Using Codon Bias to Validate the Proteobacterial Origins of Mitochondria

Sebastian Espinoza

***Abstract*—Codon bias can serve as a classifier of various organisms. Specifically, this study focuses on classifying genomes across four classes of the proteobacteria phylum. Three machine learning classification models were trained using over 500 proteobacterial genomes. These models were then able to classify proteobacteria with high accuracy on a testing set and showed that it is possible to predict a proteobacterium's class by simply calculating its codon counts. These machine learning models were then applied to mitochondrial genomes to see if the classification results would suggest that mitochondria are closely related to α-proteobacteria. Although the mitochondria classification results were mixed, this study demonstrated that codon counts can be used to train machine learning models to identify taxonomic groups reliably. The datasets and machine learning models are available on GitHub (https://github.com/seespinoza/bioiSeniorProject).**

I. Introduction

The endosymbiotic theory has been established as the most likely explanation for the origin of eukaryotes and their mitochondria. There is overwhelming evidence suggesting that the event took place around 1.5 billion years ago and was why O_2 levels increased drastically [1]. In general, scientists believe that mitochondria arose from ancient α-protobacteria that became symbionts with an archaeon [2]. As a result of multiple evolutionary stages, mitochondria have integrated biochemical pathways and transport proteins within the host cell as a direct result of the endosymbiosis. This endosymbiotic event is considered the genesis of eukaryotic organisms, as the absorbed symbiont provided an important evolutionary advantage.

There are alternative hypotheses for the genesis of mitochondria that do not mention phagocytosis. One theory is that anaerobic syntropy was responsible for the creation of mitochondria and not phagocytosis. This theory would also explain the common ancestors that mitochondria and hydrogenosomes share [1]. According to this theory, scientists believe that the archaeon would have no benefit from becoming phagotrophic, and its bioenergetic situation would deteriorate. It is still unsure how likely this theory is due to discontinuities in prokaryotic evolution, mostly to do gene transfers between bacteria and archaea. Despite the uncertainty surrounding this theory, it suggests that there is a need for further comparing mitochondria with their suspected ancestors.

Mitochondria possess their own genome, separate from the nuclear genome. The Human mitochondrial genome is composed of 37 genes which span around 16.6-kb of circular DNA [3]. These genes are used to produce 13 essential subunits of oxidative phosphorylation and essential components of the mitochondrial translational machinery. It is believed that the size of the mitochondrial genome was substantially reduced due to gene loss or endosymbiotic gene transfer (EGT) [3]. Moreover, the mitochondria require around 13000 nuclear-encoded proteins to produce, assemble, and support the 5 complexes of oxidative phosphorylation. Mitochondria can undergo replication using RNA intermediate throughout the lagging-strand. Previously, scientists erroneously thought that the strand displacement model described the process of mitochondrial replication. There are many similarities between the mitochondrial genome and prokaryotic genomes. Some bacteria support the presence of multiple plasmids that could encode for genes involved in antibiotic resistance. Similarly, human mitochondria contain multiple copies of circular DNA. Another similarity is that mitochondrial and bacterial DNA associates with proteins to created nucleoids, which also supports the endosymbiotic theory [4]. These similarities provide both a basis for linking prokaryotes with mitochondria.

All the putatively conserved genetic and structural similarities between mitochondria and bacteria suggest that there may be additional similarities which have yet to be discovered. Although codon bias can sometimes be species-specific, studies have shown that codon bias is conserved among species in orthologous genes. This study will attempt to classify prokaryotes based on their codon frequencies using multiple machine learning approaches. Specifically, the models will be trained using codon frequency profiles from prokaryotes in the proteobacteria phylum: alphaproteobacteria, betaproteobacteria, gammaproteobacterial, deltaproteobacteria, epsilonproteobacteria, zetaproteobacteria, oligoflexia, acidithiobacillia, and hydrogenophilia. Finally, the models will be used to classify mitochondrial genomes from different eukaryotes to see if their codon biases align with that of alphaproteobacteria, supporting the endosymbiotic theory. Three different algorithms—Artificial Neural Networks, XGBoost, and Naïve Bayes—have been chosen due to their varying complexity and expected difference in performance.

II. Material and Methods

This research consisted of two main steps: calculating the codon counts for genomes and training the machine learning algorithms on that codon data. Finally, the trained classification models were applied to mitochondrial genomes. Overall, three types of models were trained and tested: Artificial Neural Network, XGBoost, and Naive Bayes.

The classification models were trained to assign one of four classes, found within the proteobacterial phylum, to each genome. These classes included α-protobacteria,

β-protobacteria, δ-protobacteria, and ϵ-protobacteria. Other classes found within the proteobacterial phylum were omitted due to data quality issues such as having too little records. Each class sample was reduced to the size of the smallest class. 139 samples were taken from each class to create the training set.

The training dataset was split into a 20% and 80% subsets, where the former was used as a testing set and the latter was used for training the models. The Artificial Neural Network and Naive Bayes models were built using the scikit-learn package in Python, and the XGBoost model was built using the xgboost package in Python. Statistical metrics were then calculated for comparison of all models including F1 score, recall, precision, and accuracy.

The three models were then tested on 37 mitochondrial codon counts which included animal, fungi, plant, and protist mitochondrial genomes.

A. Data

Counts for all 64 codons were calculated from 2,923 reference proteobacterial genomes fetched with NCBI's dataset command-line tool. All codon counts were calculated only by using the coding sequences (CDS) of the genomes. Coordinates for each CDS were retrieved from GFF3 that contained genomic annotation. Each genomic record was then assigned one of four classes: α-protobacteria, β-protobacteria, δ-protobacteria, and ϵ-protobacteria. The same data collection process was followed for the mitochondrial genome dataset.

B. Statistical Metrics

The following statistical metrics were used to compare the performance of the machine learning models.

Accuracy denotes the ratio of total correct predictions (TP and TF) to the sum of all correct and incorrect predictions.

$$Accuracy = \frac{TP+TF}{TP+TF+FP+FN} \text{ (1)}$$

Precision denotes the accuracy of all positive predictions.

$$Precision = \frac{TP}{TP+FP} \text{ (2)}$$

Recall denotes the ratio of TP found to the amount of total positives in the testing set. It is also called the True Positive Rate.

$$Recall = \frac{TP}{TP+FN} \text{ (3)}$$

Although both precision and recall are adequate measures of model performance, it is often desired to have a balance between both metrics. The F1 score allows us to look at both metrics at the same time. The F1 score is the harmonic mean of precision and recall [5].

$$F1score = \frac{2*Precision*Recall}{Precision+Recall} \text{ (4)}$$

The AUC or area under the curve is the calculated area under the receive operating characteristic (ROC) curve. This is a measure of how well the model can distinguish between all of the classes. The ROC curve is created by plotting the True Positive Rate against the False Positive Rate. The AUC value is between 0 and 1. A value of 1 indicates that the model is able to correctly classify all members of a certain class, and exclude those that do not belong to the class. The opposite is true for an AUC of 0. The AUC can be calculated using the trapezoidal rule [6].

C. Naive Bayes

As the name suggests, Naive Bayes is a probabilistic classifier that is based on the Bayes' Theorem. Since the model is simpler than other models, it can be trained and run relatively quickly. This is also most likely why the model often suffers in performance. The Naive Bayes algorithm can be represented by the following equation.

$$P(class|x) = \frac{P(x|class)P(class)}{P(x)}$$

In the equation above, $P(x|class)$ represents the posterior probability of a sample, which determines which class it belongs to. In the numerator, $P(x|class)$ represents the probability that a sample would be in a certain class, while $P(class)$ is the prior probability of a certain class. $P(x)$ is the sample's prior probability [7]. Additionally, the Naive Bayes algorithm makes an important assumption: all properties that contribute to a classification are independent from one another. This may certainly not be the case in this classification problem.

D. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a method which combines several techniques found in other machine learning models. XGBoost is a gradient boosting method that utilizes weak learners or small decision trees to gradually improve the model [8]. Each adjustment is made by optimizing a loss function by using gradient descent.

E. Artificial Neural Network

Artificial Neural Networks consist of layers of interconnected nodes or neurons. Specifically, a multilayer perceptron was used for this project (MLP). MLPs considered the simplest type of neural network that consist of an input, hidden and output layer. Training data is fed into the model through the input layer and, in the hidden layer, the model is able to elucidate complex relationships between the features and determine the class of a sample [9]. Artificial neural networks excel with high-dimensional data and non-linear relationships.

F. Result

First, the models were trained and tested on the proteobacterial dataset to benchmark performance. Below Fig. 1 and Fig. 2 show the classification performance results of the 3 classifiers. The AUC value for each model was calculated as the average of the AUC for all classes. The ROC curves visualize the classification trade-off between the True Positive Rate and the False Positive Rate. No ROC AUC curve was created for the XGBoost model as the library does not support such plots. The average AUC value for the XGBoost value was 0.9926.

Although not shown, XGBoost has the highest AUC value out of the three models which indicates that the XGBoost model is best at differentiating from different classes.

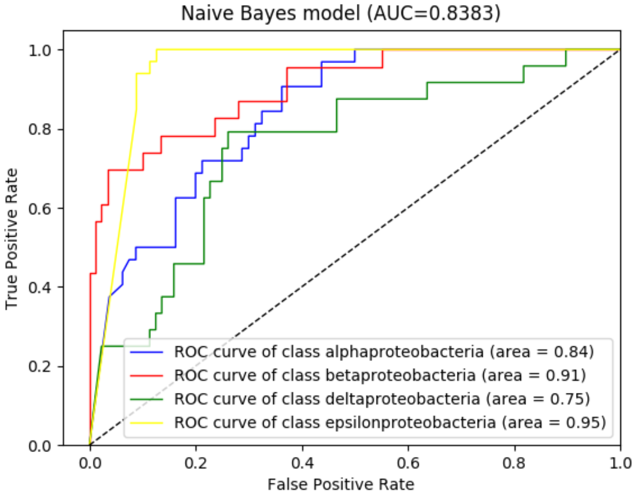

Fig. 1: AUC ROC curves for the Naive Bayes Model tested using proteobacteria genomes

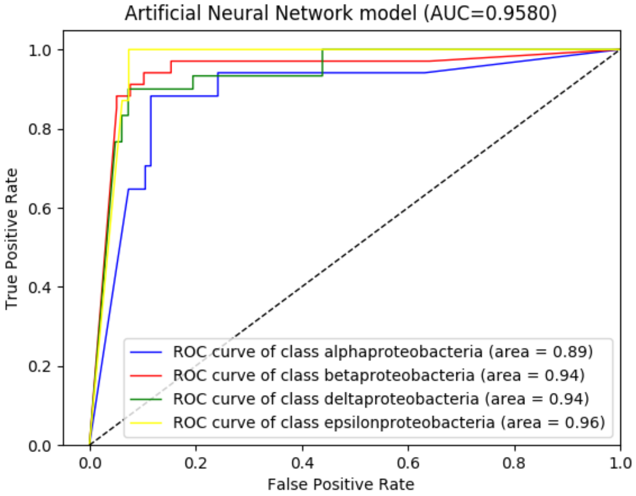

Fig. 2: AUC ROC curves for the Artificial Neural Network Model tested using proteobacteria genomes

Table 1 gives a more detailed overview of the performance of each model using the statistical metrics previously described. Both XGBoost and the Artificial Neural Network have relatively high precision and accuracy scores. Even though the Artificial Neural Network had a lower AUC values than XGBoost, its higher F1-score indicate it provides a better trade-off between Precision and Recall than XGBoost.

After analyzing the performance of the models trained on proteobacterial data, the mitochondrial dataset was both analyzed and then classified by the Artificial Neural Network model. Fig. 3 displays a heatmap visualization of log-transformed codon counts for the mitochondrial genome dataset. The horizontal axis represents all of the codons and their respective log-transformed counts, while the vertical axis represents all of the species used in the dataset. The codon counts were log transformed to be normally distributed, and thus help with interpretation. The mitochondrial genomes were then hierarchically clustered based on their codon counts. The mitochondrial genomes were split into three major clades, as they did not show uniformly distributed codon counts.

Table 2 shows the classifications of the Artificial Neural Network model on the dataset containing 37 mitochondrial codon counts. Of the 37 mitochondrial genomes, 58.30% were classified as α-proteobacteria, 27.78% as δ-proteobacteria, 11.11% as ϵ-proteobacteria, and 2.78% as β-proteobacteria.

After analyzing the Artificial Neural Network model results, the XGBoost model was analyzed for interpretability. The Naive Bayes model was ignored due to poor performance when tested on the proteobacterial data. The Artificial Neural Network was also excluded due to the complexity of the algorithm, making it difficult to interpret the importance of each variable. This left the XGBoost model, which is also quite uninterpretable, but has a metric that is similar to variable coefficients in logistic regression. Table 3 and Table 4 show the top and bottom 7 features ranked by their gain values, respectively. In the XGBoost model, the gain value is the average weight a variable is a assigned throughout all weak learners generated by the algorithm. Gain values provide a general idea of the importance each feature holds in the model, but they are not very reliable because XGBoost relies on other sets of weights to generate predictions. The codon with the highest gain value was TGG, which is the only codon that results in tryptophan.

Model	Precision	Recall	Micro F1-score	Macro F1-score	Accuracy	AUC
Naive Bayes	0.76	0.75	0.75	0.72	0.75	0.8383
XGBoost	0.97	0.96	0.96	0.96	0.96	0.9926
Artificial Neural Network	0.98	0.98	0.98	0.98	0.98	0.9580

TABLE I: Proteobacteria Classification Results.

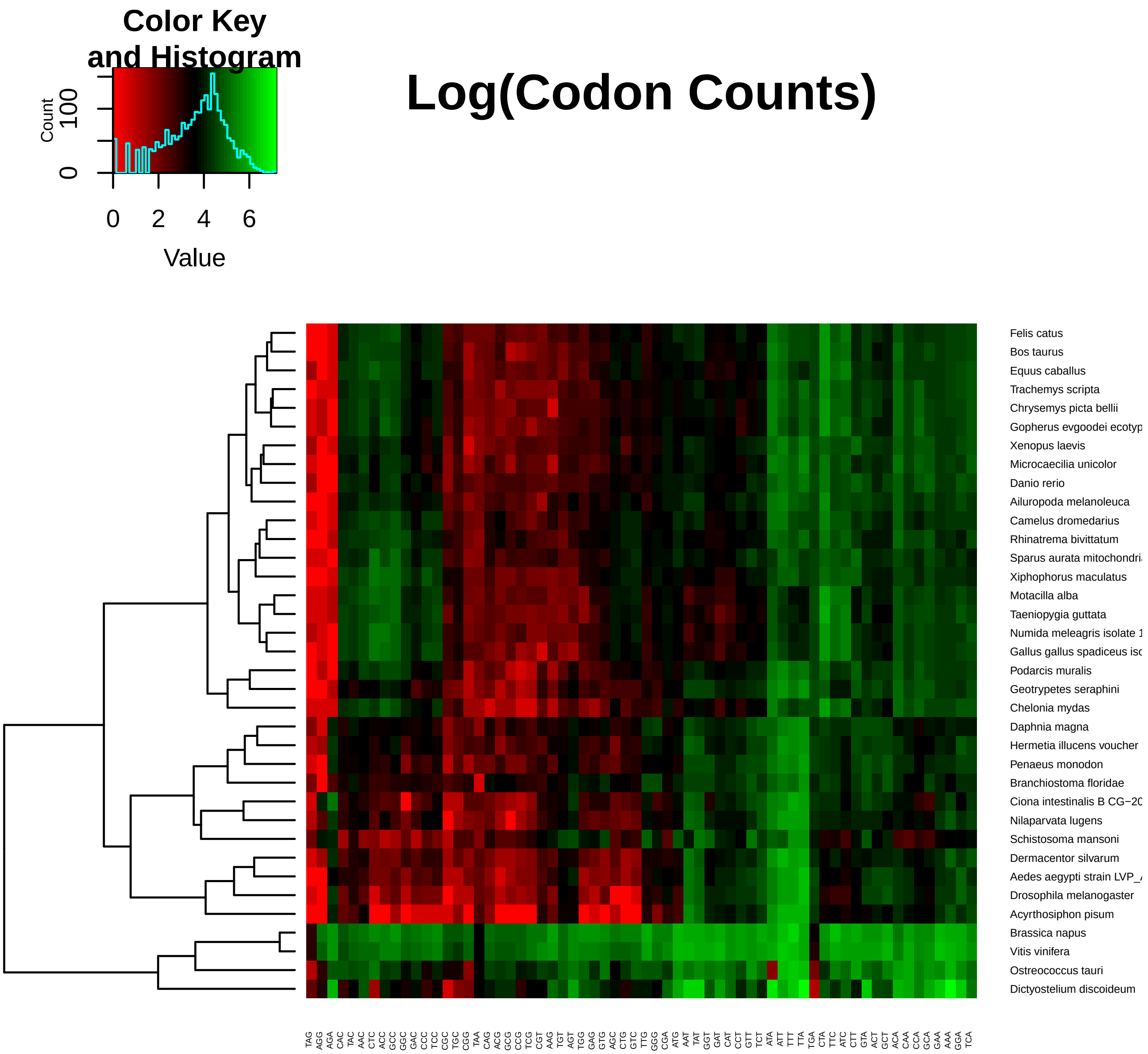

Fig. 3: Heatmap of mitochondrial log-transformed codon counts

Predicted Class	Species
α	Dermacentor silvarum
α	Geotrypetes seraphini
α	Drosophila melanogaster
ϵ	Numida meleagris
α	Schistosoma mansoni
α	Podarcis muralis
ϵ	Gallus gallus
α	Penaeus monodon
α	Trachemys scripta
δ	Vitis vinifera
δ	Ailuropoda melanoleuca
α	Bos taurus
δ	Xenopus laevis
α	Gopherus evgoodei
α	Camelus dromedarius
α	Rhinatrema bivittatum
δ	Sparus aurata
ϵ	Xiphophorus maculatus
α	Aedes aegypti
α	Felis Catus
δ	Daphnia magna
α	Danio rerio
δ	Brassica napus
α	Acyrthosiphon pisum
ϵ	Motacilla alba
δ	Hermetia illucens
β	Taeniopygia guttata
α	Nilaparvata lugens
δ	Ostreococcus tauri
α	Chrysemys picta bellii
α	Chelonia mydas
α	Microcaecilia unicolor
δ	Ciona intestinalis
α	Equus caballus

TABLE II: Classification results of applying the Neural Network Model to the mitochondrial dataset

Codon	Gain	Amino Acid
TGG	13.5120	Tryptophan
AGG	4.5040	Arginine
GCC	3.4293	Alanine
GAA	3.4293	Glutamic Acid
TAC	3.1439	Leucine
CTT	3.0955	Leucine
AGC	3.0420	Serine

TABLE III: Top 7 XGBoost features ordered by gain values

Codon	Gain	Amino Acid
GGT	0.7542	Glycine
GCT	0.8375	Alanine
ATT	0.8394	Isoleucine
ATA	0.8790	Isoleucine
AAC	0.9134	Asparagine
ATC	1.0035	Isoleucine
CTC	1.0377	Leucine

TABLE IV: Bottom 7 XGBoost features ordered by gain values

III. Discussion

Codon counts have potential to become reliable predictors of taxonomic groups and phylogenetic relationships. Further research might look to integrate codon counts in phylogenetic studies, along with DNA sequence alignment, to better understand evolutionary relationships between species. For best performance, Artificial Neural Networks or XGBoost should be used for the codon classification task.

Although reliable classification using codon bias is feasible, it is difficult to understand the reasoning behind the predictions due to the complexity of the models and the vast amount of features in the training dataset. Specifically, it was difficult to understand which codons were most important when classifying proteobacterial genomes. Understanding feature importance may lead to a better understanding of complex mitochondrial evolutionary history. For example, studies have shown that some evolutionary lineages prefer codons that end in T/A while other prefer G/C [10]. If such pattern was found in mitochondrial genomes, it could suggest that mitochondria arose from several different evolutionary lineages or be a way to group mitochondria that have evolved differently.

To address the issue of model interpretability, future work may aim to perform Principal Component Analysis (PCA) to reduce the high dimensionality of the training dataset and better understand the codon bias differences between samples. Using a simpler machine learning model such as random forests or logistic regression could also help with interpretability as variable importance can be calculated easily in those models.

Although the results in Table 2 did support the endosymbiotic theory, there are several issues in this study that prevent the results from being interpreted at face value. The first issue with this research is that it assumes that mitochondria have retained their original genome since they became endosymbionts. In fact, mitochondrial genomes are variable across the many different types of eukaryotes. Mitochondrial protein coding genes can vary from 3 to 67 and tRNA gene content from 0 to 27 [11]. Much of this variation is theorized to be due to gene transfer to the nucleus. The fact that mitochondria have lost parts of their genome to different extents introduces an additional element of randomness in the mitochondrial genome set. This randomness could explain why only 58.30 % of the mitochondrial dataset was predicted to be an α-proteobacteria. Fig 3. supports this notion, as the hierarchical clustering produced several large clades across the mitochondrial dataset. Additionally, patterns in codon bias in each clade are visible from the heatmap alone. For example, the clade including *Brassica nupus* and *Vitis vinifera* show relatively high codon counts for all codons except for 3 codons. These two species starkly contrast with the rest of the species in the dataset that do not have their codon counts so uniformly distributed. Table 2 also showed that no plants were classified as α-proteobacteria which could be due to gene loss. Plant mitochondrial genomes are also much larger than other mitochondrial genomes and frequently undergo recombination activities that promotes evolution [12].

A potential solution for the issue of mitochondrial gene loss could be including nuclear genes which are believed to be of mitochondrial origin within codon count calculation. This solution would require much manual curation and could introduce even more noise due to subjectivity. This approach would only be viable with well-studied organisms such as *Homo sapiens*.

Another issue which could have altered the results of the mitochondrial classification was that the entire α-proteobacterial class was used to train the models. It is believed that a single order, within the α-proteobacteria phylum, called Rickettsiales is a close relative of mitochondria [13]. By including the entire α-proteobacteria phylum the models could have been training on codon counts that were not representative of mitochondrial ancestors. This issue can be simply addressed in the future by excluding any α-proteobacteria that do not belong to Rickettsiale.

Despite all of the issues listed, this study attempted to explain the predictions from the XGBoost model. Table 3 showed the seven most important codons used for classifying proteobacteria by their gain value. The codon TGG had a gain value of 13.5120, which was almost triple the second-highest value. TGG is the only codon used for the amino acid tryptophan. Tryptophan is an aromatic amino acid that contains a benzene ring which is hydrophobic. Tryptophan is sparingly used in protein synthesis since it is very bioenergetically costly to use. Unlike other amino acids, altering the third nucleotide in the TGG codon does not code for tryptophan. Changing the last nucleotide of TGG results in amino acids of very different properteries, both which are aliphatic [14]. This could suggest that tryptophan is used only when it is absolutely needed, and its use is highly conserved in organisms. Therefore, the usage of the tryptophan codon could be one of the most useful signals to use when classifying different types of proteobacteria and mitochondria.

Table 3 and Table 4 also showed that there is not complete agreement between how proteobacteria use synonymous codons. For example, GCC, which results in Alanine, has a gain value of 3.4293 while GCT, which also results in Alanine, has a gain value of 0.8375. This discrepancy would seem to support the notion that more biased genomes show the biggest differences in codon usage, which suggests that variation among species is mainly caused by horizontal gene transfer [10].

This exploratory study shows that there is much that can be learned from codon bias datasets, especially regarding the interpretation of the variability within the data.

IV. Conclusion

The models created in this study indicate that codon usage counts can be used as a means of classifying organisms. Based on the statistical metrics used to benchmark the models, XGBoost and Artificial Neural Networks were the best models to use when classifying proteobacteria. Although the Artificial Neural Network had a lower AUC than XGBoost, it was able to achieve both a higher precision and recall than all of the models. The mitochondrial classification results suggested that a majority of them have α -proteobacteria ancestors although there are still many assumptions this study made that still need to be addressed. Overall, the Artificial Neural Network can be used as a reliable and efficient way to build a model for predicting taxonomic and evolutionary relationships among organisms.

References

[1] Martin, W. F., Garg, S., & Zimorski, V. (2015). Endosymbiotic theories for eukaryote origin. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 370(1678), 20140330. https://doi.org/10.1098/rstb.2014.0330

[2] Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017). The Origin and Diversification of Mitochondria. Current Biology, 27(21). doi:10.1016/j.cub.2017.09.015

[3] Gray, M. W., Burger, G., Derelle, R., Klimeš, V., Leger, M. M., Sarrasin, M., Vlček, Č., Roger, A. J., Eliáš, M., & Lang, B. F. (2020). The draft nuclear genome sequence and predicted mitochondrial proteome of Andalucia godoyi, a protist with the most gene-rich and bacteria-like mitochondrial genome. BMC biology, 18(1), 22. https://doi.org/10.1186/s12915-020-0741-6

[4] Boguszewska, K., Szewczuk, M., Kaźmierczak-Barańska, J., & Karwowski, B. T. (2020). The Similarities between Human Mitochondria and Bacteria in the Context of Structure, Genome, and Base Excision Repair System. Molecules (Basel, Switzerland), 25(12), 2857. https://doi.org/10.3390/molecules25122857

[5] Chicco, D., & Jurman, G. (2020). The advantages of the MATTHEWS correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1). doi:10.1186/s12864-019-6413-7

[6] Sun, L., Wang, J., & Wei, J. (2017). AVC: Selecting discriminative features on basis of AUC by Maximizing Variable complementarity. BMC Bioinformatics, 18(S3). doi:10.1186/s12859-017-1468-4

[7] Zhang Z. (2016). Naïve Bayes classification in R. Annals of translational medicine, 4(12), 241. https://doi.org/10.21037/atm.2016.03.38

[8] Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017). Machine learning–xgboost analysis of language networks to classify patients with epilepsy. Brain Informatics, 4(3), 159-169. doi:10.1007/s40708-017-0065-7

[9] Castro, W., Oblitas, J., Santa-Cruz, R., & Avila-George, H. (2017). Multilayer perceptron architecture optimization using parallel computing techniques. PLOS ONE, 12(12). doi:10.1371/journal.pone.0189369

[10] Khomtchouk, B. B. (2020). Codon usage bias levels predict taxonomic identity and genetic composition. doi:10.1101/2020.10.26.356295

[11] Adams, K. (2003). Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. Molecular Phylogenetics and Evolution, 29(3), 380-395. doi:10.1016/s1055-7903(03)00194-5

[12] Chevigny, N., Schatz-Daas, D., Lotfi, F., & Gualberto, J. M. (2020). DNA Repair and the Stability of the Plant Mitochondrial Genome. International journal of molecular sciences, 21(1), 328. https://doi.org/10.3390/ijms21010328

[13] Gray M. W. (2012). Mitochondrial evolution. Cold Spring Harbor perspectives in biology, 4(9), a011403. https://doi.org/10.1101/cshperspect.a011403

[14] Barik, S. (2020). The uniqueness of tryptophan in biology: Properties, metabolism, interactions and localization in proteins. International Journal of Molecular Sciences, 21(22), 8776. doi:10.3390/ijms21228776