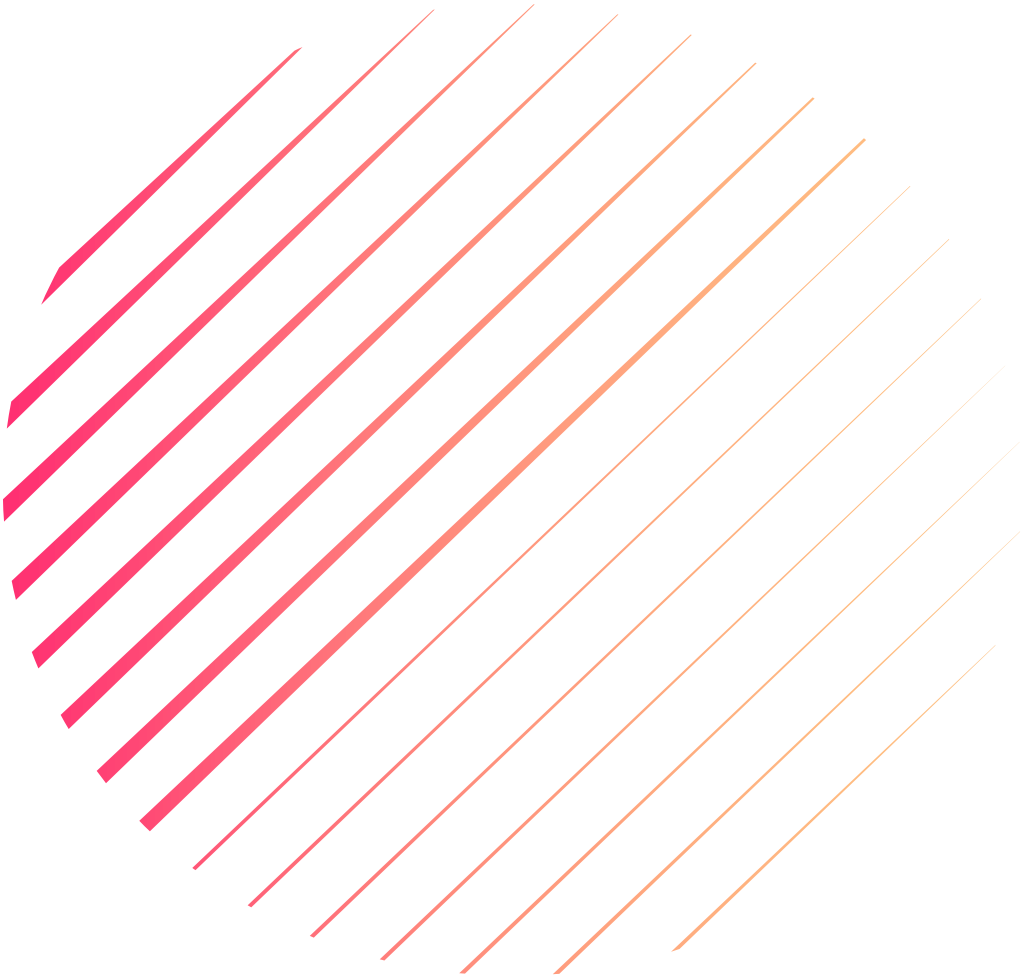
Principal Component Analysis of Gram-Positive 16s rRNA Sequences



Syed Ali

April 9th, 2021

Introduction

To understand relationships between organisms a considerable number of methods can be applied. The most popular methods are clustering methods (hierarchical and non-hierarchical) and within this unsupervised learning methods realm are several options of which biologist prefer to use called phylogenetic trees (hierarchical) (1). The construction of phylogenetic trees is primarily on similarities and differences between organisms based on genetic and physical features. (1). Phylogenetic trees like all other approaches come with certain disadvantages such as, if the sequences provided is not up to good quality, then false relationships could be built leading to false conclusions (1). Within the domain of construction of phylogenetic trees there are different methods as well, such as neighbour-joining and maximal likely hood and depending on which methods you use conclusions about evolutionary relationships can be different. Phylogenetic trees though provide good visualization they are limited to datasets or sequences of small number otherwise if you have a large number of sequences the tree might be built but becomes too difficult to analyze. Thus, for data sets that are extraordinarily large two machine learning methods can be used, one which is in the realm of deep learning called Sequence Similarity Networks and the other a division of unsupervised learning method called Principal Component Analysis.

Principal Component Analysis is recognized as a standard and useful tool for multivariate data analysis, and it is applicable in many different kinds of big data (2). The advantage that comes with the PCA is the reduction in dimension number but still retains data variation as much as possible (2). Each of these dimensions is known as components and each component has a certain variance explained (2). The variability of each component can be visualized either as a special graph known as scree plots and it allows to understand which components need to be taken into consideration. PCA indeed can be useful at analyzing the distances between 16s rRNA sequences and the reason why you want to visualize it in this manner is that it gives a good indication of how similar the species or genus or group of organisms are (2). It is also indicating to what degree some members of particular genus sway or diverge towards a different member and understanding that particular variation and what causes it to sway could potentially high light something important about certain members of a particular genus.

Considering that PCA provides a useful analysis about nucleotide sequences and protein sequences this is what has been explored. Particularly 16s rRNA sequences of gram-positive bacterial species were obtained from NCBI and it was explored to what degree they have relevance to a Mycobacterium genus an acid-fast stain group.

Material and Methods

300 sequences were acquired from the NCBI database using the Entrez package in R and the distance similarity matrix was computed. A total of 8 genus *Streptococcus*, *Staphylococcus*, *Enterococcus*, *Clostridium*, *Mycobacterium*, *Bacillus, Listeria* and *Lactobacillus* were considered. Once the similarity matrix was computed the PCA was computed and using the auto plot package the PCA was visualized as a plot. The distance matrix was subject to clustering as well. The appropriate number of clusters was determined using the clValid package in R – a package designed for different types of validation methods. The validation that was conducted was internal validation which is normally used to assess the appropriate number of clusters. The validation is determined using indexes known as silhouette and Dunn index and the clustering method used was k-means clustering. Preliminary assessment of sequences was determined as well which were the GC content of the sequences as well as sequence length distribution.

Results

As previously mentioned, 300 sequences were obtained from NCBI, and given the unique function was applied in R, 296 unique sequences were obtained. The highest GC content was claimed by the genus *Mycobacterium* with approximately 58.5%. Most genera of *Bacillus* had approximately 56% GC content but surprisingly *Bacillus thermozeamaize* had a GC content of 59.8%. Principal Component 1 showed a variance of 68.8% and PC2 represented a variance of 16%. There was a total formation of 5 clusters where 1 large cluster was composed of *Enterococcus*, *Bacillus* and *Staphylococcus*, *Listeria* Species. The second cluster was composed of *Streptococcus* species, 3rd composed of *Lactobacillus* species, 4th composed of *Clostridium* species and 5th of *Mycobacterium* Species. The *Streptococcus* genus was considered to be much closer to *Lactobacillus*, *Staphylococcus*, *Bacillus* and *Enterococcus* genus. The *Mycobacterium* genus was considered to be vastly different from all the genus mentioned. Surprisingly there were species of *Clostridium* that seemed to sway from their designated cluster towards the *Mycobacterium* cluster. The two species of *Clostridium* were *C. merdae* and *C. jeddahanse.* The validation results showed that according to the K-means clustering 5 clusters formation were appropriate enough having the highest indexes.

Discussion

According to Fu and Liu 2002, the Mycobacterium species is more related to gram-negative than it is too gram-positive, and the results found from the PCA confirm this as the Mycobacterium is set much further apart in the PCA plot along the lines of PCA 1 which shows 69.3% variance. Surprisingly the *C.merdae* and *C.jeddahanse* are two species that are confirmed to be much closer to the *Mycobacterium* genus and are the only members that sway to a large degree from their designated cluster which fully comprises of Clostridium species. The reason perhaps that this is the result could be because these species are found in the human gut microbiome while other members are found in the soil. Due to the consistently different environments, their sequences are much similar to *Mycobacterium* whose organism popularly infect humans such as the infamous *M.tuberculosis* species. Another surprising discovery made

was that *Bacillus* species are much closer to *Staphylococcus* species. Which is something not discovered yet or mentioned in the literature. *Bacillus* species are either found in the soil or human gut microbiome and they are indeed a spore-forming species while *Staphylococcus* do not have any of these qualities. There is mentioned that *Bacillus* species do compete with *Staphylococcus* with regards to the human microbiome by producing an antibiotic and perhaps this competition could cause them to be more similar in nature.

Reflection

Considering that I have never applied PCA on sequences before this was an interesting experience and when it comes to analyzing the 16s rRNA sequences provides much more stable answers and in-matter more unique answers as well. The reason why I chose this particular analysis was of the curiosity to what extent the Mycobacterium species was different from gram-positive species as it has not been fully explored in the many pieces of literature I have read. The workflow was inspired by Konishi *et al* 2019 and was quite easy to follow. The code, however, they posted in GitHub posed significant challenges with regards to using Boolean vectors and the results were largely presented oddly thus that particular part of their workflow was disregarded. I had hoped to make a neural network of the sequences as well which indeed is a unique deep learning method but unfortunately was unsuccessful. In future, I would like to further apply deep learning method approaches to understanding complicated methods and how to use the R-shiny package to create interactive web applications to analyze 16s rRNA sequences with ease.

References

1. Wan, P., & Che, D. (2013). Constructing phylogenetic trees using interacting pathways. *Bioinformation*, *9*(7), 363–367. https://doi.org/10.6026/97320630009363

2. Konishi, T., Matsukuma, S., Fuji, H., Nakamura, D., Satou, N., & Okano, K. (2019). Principal Component Analysis applied directly to Sequence Matrix. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-55253-0

3. Fu, L. M., & Fu-Liu, C. S. (2002). Is Mycobacterium tuberculosis a closer relative to Gram-positive or Gram-negative bacterial pathogens? Tuberculosis, 82(2–3), 85–90. https://doi.org/10.1054/tube.2002.0328

Figures

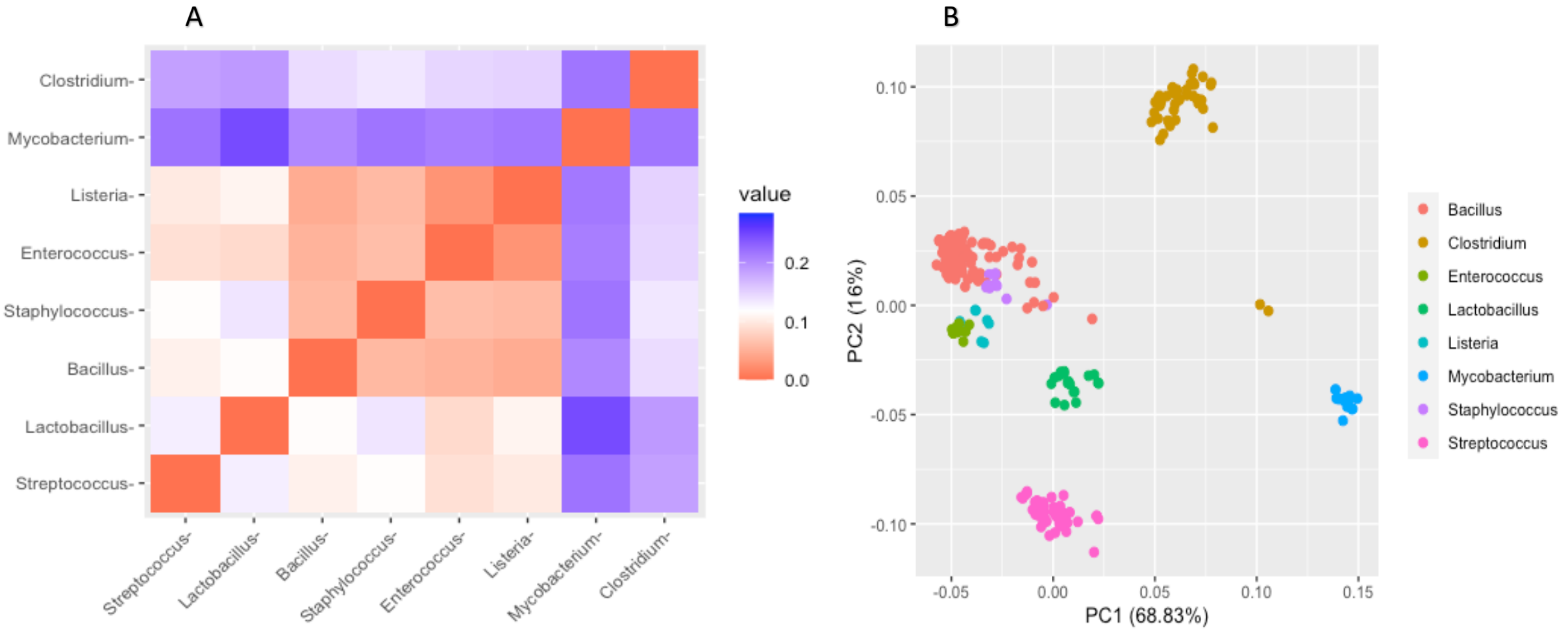


Figure 1 distance Matrix Visualization. A) Shows the distance matrix with visualization on similarity. Red indicated high similarity while Purple indicates high dissimilarity. B) Show the PCA plot of distances computed between 16s rRNA sequences.

```
---
Author : Syed Shahzaib Ali
title: "PCA Analysis of Gram-Positive"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(rentrez)
library(stringr)
library(DECIPHER)
library(tidyverse)
library(ggplot2)
library(seqRFLP)
library(ape)
library(ggfortify)
library(cluster)
library(bpca)
library(clValid)
library(factoextra)
library(autoplotly)
library(seqinr)
```

Searching for Sequneces and Building a Data-frame

The Renterez package is applied here to search the sequences and and
extract them from the NCBI database. The sequneces were saved in E.data
object and finally written in Fasta format and saved in your current
directory. The fasta format was read as a DNA string and converted to a
data-frame format where the names were trimmed to produce species names and
Genus names in the data-set.

```{r setup, include=FALSE}
E.search <- entrez_search(db = "nuccore", term = "Streptococcus[ORGN] OR
Staphylococcus[ORGN] OR Mycobacterium[ORGN] OR Clostridium[ORGN] OR
Bacillus[ORGN] OR Enterococcus[ORGN] OR Listeria[ORGN] OR
Lactobacillus[ORGN]   AND 16S AND biomol_rRNA[PROP]", retmax = 300,
use_history = TRUE)

E.data <- entrez_fetch(db = "nuccore", rettype = "fasta", id =
E.search$ids)

```


```{r setup, include=FALSE}

write(E.data, "Gram.fasta", sep = "\n")

Gram.string = readDNAStringSet("Gram.fasta")

Gram.string.1 <- data.frame(Title = names(Gram.string), Sequence =
paste(Gram.string))
```
```

```
Gram.string.1$Species_Name <- word(Gram.string.1$Title, 2L, 3L)
Gram.string.1$Genus <- word(Gram.string.1$Species_Name, 1L)


```

PreLimnary Analysis Part 1

When data is retrieved some of the sequneces are reperated thus the unique function is applied to keep the diversity and representation of choosen Genus to a considerable high level. The missing nucleotides and the gaps in the sequence are asseed to judge the quality of these sequences as low quality sequneces interfere with accurate analysis.

```{r setup, include=FALSEE}
Gram.string.1 <- Gram.string.1[, c("Species_Name", "Sequence","Genus")]
Unique.Gram <- unique(Gram.string.1)

View(Unique.Gram)

mean(nchar(Unique.Gram$Sequence))

str_count(paste(Unique.Gram$Sequence, collapse=""), 'N')
str_count(paste(Unique.Gram$Sequence, collapse=""), '-')

summary(nchar(Unique.Gram$Sequence))

Edit.Gram <- Unique.Gram %>% mutate(S_Sequence = str_remove(Sequence, "^[-N]+")) %>% mutate(Sequence = str_remove(Sequence, "^[-N]+$"))

str_count(paste(Edit.Gram$S_Sequence, collapse=""), 'N')
str_count(paste(Edit.Gram$S_Sequence, collapse=""), '-')

Gram.string.2 <- data.frame(Title = names(Gram.string), Sequence = paste(Gram.string))

Gram.string.2$Title <- word(Gram.string.2$Title, 2L, 3L)

Final.u = unique(Gram.string.2)

```

Pre-Liminary Analysis Part 2

In this particular step ggplots are created. 1) A histogram of the lenghts of the sequneces is created because if the lenghts are extremely un-equal in interferes with computation of distance matrix with all the values or most values being generated as NAN. The data-frame is saved as a fasta file. The GC-content for each sequecnce is also determined as well.

```{r setup, include=TRUE}
ggplot(Edit.Gram) +
  geom_histogram(aes(x = nchar(S_Sequence)), binwidth = 100, color = "plum2", fill = "plum2") +
```

```
  labs(x = "Sequence Length (nucleotides)", y = "Number of Sequences",
title = " 16S rRNA Sequence Length Distribution")

dataframe2fas(Final.u, "Gram_P.fasta")

DNA.Gram <- readDNAStringSet("Gram_P.fasta")

DNA.Gram.1 <- ((letterFrequency(DNA.Gram, letters = c("G"), as.prob =
FALSE)) + (letterFrequency(DNA.Gram, letters = c("C"), as.prob = FALSE))) /
((letterFrequency(DNA.Gram, letters = c("A"), as.prob = FALSE))+
(letterFrequency(DNA.Gram, letters = c("T"), as.prob = FALSE)) +
(letterFrequency(DNA.Gram, letters = c("C"), as.prob = FALSE)) +
(letterFrequency(DNA.Gram, letters = c("G"), as.prob = FALSE)))

Gram.Names <- Unique.Gram$Species_Name

GC.data <- data.frame(Gram.Names, DNA.Gram.1)

View(GC.data)

ggplot(GC.data, aes(Gram.Names, G)) + geom_col(show.legend = TRUE) +
  labs(x = "Species Name", y = "GC proportion", title = " 16S rRNA Sequence
GC proportion")


```

The sequence Alignment is conducted below and the distance matrix is computed. The TN93 model was used as it provides accurate distance matrix numbers.


```{r setup, include=FALSE}

DNA.Gram  <- OrientNucleotides(DNA.Gram , processors = NULL)
DNA.Gram.alignment <- AlignSeqs(DNA.Gram)
DNA.Ad <- AdjustAlignment(DNA.Gram.alignment)

Pair.Gram <- as.DNAbin(DNA.Ad)

class(Pair.Gram )

Gram.Dist.DNA <- dist.dna(Pair.Gram, model = "TN93", variance = FALSE,
gamma = FALSE, pairwise.deletion = FALSE,
                            base.freq = NULL, as.matrix = TRUE)

```

The prinncipal component analysis is conducted based on the distance matrix and the autoplot is used as to plot the principal component analysis. The scree plot is generated to understand how much variation each PC represents and the clustering formation if also conducted using clusplot based on the pam method which is a robust version of k-means method.

```{r setup, include=TRUE}

Gram.pca1 <- prcomp(Gram.Dist.DNA, scale. = FALSE)

autoplot(Gram.pca1, data = Unique.Gram, colour = "Genus", size = 2)

fviz_dist(as.dist(Gram.Dist.DNA))

screeplot(Gram.pca1)

K.gram <- pam(Gram.Dist.DNA , 5)

Gram.plot.clust <- clusplot(K.gram, labels = 2, sub = '', main = "5
Clusters Using K-means")

Gram.plot.clust

```


Here the valdiation of the cluster is generated and it was determined that 5 clusters used are appropriate based on k-means. The Principal component plot is also generated, a more interactive version to determine which species are present and which are swaying from their designated clusters. The heatmap is also generated for a different visualization of similarity between species.

```{r setup, include=TRUE}

Gram.valid = clValid(DNA23S.Dist.DNA, 2:6,
clMethods=c("hierarchical","kmeans"), validation="internal")

summary(Gram.valid)

PCA.species <- autoplotly(prcomp(Gram.Dist.DNA), data = Unique.Gram,frame =
TRUE, colour = "Species_Name")
PCA.species

heatmap(Gram.Dist.DNA)

```