

Building Models for Phylogenetic Classification using Species Codon Usage Frequencies

Rahul Koonantavida
CMPE252

San Jose State University
San Jose, United States of America
rahul.koonantavida@sjsu.edu

Abstract—Codon usage frequencies, or the biased usage of synonymous codons in the genomes of species, carry valuable information that can be utilized for classification tasks in the space of biology and machine learning. In this report, these frequencies are explored as features for phylogenetic classification by building and evaluating machine learning models. Specifically, multiple random forest models are trained on codon usage data from a diverse set of organisms to predict their phylogenetic kingdom. Random forest models demonstrate high classification accuracy and robust performance, highlighting their effectiveness in capturing the complex patterns embedded in highly dimensional datasets such as genomic sequences. Results suggest that codon usage statistics, combined with interpretable machine learning models, offer a promising direction for computational phylogenetics and genome analysis.

Keywords—Machine Learning, Classification, Random Forest, Biology, Genetics, Codon Usage Frequencies, Phylogeny

I. INTRODUCTION

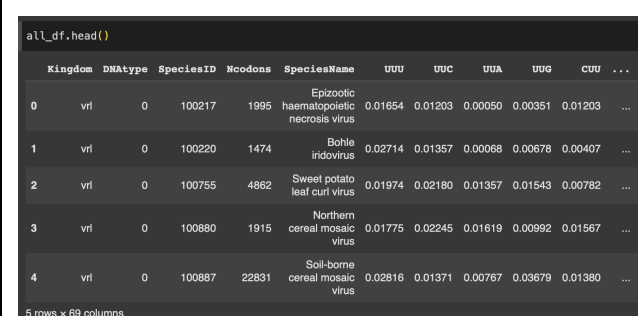
Codon usage bias—the preferential use of specific synonymous codons over others encoding the same amino acid—is a well-documented phenomenon observed across all domains of life. While the genetic code is universal, the frequency with which organisms use particular codons varies widely and is influenced by multiple evolutionary pressures. In recent years, codon usage statistics have emerged as a rich source of information for genomics and phylogenetics. They provide a quantitative feature set that can generalize across large and diverse datasets. This makes them particularly attractive for classification tasks.

In this project, the use of codon usage frequencies as features for phylogenetic classification is investigated using supervised machine learning techniques. The primary goal is to build a classification model capable of predicting the phylogenetic kingdom of a species given that species’ codon usage profile. Specifically, random forest models, known for their success with highly dimensional data, are implemented. Using curated codon usage data, these models are trained and evaluated on the task of classifying organisms into phylogenetic categories. Results demonstrate that codon usage contains sufficient discriminatory information to support accurate classification.

By framing phylogenetic classification as a machine learning problem and leveraging codon bias as the feature space, this project bridges genomics and artificial intelligence. It contributes to the growing body of work advocating for data-driven, scalable, and interpretable methods at the intersection of biology and computer science.

II. “CODON USAGE” DATASET

In order to build an effective phylogenetic classification model, high quality data is of the utmost importance. This project utilizes the “Codon usage” dataset [1] provided by the UC Irvine Machine Learning Repository. It provides 13028 instances of species of various phylogenetic kingdoms and their corresponding codon usage frequency values, some of which are illustrated in Figure 1.



	Kingdom	DNatype	SpeciesID	Ncodons	SpeciesName	UUU	UUC	UUA	UUG	CUU	...
0	vrl	0	100217	1995	Epizootic haematopoietic necrosis virus	0.01654	0.01203	0.00050	0.00351	0.01203	...
1	vrl	0	100220	1474	Bohle iridovirus	0.02714	0.01357	0.00068	0.00678	0.00407	...
2	vrl	0	100755	4862	Sweet potato leaf curl virus	0.01974	0.02180	0.01357	0.01543	0.00782	...
3	vrl	0	100880	1915	Northern cereal mosaic virus	0.01775	0.02245	0.01619	0.00992	0.01567	...
4	vrl	0	100887	22831	Soil-borne cereal mosaic virus	0.02816	0.01371	0.00767	0.03679	0.01380	...

5 rows x 69 columns

Fig. 1. Example instances of species data in the “Codon usage” dataset

Examples of classification targets are viruses, bacteria, mammals, and plants. The wide variety of types of species included in the dataset is a valuable trait in terms of implementing a generalized and unbiased model.

III. EXPLORATORY DATA ANALYSIS

With a suitable dataset in place, exploratory data analysis was conducted to identify potential routes for further optimization of the model training process. A distribution of the target classes is illustrated in Figure 2.



Fig. 2. Distribution of target classes

The abbreviations in Figure 2 in full are as follows, from left to right on the x-axis: bacteria, virus, plant,

vertebrate, invertebrate, mammal, bacteriophage, rodent, primate, archaea, and plasmid. Evidently, the dataset suffers from imbalancing issues. Given more time/resources, effort could be contributed towards acquiring supplementary data for underrepresented classes. However, given the scope of this project, no data augmentation steps were pursued. The most impacted target class is the “plasmid” class, which unfortunately only accounts for 18 of 13028 total instances in the dataset. Outside of the “plasmid” class, all target classes have over 100 training samples. Regardless, all target classes were included in the training process for the final model.

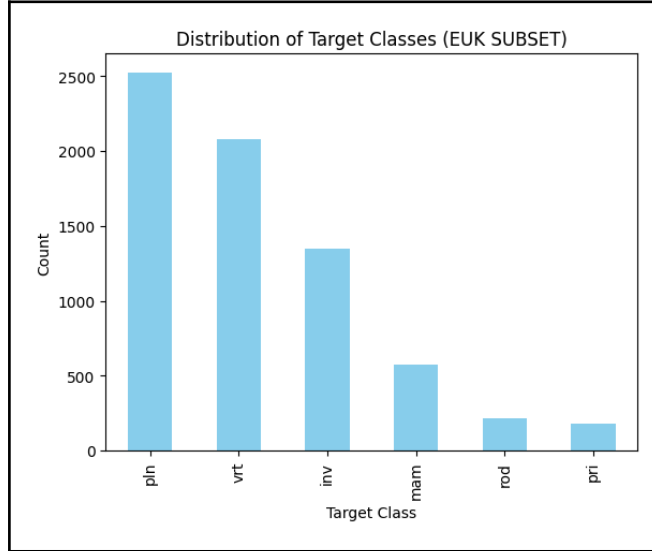


Fig. 3. Distribution of target classes for eukaryotic dataset subset

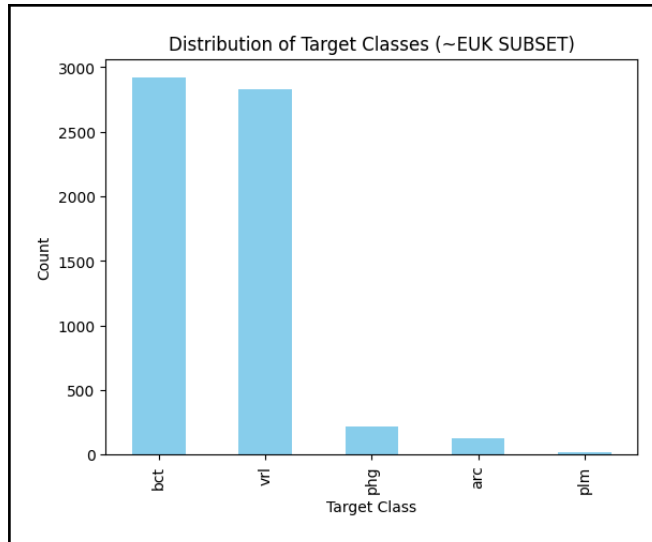


Fig. 4. Distribution of target classes for non-eukaryotic dataset subset

The distribution of target classes for two additional subsets of the primary dataset are illustrated in Figure 3 and 4. These subsets, specifying eukaryotic and non-eukaryotic species respectively, will be utilized for training additional classification models in an attempt to increase model performance metrics. The dataset imbalance issue mentioned prior is especially pronounced for the non-eukaryotic dataset subset, with the “bacteria” and “virus” classes dominating the distribution. It will be interesting to observe the impact of this during evaluation of the final model.

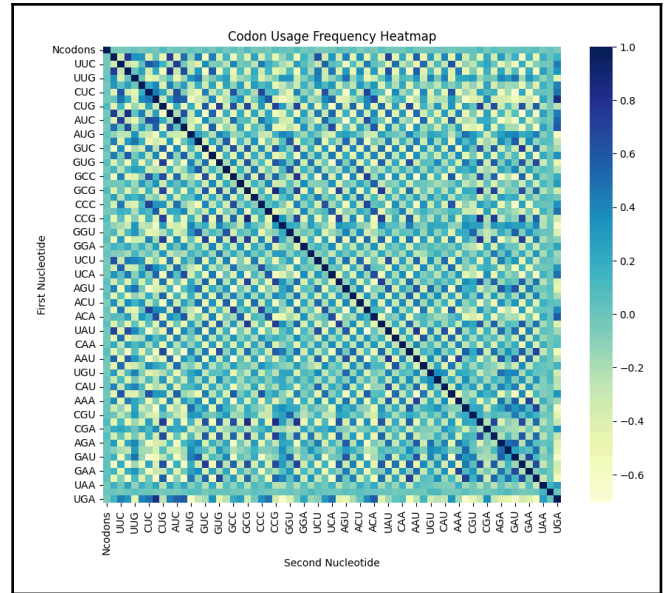


Fig. 5. Codon usage frequency correlation heatmap

Figure 5 illustrates a correlation heatmap of the features utilized for model training. The details of the heatmap need not be inspected for the scope of this project, but it is clear that particular codons have considerable correlation with others. This, paired with the high dimensionality of the dataset, suggests that principal component analysis (PCA) is a worthwhile pursuit for feature engineering. In terms of overall scalability, this is also valuable as achieving efficient model performance while decreasing computational cost is always desirable.

IV. DATA PREPROCESSING

Upon running initial experiments with the dataset, corrupted instances were uncovered. These were manually removed as the first steps of data preprocessing. Fortunately, only two of 13028 instances in the dataset were corrupted, not meaningfully impacting the model training process. After removing corrupted data, column datatypes were standardized and the dataset was partitioned into the dataset subsets mentioned in SECTION III—general, eukaryotic species, and non-eukaryotic species. Furthermore, specific training subsets not including target columns were also created. These initial steps are illustrated in the code snippet in Figure 6.

```
[3] df = pd.read_csv('codon_usage.csv')

# REMOVAL OF ROWS WITH CORRUPTED DATASET COLUMNS (manually inspected)
# FIXING COLUMN DATATYPES
df = df.drop([486, 5863])
df = df.astype({'UUU': np.float64, 'UUC': np.float64})

# DATASET SUBSETS
all_df = df
euk_df = df[df['Kingdom'].isin(['pln', 'inv', 'vrt', 'mam', 'rod', 'pri'])]
non_euk_df = df[~df['Kingdom'].isin(['pln', 'inv', 'vrt', 'mam', 'rod', 'pri'])]

# CODON DATAFRAMES FOR MODEL TRAINING (dropping identification/target values)
codon_df = all_df.drop(columns=['Kingdom', 'SpeciesName', 'SpeciesID', 'DNAType'])
euk_codon_df = euk_df.drop(columns=['Kingdom', 'SpeciesName', 'SpeciesID', 'DNAType'])
non_euk_codon_df = non_euk_df.drop(columns=['Kingdom', 'SpeciesName', 'SpeciesID', 'DNAType'])
```

Fig. 6. Initial data preprocessing (code snippet)

Additionally, the StandardScaler [2] class provided by sci-kit learn was utilized to standardize training data. This was an especially important step for this dataset in particular because the “Ncodons” feature, which represents the number of codons (usually in the thousands) in a particular species genome, is on a wildly different scale than all of the other features in the training dataset, which are codon usage frequencies between zero and one.

V. FEATURE ENGINEERING

As mentioned in SECTION III, given the high correlation between many features in the “Codon usage” dataset, PCA is a worthwhile pursuit. PCA was conducted using the PCA [3] class provided by sci-kit learn. Furthermore, it was conducted separately for each data subset, leading to similar but slightly different results.

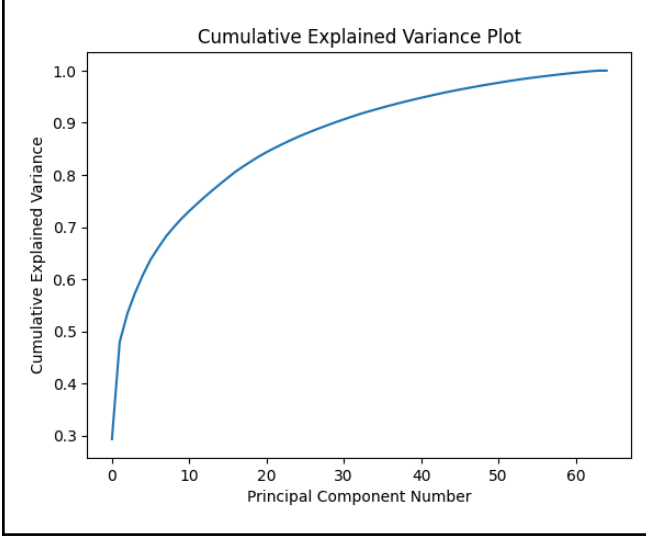


Fig. 7. Cumulative explained variance plot for PCA on primary dataset

The overall goal with PCA was to decrease the amount of features utilized for training while still capturing 95% of the explained variance for our dataset. A plot of the cumulative explained variance for PCA on the primary dataset is displayed in Figure 7. As illustrated, 95% of the variance in the primary dataset can be explained by 42 principal components. This reduction in dimensionality will decrease overall computational cost at the expense of a minor drop in performance, especially at scale. An identical approach was utilized for the eukaryotic and non-eukaryotic dataset subsets, and those subsets were reduced in dimensionality to 43 and 40 features respectively.

Principal Component (PC)	Representative Codon	PC Explained Variance
0	GGC	0.293256
1	CUA	0.186205
2	AGG	0.054969
3	GGU	0.039705
4	CUU	0.034015
5	CGA	0.029563
6	GCA	0.022922
7	GGA	0.021630
8	UAG	0.017590
9	UGC	0.016332

Fig. 8. Representative feature for PCs of primary dataset

Principal Component (PC)	Representative Codon	PC Explained Variance
0	GAG	0.239086
1	GCC	0.235572
2	GCU	0.046825
3	GGG	0.041022
4	UGC	0.032745
5	GGA	0.029567
6	CAA	0.026520
7	GGG	0.026054
8	UAG	0.018002
9	CAA	0.016830

Fig. 9. Representative feature for PCs of eukaryotic subset

Principal Component (PC)	Representative Codon	PC Explained Variance
0	GCC	0.428408
1	AGG	0.074417
2	CGU	0.040369
3	AAC	0.036340
4	CGA	0.032319
5	UUG	0.028955
6	AGC	0.024134
7	CGA	0.022117
8	Ncodons	0.018435
9	Ncodons	0.016890

Fig. 10. Representative feature for PCs of non-eukaryotic subset

Figures 8, 9, and 10 illustrate the top ten principal components (PCs) of each dataset subset as well as the feature from the original dataset that was the most representative for that particular PC. Analysis of the most influential codons with respect to phylogenetic classification is an area for extensive potential research. However, for the scope of this project, this analysis was primarily conducted out of curiosity and the hope to lay potential groundwork for further inquiry. An interesting observation to note is that for non-eukaryotic organisms, “Ncodons” or the number of codons in a particular genome is so influential that it was the most representative feature for two different PCs. Another observation is that there is considerable bias towards guanine and cytosine in terms of nucleotide representation.

VI. MODEL TRAINING

SECTIONS VI and VII are divided into sections highlighting each of the different models that were trained and evaluated over the course of this project. The baseline logistic regression [4] model is trained on the primary dataset prior to standardization and PCA. Hyperparameters for the logistic regression model were not tuned. The random forest [5] models are trained on each data subset after standardization and PCA. Hyperparameters for the random forest models are as follows: `n_estimators=1200`, `min_samples_split=2`, `min_samples_leaf=1`, `max_features='sqrt'`, `max_depth=100`, and `bootstrap=False`. These hyperparameters were adopted from the creators of the original “Codon usage” dataset; further hyperparameter tuning is a potential avenue for future experimentation. Model classification reports and brief discussions on the training processes are provided below.

A. Baseline Logistic Regression Model

The baseline model is intended to illustrate a lower bound on the classification task. Evidently, the high

dimensionality of the dataset led to the logistic regression model struggling, achieving only a 23% accuracy.

	precision	recall	f1-score	support
arc	0.00	0.00	0.00	23
bct	0.23	1.00	0.38	604
inv	0.00	0.00	0.00	291
mam	0.00	0.00	0.00	112
phg	0.00	0.00	0.00	46
plm	0.00	0.00	0.00	5
pln	0.00	0.00	0.00	495
pri	0.00	0.00	0.00	40
rod	0.00	0.00	0.00	40
vrl	0.00	0.00	0.00	575
vrt	0.00	0.00	0.00	375
accuracy			0.23	2606
macro avg	0.02	0.09	0.03	2606
weighted avg	0.05	0.23	0.09	2606

Fig. 11. Baseline logistic regression classification report

B. General Random Forest Model

The general random forest model immediately showed signs of improvement over the baseline, achieving an 89% accuracy when including all of the dataset's target classes.

	precision	recall	f1-score	support
arc	0.85	0.48	0.61	23
bct	0.88	0.96	0.92	604
inv	0.93	0.70	0.80	291
mam	0.88	0.83	0.85	112
phg	0.88	0.50	0.64	46
plm	0.00	0.00	0.00	5
pln	0.90	0.93	0.91	495
pri	0.87	0.50	0.63	40
rod	0.93	0.68	0.78	40
vrl	0.89	0.96	0.93	575
vrt	0.89	0.96	0.93	375
accuracy			0.89	2606
macro avg	0.81	0.68	0.73	2606
weighted avg	0.89	0.89	0.89	2606

Fig. 12. General random forest classification report

C. Eukaryotic Random Forest Model

The eukaryotic random forest model illustrated an accuracy improvement of two percent over the general random forest model amidst the limiting of target class scope to strictly invertebrates, mammals, plants, primates, rodents, and vertebrates.

	precision	recall	f1-score	support
inv	0.91	0.82	0.86	285
mam	0.91	0.83	0.87	117
pln	0.92	0.97	0.95	516
pri	1.00	0.65	0.79	34
rod	0.91	0.62	0.74	34
vrt	0.89	0.97	0.93	397
accuracy			0.91	1383
macro avg	0.92	0.81	0.86	1383
weighted avg	0.91	0.91	0.91	1383

Fig. 13. Eukaryotic random forest classification report

D. Non-Eukaryotic Random Forest Model

The non-eukaryotic random forest model obtained the best accuracy of all random forest models at 95%. Further discussion of model performance and metrics will take place in SECTION VII.

	precision	recall	f1-score	support
arc	0.84	0.53	0.65	30
bct	0.94	0.97	0.96	552
phg	0.92	0.60	0.73	40
plm	0.00	0.00	0.00	2
vrl	0.97	0.98	0.98	599
accuracy			0.95	1223
macro avg	0.74	0.62	0.66	1223
weighted avg	0.95	0.95	0.95	1223

Fig. 14. Non-eukaryotic random forest classification report

VII. EVALUATION

Overall, the random forest model succeeded with the task of phylogenetic classification of species given codon usage frequencies. It excels with the highly dimensional datasets that are commonplace in biological problem spaces, while being relatively computationally cheap compared to more elaborate models such as neural networks. Figure 15 provides an overview of model accuracies from the project. The best performing model was the non-eukaryotic random forest classifier, which achieved an accuracy of 95%. This was marginally better than the general random forest classifier (89%) and the eukaryotic random forest classifier (91%). This difference in performance could potentially be attributed to the relative simplicity of non-eukaryotic genomes, leading to more straightforward predictive capabilities for phylogenetic classification models.

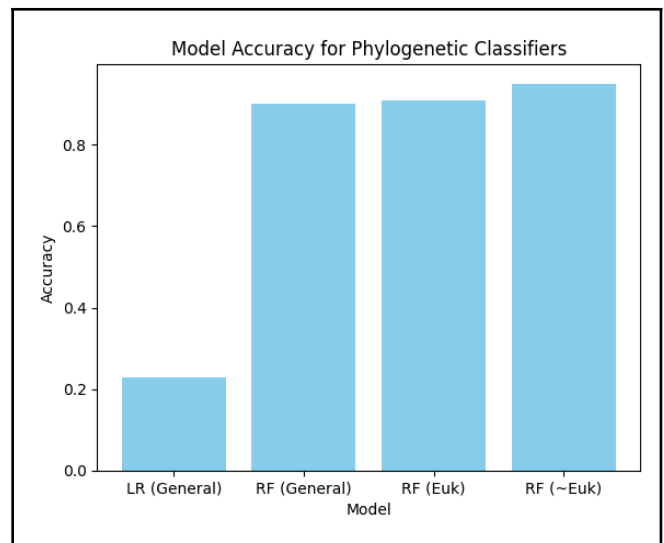


Fig. 15. Model accuracies

A. Baseline Logistic Regression Model

The rudimentary capabilities of the baseline model are illustrated in Figure 16. The model consistently predicts the "bacteria" class, leading to its extremely low overall accuracy.

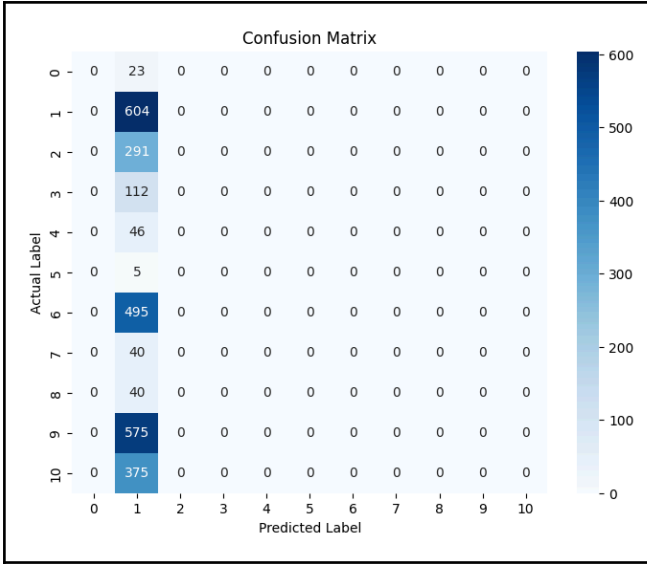


Fig. 16. Baseline logistic regression model confusion matrix

B. General Random Forest Model

The confusion matrix for the general random forest model in Figure 17 is far more appealing. The model is consistently accurate, performing the best on target classes such as “bacteria,” “invertebrate,” “plant,” “virus,” and “vertebrate.” This was predictable, as these target classes suffer far less from the data imbalance issues mentioned in SECTION III compared to other target classes in the dataset. For example, regardless of model or data subset, performance on the “plasmid” target class was poor. There were simply not enough data samples to achieve sufficient performance metrics for this class; in the future, plasmid species in particular could be explored for data augmentation. The confusion matrices for the eukaryotic and non-eukaryotic random forest models are included in the following Sections C and D. Similar observations can be made for those models.

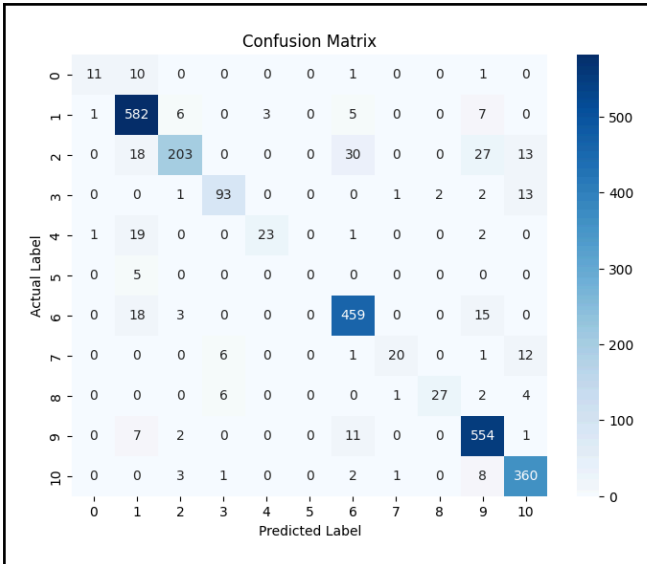


Fig. 17. General random forest model confusion matrix

C. Eukaryotic Random Forest Model

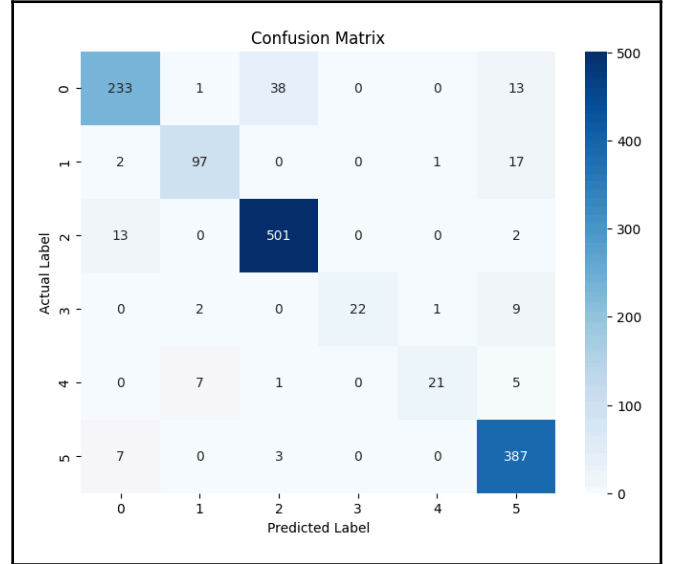


Fig. 18. Eukaryotic random forest model confusion matrix

D. Non-Eukaryotic Random Forest Model

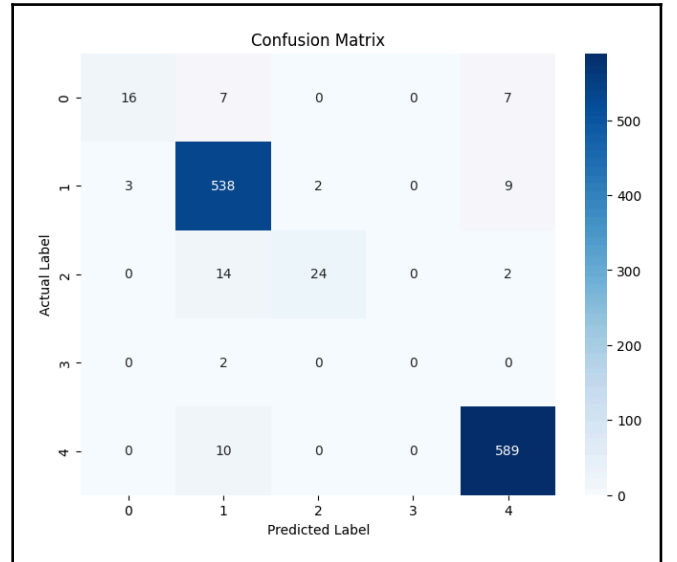


Fig. 19. Non-eukaryotic random forest model confusion matrix

VIII. CONCLUSION

In conclusion, random forest models are efficient and effective models for the classification of the phylogenetic kingdom of various different species. By leveraging the “Codon usage” dataset and conducting various data preprocessing and feature engineering techniques, I was able to build random forest models that achieved 89% accuracy or greater on the task of classifying a subset of the bacteria, virus, plant, vertebrate, invertebrate, mammal, bacteriophage, rodent, primate, archaea, and plasmid target classes given species codon usage frequencies. Avenues for future exploration are: implementing techniques for dataset balancing, analyzing representative codons across different phylogenetic kingdoms, conducting further feature engineering and hyperparameter tuning. Overall, the ability to build powerful phylogenetic classifiers is valuable for furthering society’s understanding of genetics and the relationships between the myriad of species on our planet.

ACKNOWLEDGMENTS

I would like to acknowledge Logan Hallee and Bohdan B. Khomtchouk for curating the “Codon usage” dataset and providing spectacular references for working with the data. Their foundational work was extremely valuable in my pursuit of exploring the capabilities of codon usage frequencies to predict phylogenetic kingdoms.

REFERENCES

1. Hallee, L., Khomtchouk, B.B. Machine learning classifiers predict key genomic and evolutionary traits across the kingdoms of life. Sci Rep 13, 2088 (2023). <https://doi.org/10.1038/s41598-023-28965-7>
2. “StandardScaler,” scikit, <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
3. “PCA,” scikit, <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
4. “LogisticRegression,” scikit, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
5. “RandomForestClassifier,” scikit, <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>