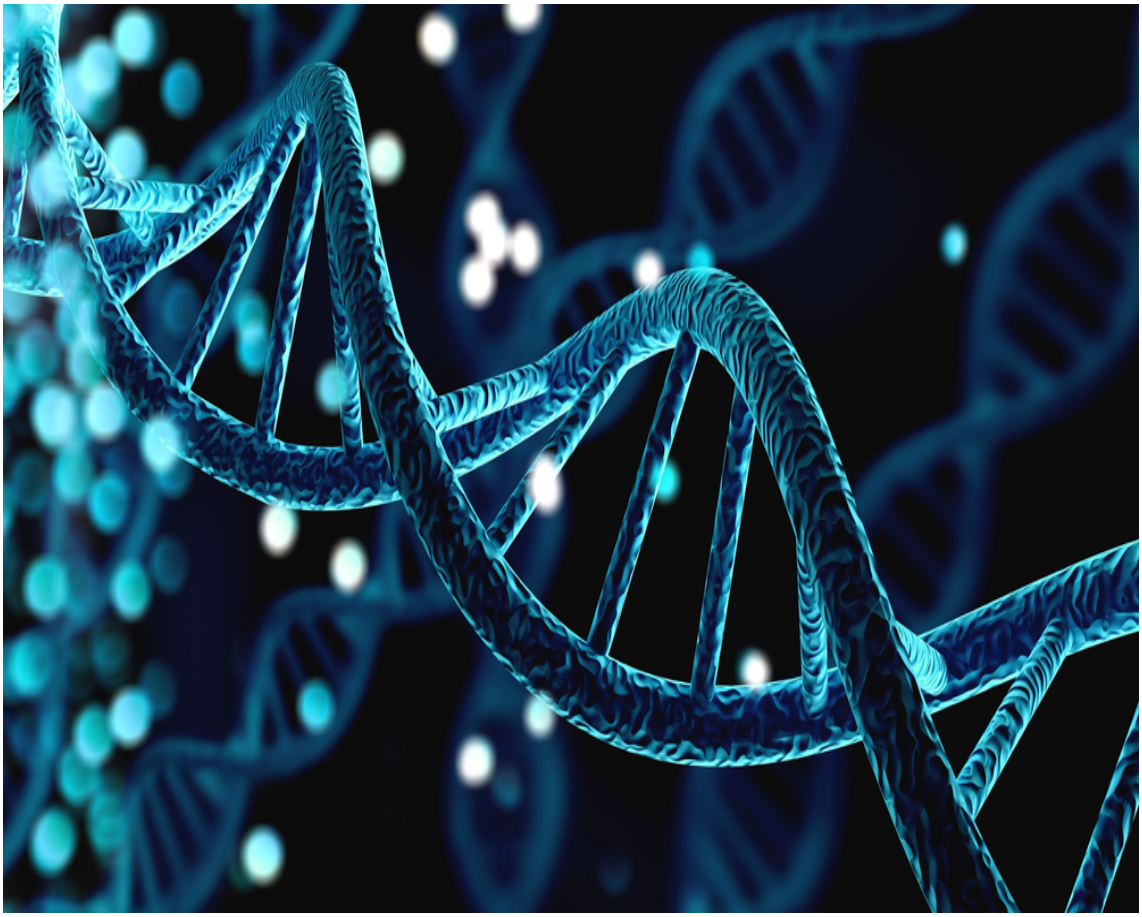
Genetic Analysis of Bacteria Using PCA

Group 10



Kashish Kothari *(202118011)*
MSc DS, DAIICT
Gandhinagar, Gujarat
kashishk9.9.00@gmail.com

Pratik Patil *(202118023)*
MSc DS, DAIICT
Gandhinagar, Gujarat
prattikk.pattill@gmail.com

Jeffrey James *(202118031)*
MSc DS, DAIICT
Gandhinagar, Gujarat
trij806175@gmail.com

Tanya Jagyasi *(202118039)*
MSc DS, DAIICT
Gandhinagar, Gujarat
tanyajagyasi@gmail.com

April 25, 2022

Abstract

The importance of bacteria in medicine, pharmaceutical businesses, and biotechnology make the study and analysis of bacterium species vital.We are going to examine the genetic information of 10 different species of bacteria using PCA and try to present the results with good accuracy rate.

1 Introduction

Bacteria is a unicellular, microscopic organisms which survive in huge numbers, approximately millions. Few of these bacteria are harmful but majority of them are useful to human lives and the environment. These organisms have been found to be the first living thing on the planet for around 4 billion years ago. They support many life forms on earth including humans, plants and animals.

Bacteria Genetics is study of composition,structure and inherited data present in the microorganisms. This study is been used across the world in various industries which includes medical industries, food and beverages industries, agricultural industries and many more.

In this project we will examine genetic information applying Principal Components Analysis (PCA) and build a model that predicts and can classify species of bacterial.

2 Data Description

This section contains all the aspects of data from data description, collection and pre- processing. To build the models, we took the dataset of train dataset and test datasets seperately . The attributes in the dataset are :

Field Name	**Description**
row$_i d$	Serial Number
A0T0G0C10	Gene
A0T0G1C9	Gene
...	....
A10T0G0C0	Gene
Target	10 different species of Bacteria

	row_id	A0T0G0C10	A0T0G1C9	A0T0G8C2	...	A9T0G1C0	A9T1G0C0	A10T0G0C0	target
0	0	-9.536743e-07	-0.000010	-0.000043	...	-0.000010	-0.000010	-9.536743e-07	Streptococcus_pyogenes
1	1	-9.536743e-07	-0.000010	-0.000043	...	-0.000010	-0.000010	-9.536743e-07	Salmonella_enterica
2	2	-9.536743e-07	-0.000002	0.000001	...	0.000008	0.000019	1.046326e-06	Salmonella_enterica
3	3	4.632568e-08	-0.000006	-0.000003	...	0.000015	0.000046	-9.536743e-07	Salmonella_enterica
4	4	-9.536743e-07	-0.000010	-0.000043	...	-0.000010	-0.000010	-9.536743e-07	Enterococcus_hirae

5 rows × 288 columns

Figure 1: Dataset

The data in this study is based on compressed DNA snippet measurements. Raman spectroscopy, which calculates the histogram of bases in the snippet, was used to examine snippets of length 10. The DNA sequence $ATATGGCCTT$ is translated to $A2T4G2C2$ using this method.

The training set consists of 2,00,000 bacteria and testing data set consists of 1,00,000 rows across 10 different species including Bacteroides fragilis, Campylobacter jejuni, Enterococcus hirae, Escherichia coli, Escherichia fergusonii, Klebsiella pneumoniae, Salmonella enterica, Staphylococcus aureus, Streptococcus pneumoniae, Streptococcus pyogenes.

2.1 Data Pre-Processing

In data pre-processing, we have check for the testing set whether we have any null values with help of 'isna' and the later we have also checked for the duplicate values to remove it. Our data has no null or duplicate values to be removed, hence we move ahead doing the same process with training set.

3 Exploratory Data Analysis

We have done the analysis of the data :

3.1 Bar Graph

The bar graph shows that the class distribution in our target variable is evenly balanced with 10% of each species represented in the data.

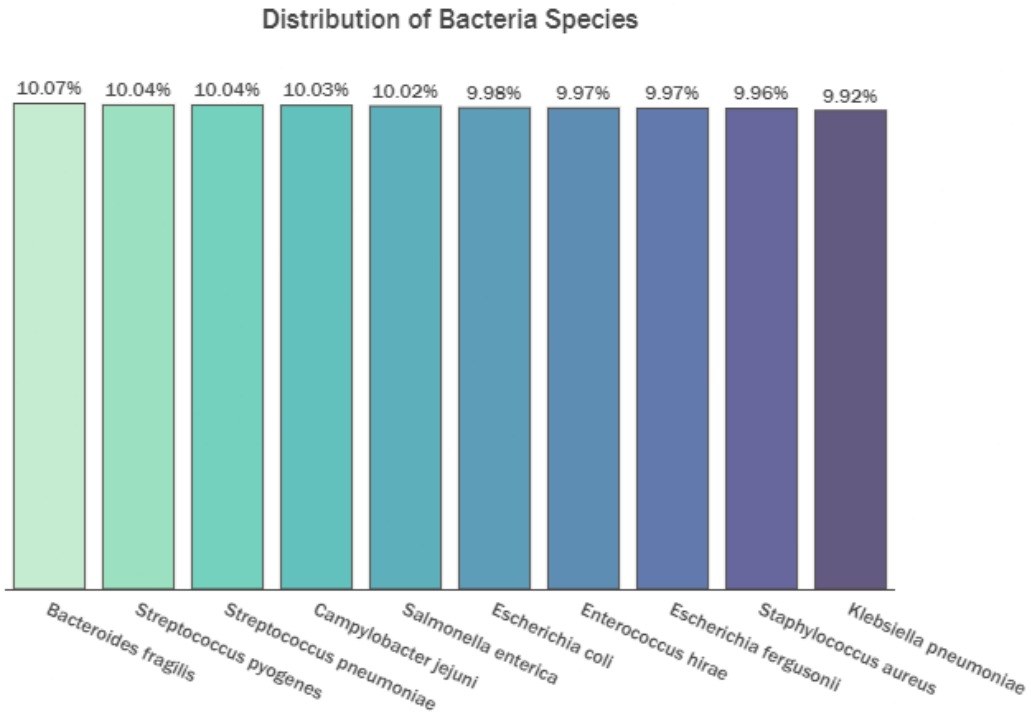


Figure 2: Distribution of species

3.2 Correlation between Genes

Using corr(), we have plotted a corrplot which finds the correlation between each gene enlisted. The more negative the correlation become the deeper the colour of plot gets and lighter the colour, positive the value is.

	row_id	A0T0G0C10	A0T0G1C9	A0T0G2C8
row_id	1.000000	-0.000480	0.000862	0.001132
A0T0G0C10	-0.000480	1.000000	-0.000668	-0.001438
A0T0G1C9	0.000862	-0.000668	1.000000	0.000018
A0T0G2C8	0.001132	-0.001438	0.000018	1.000000

Figure 3: Correlation between different genes

4 Methodology

We have a data set whose dimension is very large and so applying models took a long time. We used the concept of Principal Component Analysis to solve this problem (PCA).

4.1 Principal Component Analysis

In PCA, the dimensional of a data set consisting of a large number of interrelated variables is reduced.This is done by retaining maximum possible variation from the data. Its main motive is to reduce the size and similarity should be also less.

4.2 Why to use PCA ?

1. We can simply determine the association between observations using PCA.
2. Extracting the most important information from the data.
3. The most powerful use is that it helps in the reduction of data dimension by preserving crucial information.

4.3 Mathematical Formulation

Firstly, we define the vector space V

$$v = \sum_{i=1}^{n} a_i u_i$$

where,
a= 'n' scalars
u=the basis vectors.
So our feature vector can be transformed into a set of principal components(PC's) PCA works on Eigen value Decomposition of COV matrix. The steps to calculate PCA are:

1. We first standardize the data set by finding the means.
2. Find the correlation matrix using the formula:

$$C = \left(\frac{X \cdot X^T}{N-1}\right)$$

Using this equation:

$$|C - \lambda_i| \mid = 0$$

where,
λ = eigen value
I=identity matrix
On solving further, we take determinant on the other side and get the value of λ_1 andλ_2 which is the eigen value.
Sum of λ_1 and λ_2 =Total variance occured.

3. Find eigen vectors for eigen values by using the equation:

$$C \cdot X = \lambda \cdot X$$

4. We keep the eigen value with highest eigen vector as our Principal Component.
5. Now we transform our original data using Z=X*V, where, X = Row zero mean Data.
6. Further we reconstruct the data using:

$$\mathrm{X} = Z * V^1$$

so for ,
Row Original Data Set = Row Zero Mean Data + Original Mean

Therefore, Using the eigen vector of 1st eigen value we reconstruct the data set similar to the original data set. Thus concluding that the Principal component of the data is λ_1.

5 Experimental Results and Comparison with other models

After developing two models to predict the bacteria species using genetic sequences, this paper calculated the correlation for each corresponding genetic sequence. The genetic sequences ***A0T0G8C2*** and were found ***A1T0G0C9*** had the highest correlation coefficient of ***0.995620***. Similarly, the top 10 gene sequence pairs having the highest correlation coefficients were recognized that could be used to quickly predict the species of bacteria with high accuracy.

Along with correlation, the graph of Principal Components using cumulative and individual Variance also gives a good insight into the overall data. Similar to the Pareto chart, here too, it can be seen that as the cumulative variance increases, individual variance decreases correspondingly. This graph essentially represents that 80% of the cumulative variance is explained only by 17% datapoints.

6 Conclusion

This paper has applied the proposed PCA to produce more accurate prediction results and find sets of features (variables) that contribute the most to the classification models. The model was applied To predict bacteria species using genetic sequences, and two models were developed: one containing the information of all 286 genetic segments and a simpler method using 100 Principal Components.Applying PCA explores the data and sees how well the model would perform using a smaller set of variables. Out of the two methods, the model with all variables included achieved the highest accuracy with a score of nearly 96% on the test set. With just 100 Principal Components, 85% of the variance in the data was accounted for, and though we experienced a slight decrease in performance, the model was able to predict the species of bacteria with an accuracy of over 87%

References

[1] https://www.kaggle.com/code/kellibelcher/genetic-analysis-of-bacteria-with-pca

[2] https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643

[3] https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/

[4] https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186

[5] https://www.geeksforgeeks.org/mathematical-approach-to-pca/