Predicting Bacterial Pathogenicity Random Forest with k-mers and Chaos Game Representation Features

Zeyad Ahmed

Abstract

In this paper, we present a machine learning approach to predict the pathogenicity of bacterial species based on their DNA sequences. Supervised machine learning algorithms were used to analyze genomic signatures, each computed as the k-mer frequency vector of a DNA fragement between 8 kbp and 500 kbp. Our model achieved a mean accuracy of 87.5% in binary classification of pathogenic and non-pathogenic bacteria.

1 Introduction

Pathogenic bacteria are microorganisms capable of causing disease in humans, animals, or plants. These diseases range from mild infections to life-threatening conditions, which impacts the public health, agricultural productivity, and environmental integrity. Given the pervasive nature of bacterial pathogens, accurate prediction of their pathogenicity stands as a critical endeavor with far-reaching implications across multiple domains. In this project, we aim to predict bacterial pathogenicity using machine learning techniques applied to DNA sequences.

2 Materials and Methods

2.1 Data Collection and Pre-processing

We collected bacterial species names and pathogenic labels from Wikipedia and PubMed. Bacterial DNA sequences were obtained from the GTDB database, and only GTDB species representatives with reported completeness over 95%, and contamination under 5% were downloaded. We cleaned and pre-processed the sequences by extracting k-mers (sub-sequences of length $k = 7$) as features for classification. To generate representative k-mers for each species - mainly for EDA, the contigs of the genome were either grouped together until reaching a length of 500 kbp, or by sampling a 500 kbp fragment from a long contig. For representatives using the first method, we computed their respective k-mer frequency vector at k=7 and summed these vectors to create a representative

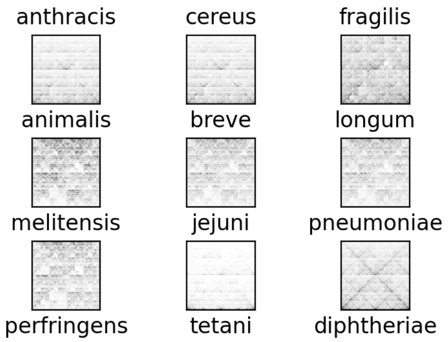

Figure 1: $fCGR_7$ of selected species in the dataset

vector for each species. To account for potential strand orientation inconsistencies, the k-mer frequencies of the reverse complement of the sequence were added to the computed k-mer frequencies in the previous step.

2.2 Exploratory Data Analysis (EDA)

To generate a visual representation of the computed k-mer frequencies we generated frequency Chaos Game Representation (fCGR) plots for each feature vector. Each plot is a gray-scale image such that the intensity of each pixel corresponds to the abundance of the sub-word with length k=7, relative to the most abundant sub-word, in other words normalized on the maximum.

The generated fCGR plots revealed patterns in DNA sequences as shown in figure 1. We explored the C+G%, G+T%, A+T%, A+G%, A+C%, and C+T% content in the dataset.

Table 1: Summary of Bases Contents

	C+G	G+T	A+T	A+G	A+C	C+T
mean	0.402999	0.496401	0.597001	0.500071	0.503599	0.499929
std	0.097238	0.015136	0.097238	0.028523	0.015136	0.028523
min	0.260961	0.442643	0.308243	0.425315	0.472469	0.391626
25%	0.338545	0.486729	0.550185	0.486555	0.492391	0.485741
50%	0.380102	0.495440	0.619898	0.497130	0.504560	0.502870
75%	0.449815	0.507609	0.661455	0.514259	0.513271	0.513445
max	0.691757	0.527531	0.739039	0.608374	0.557357	0.574685

We see that these results align with findings in literature on base contents in bacteria. To investigate how the bases contribute and shaping amino the amino acid composition, we applied Principal Component Analysis, PCA, and computed the coefficient of determination, R^2, for each pair of bases and the first component of the PCA.

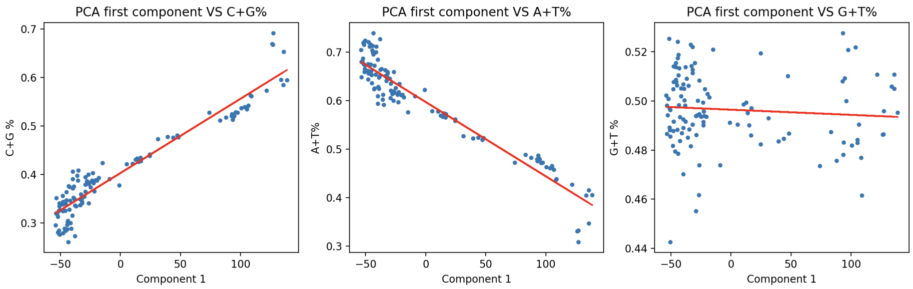

Figure 2: PCA first component VS base pairs content

Table 2: Coefficient of determination for pair bases and Component1 in PCA

	Component1	C+G	G+T	A+T	A+G	A+C	C+T
Component1	1.000000	0.905702	0.007052	0.905702	0.004582	0.007052	0.004582
C+G	0.905702	1.000000	0.012788	1.000000	0.009648	0.012788	0.009648
G+T	0.007052	0.012788	1.000000	0.012788	0.033250	1.000000	0.033250
A+T	0.905702	1.000000	0.012788	1.000000	0.009648	0.012788	0.009648
A+G	0.004582	0.009648	0.033250	0.009648	1.000000	0.033250	1.000000
A+C	0.007052	0.012788	1.000000	0.012788	0.033250	1.000000	0.033250
C+T	0.004582	0.009648	0.033250	0.009648	1.000000	0.033250	1.000000

We depict from the table that C+G and its complement, A+T, explained the most variability captured by the first component of the PCA. This result aligns with findings in literature.

2.3 Machine Learning Model

We employed the Random Forest Classifier for binary classification non-pathogenic (0) and pathogenic (1) using the fCGR plots as feature vectors after normalization. In our model, Gini-index was used as the classification criteria, with maximum depth of 3. To evaluate the model, we used Stratified K-Folds at K=10. The model accuracy ranged from 75% to 100% as shown in table 3.

3 Results

Our model demonstrated promising accuracy in predicting bacterial pathogenicity, with an accuracy range of 75% - 100% across different folds. We split the data into training and testing data sets and predicted the labels of the testing data set. The model correctly identified pathogenic bacteria 87.5% of the time.

Table 3: Random Forest Classifier Accuracy Scores on 10-Folds

	Accuracy (%)
Fold-1	75.000000
Fold-2	100.000000
Fold-3	100.000000
Fold-4	91.666667
Fold-5	91.666667
Fold-6	91.666667
Fold-7	91.666667
Fold-8	91.666667
Fold-9	91.666667
Fold-10	100.000000

4 Conclusion

This paper demonstrates the successful application of a supervised machine learning algorithm for classifying bacteria based on their pathogenicity. The use of fCGR plots as feature vectors reveals the significance of the concept of a genomic signature.

5 Future Work

The findings presented in this study opens the way for several avenues of future research:

- **Feature Engineering**: Exploring additional features or representations beyond k-mers and fCGR plots.
- **Data Augmentation**: Augmenting the dataset with additional samples, especially for underrepresented genera or species, could improve the generalization ability of the model.
- **Analysis of the DNA Structure Based on Host**: Exploring the relation between the abundance of base pairs, such as G+C, with the pathogen host to understand whether the host changes the genomic signature of the pathogen.