Birds of a Feather… Cluster Together?: Studying How Clusters of Codon Usage Reflect Taxonomic Divides

Abstract

DNA and RNA are made of nucleotide triplets called codons, which are used to make proteins. In biology, "codon usage bias" refers to the idea that some organisms "prefer" one codon over a synonymous one. If it's a known phenomenon that some organisms have characteristic codon-frequency habits, can clustering by codon frequencies inform us about the types of organisms in each cluster? In this project, I fit two types of clustering models on codon-usage data from a wide range of organisms in order to study the extent to which these clusters reflect taxonomic differences. While the resulting clusters are not always well-separated, there is some evidence to suggest an animal's Class may inform its cluster.

Introduction

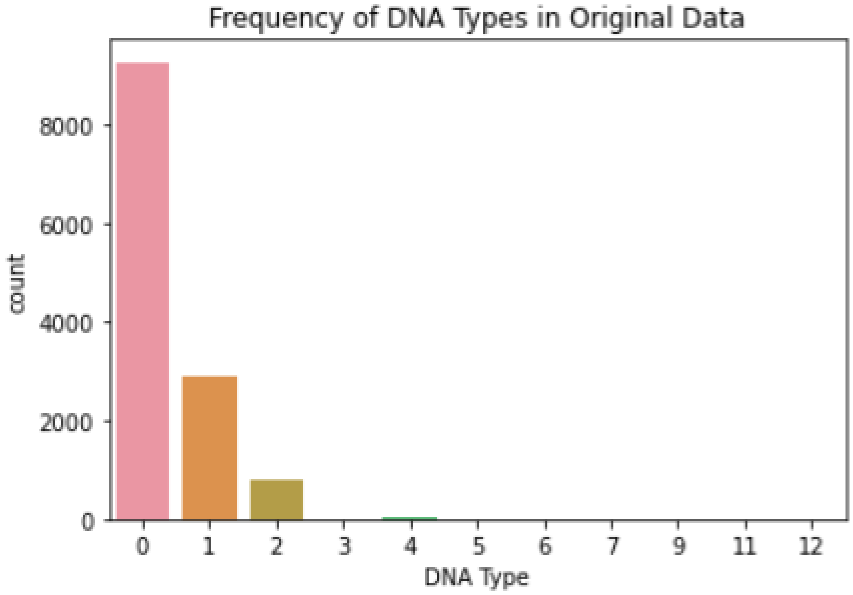
The primary goal of this project was to analyze unsupervised codon-frequency clusters to see how often the models separate organisms by different taxonomic groups (namely, Kingdoms and Classes). Will the clusters split along Kingdom divides, indicating that there is a difference in codon usage across various Kingdoms? Can an animal's Class affect what organisms it gets grouped with? If so, this knowledge could help us classify newly discovered organisms, or highlight ways in which organisms within a single Kingdom or Class are similar to each other. My secondary goal for this project was to use techniques learned in class, specifically Non-negative Matrix Factorization (NMF), KMeans Clustering, and Spectral Clustering. Thus, I performed eight sub-experiments: one for each DNA source (genome or mitochondria), data preprocessing technique ("raw" or NMF), and model type (KMeans or Spectral Clustering). While no sub-experiment had great cluster separation or unequivocally split organisms by Kingdom or Class, the mitochondrial data showed evidence of some Class-informed splitting, as well as model improvement when going from raw to NMF data.

Related Work

Scientists across multiple disciplines have been studying codon usage bias for at least the last 20 years; in fact, the data I used for this project came from one such study. Hallee and Khomtchouk compiled this data for their 2023 paper, *Machine learning classifiers predict key genomic and evolutionary traits across the kingdoms of life*, where they created an ensemble of a many models in order to “classify codon usage in terms of viral, phageal, bacterial, archaeal, and eukaryotic lineage, as well as by cellular compartments from [genomic], mitochondrial, and chloroplast DNA.” My project deviates from Hallee and Khomtchouk’s in that I a) used more specific organism classifications and b) was not concerned about “correct” model outputs but instead was interested in interpreting the results through the lens of taxonomic classifications. Other papers in the field tended to have a narrower dataset– such as the 2013 Behura and Severson study on insect genomes– and the *one* paper I found on Spectral Clustering for codon usage was a study on just the SARS-CoV-2 virus (Mitić 2022).

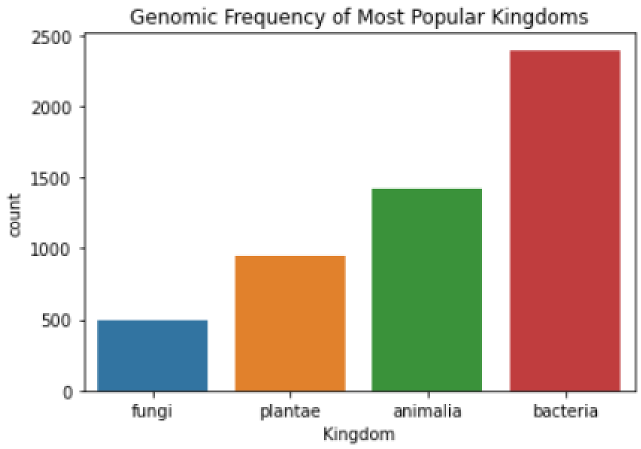
Dataset and Evaluation

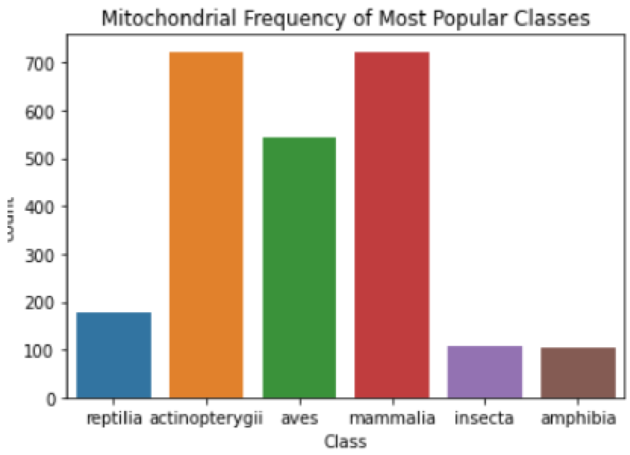
As mentioned previously, I fit my models on codon-frequency data from the Hallee-Khomtchouk study, which *they* obtained from the Codon Usage Tabulated from GenBank (CUTG) database. In addition to the 64 columns for codon frequencies, other useful columns in the Hallee-Khomtchouk dataset are “Kingdom,” “DNAtype,” and “SpeciesName.” Unfortunately, the “Kingdom” column does not



actually contain Kingdoms in the official taxonomic sense, but it did allow me to quickly filter out observations from viruses, bacteriophage, and plasmids, which are technically not alive. There were 13 possible values in “DNAtype,” the most common being 0 for genomic DNA and 1 for mitochondrial; I chose to filter out the rest of the DNA types as I worried they had too few observations to fit a reliable model. When it came time to fit the models, I split the data by DNA type.

Once the data had been filtered and I was left with 9,095 observations, I began adding the taxonomic information, from Domain down to Genus. These values came from passing a (cleaned) value from “SpeciesName” into EcoNameTranslator’s classify() function. EcoNameTranslator is a Python package for working with common and scientific names of organisms. When an organism was not found in EcoNameTranslator’s database, the taxonomic columns were left blank. Once I began fitting the models, I quickly realized that Domain was a bit too broad for either DNA type, but Kingdom seemed reasonable for the genomic data and Class had promise for the mitochondrial data. Thus, for this paper, all genomic results interpretations are with respect to Kingdoms, and the mitochondrial interpretations to Class. (Since there are many Kingdoms and Classes present in the dataset, I focused my analysis on only the most common Kingdoms/ Classes so as to not get too overwhelmed with the results).



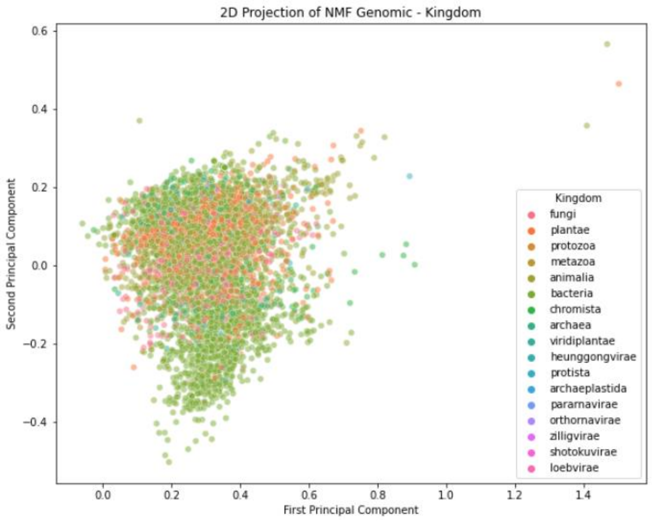


I relied on a couple performance evaluation methods for this project. To judge how reliable the clusters were, I used the average silhouette score; this also served as a tiebreaker when determining the ideal number of clusters for the Spectral Clustering models. Having an objective statistic like silhouette score (which conveys how similar points are within a group and how dissimilar they are to other groups) is useful when preparing to analyze the clusters for evidence of splitting along taxonomic lines. If the model cleanly separates *plantae* from *bacteria*, but has a low silhouette score, then we are probably less inclined to trust the idea that codon clustering can distinguish these two Kingdoms. Speaking of, how *do* we determine if the clusters split along Kingdom or Class divides? A naive approach could be to look for clusters that are mostly made of one taxonomic group, but this discounts Kingdoms and Classes that are often too small to make up a cluster on their own (especially for models with few clusters). Instead, I chose to study the cluster breakdown for each taxonomic classification. If a cluster manages to capture most of a particular Kingdom/ Class, and no other cluster has a sizable proportion of said Kingdom/ Class, then this suggests that the model may be (at least somewhat) picking up on these taxonomic differences.

Methods

All code was done in Python, particularly relying on Pandas for data handling, Sklearn for the data preprocessing and models, and Pyplot and Seaborn for data visualization. I performed eight sub-experiments, one for each combination of DNA type (genomic or mitochondrial), data preprocessing (raw or NMF), and model (KMeans or Spectral Clustering). As mentioned earlier, *only* the codon data was used for the following steps; the taxonomic information only came into play during the analysis! I will first explain the dimension reduction technique used, and then discuss the building of the models (since I had to preprocess the data before I could fit the models on it).

I chose to perform Non-negative Matrix Factorization (NMF) over other methods– such as PCA– because frequencies are non-negative. I used the NMF function from



sklearn.decomposition with random initialization; both the genomic data and mitochondrial data converged at or before 5000 iterations (at the time of the presentation, I was only letting the function run 1000 iterations). When a sub-experiment's preprocessing type is "raw," that just means no NMF was performed.

For the slearn.cluster's KMeans model, the number of clusters (up to 20) for each experiment was chosen via the elbow method with the model's inertia. Once the optimal K was picked, I refitted the model with said K, saved the assigned labels for later study, and computed the silhouette score.

I went through a similar process for the Spectral Clustering model: run tests to determine the number of clusters, refit the final model, save the labels, and calculate the silhouette score. This time, however, I had to use a different method for determining the best number of clusters. Since Spectral Clustering doesn't have an inertia statistic, I used the eigengap statistic with the silhouette score as a tiebreaker. To perform eigengap, I used a 10-nearest-neighbor adjacency matrix and only considered the 20 smallest eigenvalues (to match the options given to KMeans). When there were several large jumps, I chose the one corresponding to the highest silhouette score.

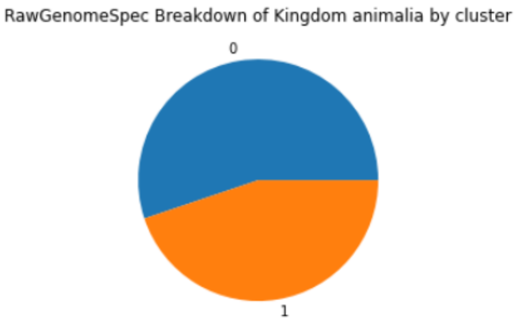
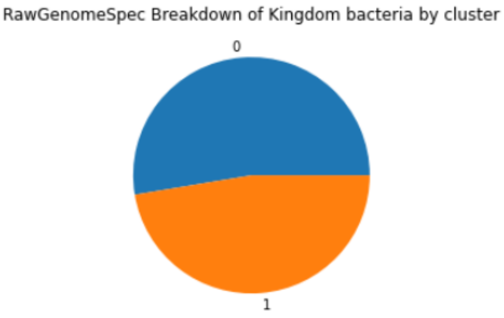
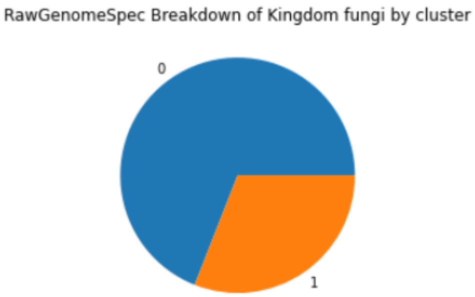
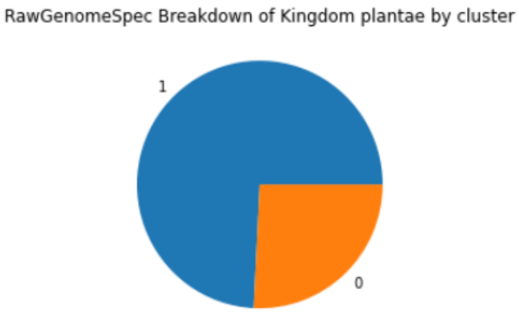
Experiments

Once all the models had been fitted and the labels collected, I was able to analyze the results. I predominantly relied on pie charts to convey how much of each Kingdom or Class went into each cluster, though sometimes it was useful to look at the clusters as a whole instead. To avoid droning-on about eight relatively similar experiments, I will discuss results for the models with highest silhouette score on the genomic data and the mitochondrial data and then conclude on general findings across all experiments.

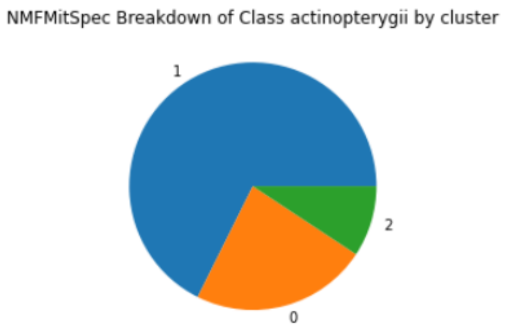
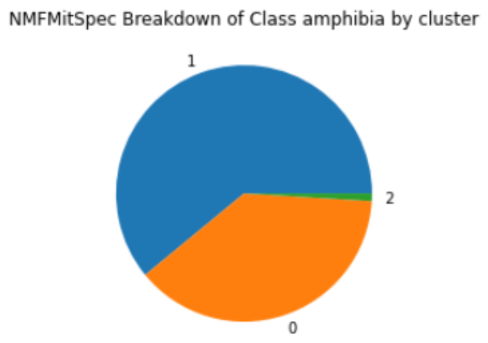
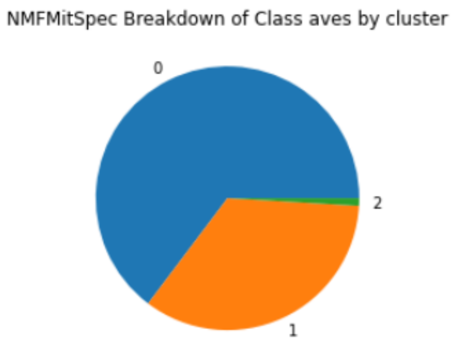
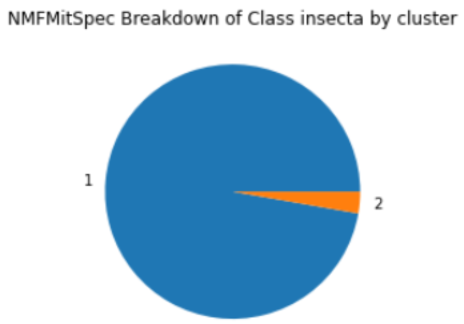
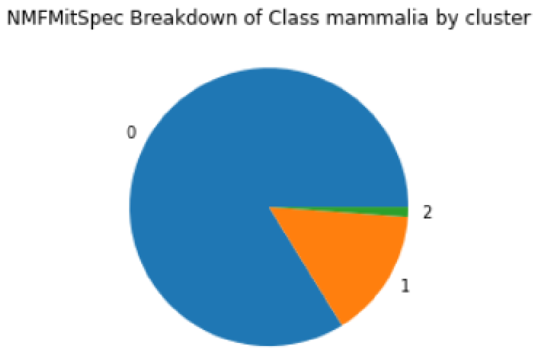
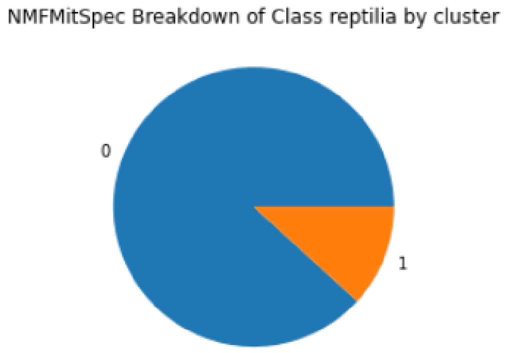
Silhouette Score for Each Model and DNA Type				
	Raw KMeans	NMF KMeans	Raw Spectral	NMF Spectral
Genomic	0.1774	0.1464	**0.2810**	NA

Mitochondrial	0.1925	0.2800	0.2056	**0.3182**

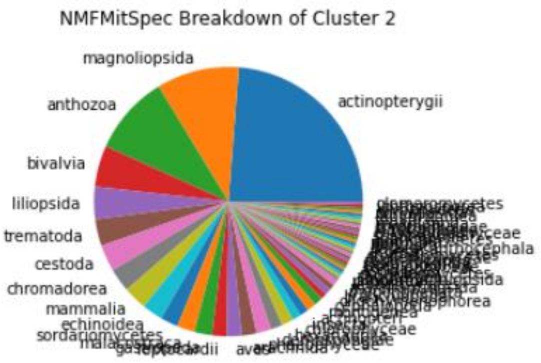
For the genomic data, the Spectral Clustering on raw data had the highest silhouette score, with a score of 0.28 and only two clusters; 0.28 indicates that this is not a great clustering but could still be informative. The clusters may very well be “informative” in general, but they are sadly not very revealing when it comes to Kingdom divisions. As we can see from the Kingdom/cluster pie charts, all four of the major Kingdoms have a decent (25%+) proportion of organisms in both clusters. It is interesting, however, that the plants are the only popular Kingdom to “prefer” cluster 1 (but animals and bacteria only slightly prefer 0 over 1). While these cluster assignments aren’t completely random (that would require 50% to be in cluster 0 and 50% in cluster 1), it is difficult to argue that the clusters are splitting along Kingdom divides when every major Kingdom got assigned to both clusters with considerable frequency.

Select Kingdoms Breakdowns by Cluster – Raw Spectral Clustering Model	
	
	

The best model for the mitochondrial data performed slightly better; the Spectral Clustering on the NMF data achieved a silhouette score of 0.32 in three clusters. Again, this score is not ideal, but it does suggest that the clusters capture some information. This time, we do see some instances of Classes that have the majority of their organisms in a certain cluster, with less than 25% in any other cluster. The mammal and reptile Classes exhibit this behavior to a certain degree, but the insect Class truly stands out, with 0 instances in cluster 0, 106 in cluster 1, and 3 in cluster 2!

Select Classes Breakdown Cluster – NMF Spectral Clustering Model	
	
	
	

It is worth noticing that these Classes were rarely put in cluster 2, if at all. When examining cluster 2’s breakdown by Class, we see that it is mostly a catch-all of the less popular Classes. Is this an indication that there is a split in codon usage between the common and uncommon Classes, or is this merely a side effect of having fewer (<100 each) examples of these Classes?



Before I get into general observations, I want to point out that the Spectral Clustering model for NMF genomic data doesn’t have a silhouette score at all because the eigengap method recommended only one cluster, so there is no “other” cluster with which to calculate the inter-cluster distance. Without any cluster splits, I also can’t analyze how the different clusters relate to different Kingdoms.

Both models on the raw mitochondrial data had multiple Classes with a dominant cluster, but unfortunately these models had the worst silhouette scores of the mitochondrial models, so I am less confident about these results. Additionally, the insects always had a clear favorite cluster, regardless of the model; this suggests that clustering *can* capture Class difference, but some Class divisions are more obvious than others.

Percentage of Insects Assigned to the Most Popular Cluster for Insects			
Raw KMeans	NMF KMeans	Raw Spectral	NMF Spectral
88.99%	93.58%	96.33%	97.25%

Even more broadly, it appears that the mitochondrial clusters were, in general, more informative about Class differences than the genomic clusters were about Kingdom differences. Additionally, performing NMF on the genomic data improved the silhouette score for both types of clustering models but the same cannot be said of the genomic data. The models for the mitochondrial data curiously picked the same number of clusters, regardless of the data preprocessing procedure (or lack thereof), this could be an indication that the different models are picking up on the same intrinsic information (a good sign for the validity of these models). Again, this is not the case for the genomic data, where each model chose a different number of clusters. Speaking of the number of clusters, for every model, the silhouette score started at around 0.3 ~ 0.35 and almost exclusively decreased as the number of clusters increased. We would generally expect the score to increase and then decrease after a certain point, but the fact that the silhouette score never increases above the starting point (2 clusters) suggests that KMeans and Spectral Clustering may not be the best models for this data.

Number of Clusters Chosen for Each Model and DNA Type				
	Raw KMeans	NMF KMeans	Raw Spectral	NMF Spectral
Genomic	4	7	2	1
Mitochondrial	5	5	3	3

Discussion

All-in-all, while we can find some evidence of clusters splitting along *some* taxonomic divides, no model or data preprocessing method was able to do this for all of the popular Kingdoms/ Classes. This is not inherently a loss, though. Clusters that do not neatly separate taxonomic groups but also don’t appear random could be used to study ways in which organisms in different Kingdoms and Classes are similar!

That being said, no model achieved a high silhouette score, meaning that the clusters were not always cleanly dividing up the data– maybe a different clustering method or different initialization seed could result in completely different clusters, and then how confident could we be that taxonomic differences really do inform cluster divisions? This may be a sign that these models are not great for the codon-usage data, or that the data itself is ill-suited for simple clustering methods. On the other hand, a good silhouette score might be unachievable because everything seems far apart in 64 dimensions… An additional concern is the fact that most of the eigen plots that helped me confirm the proposed eigengap were often gradual and ambiguous, making me less confident about the number of clusters I used.

Conclusion

To summarize, my goal for this project was to attempt different clustering models (and data preprocessing steps) in order to study the extent to which the resulting codon-usage clusters reflected different taxonomic levels. I ended up performing eight sub-experiments, one for each DNA type (genomic or mitochondrial), preprocessing method (raw or NMF), and model type (KMeans or Spectral Clustering). In general, the mitochondrial models were a bit more reliable than the genomic models, and there were some instances of Classes greatly favoring one cluster over the others. We may take this as evidence that that Class– if not Kingdom– plays a role in determining how the codon-usage clusters are constructed (for mitochondrial DNA, specifically). If time had allowed, it would have been interesting to examine the specific organisms to see if some observations were always clustered together, regardless of model. Additionally, the lackluster silhouette scores make me want to explore different clustering models, or even just different hyperparameters for the Spectral Clustering model. Finally, I think the interpretation of the clusters could be aided and enriched by the addition of non-taxonomic data such as biome or diet, which may work in conjunction with taxonomic differences to inform the clusters.

References

Behura, Susanta K, and David W Severson. “Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes.” Biological reviews of the Cambridge Philosophical Society vol. 88,1 (2013): 49-61. doi:10.1111/j.1469-185X.2012.00242.x

Hallee, Logan, and Bohdan B. Khomtchouk. “Machine Learning Classifiers Predict Key Genomic and Evolutionary Traits across the Kingdoms of Life.” Nature News, Nature Publishing Group, 6 Feb. 2023, www.nature.com/articles/s41598-023-28965-7.

Mitić, N., et al. "Large scale clustering in structural and evolutionary analysis of SARS-CoV-2 proteins." Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022). 2022

Parvathy, Sujatha Thankeswaran et al. “Codon usage bias.” Molecular biology reports vol. 49,1 (2022): 539-565. doi:10.1007/s11033-021-06749-4