Case Study 2 - Classifying DNA Barcodes (Runtime: ~60 min)

Garrett Allen, Isabella Swigart, Ayaan Patel, Matthew Cui

1 Introduction

Biodiversity research relies heavily on DNA barcoding, which is a technique used to classify and identify species using DNA sequences. DNA barcoding involves several steps, including the collection of biological material, DNA sequencing, alignment of sequences, and the assignment of a taxonomic name (Herbert et al, 2003). In recent years, DNA barcoding has become an essential tool in biodiversity research due to its speed, accuracy, and efficiency (Winterton, Wiegmann, and Schlinger 2007). In this case study, we are presented with a set of 7,000 aligned DNA sequences obtained from butterfly specimens captured in a Finnish forest. Our goal is to classify the sequences into their respective families and genera, using a historical dataset of 40,000 annotated DNA sequences for which annotations have been confidently established. The aim is to build a classification model using the historical dataset to annotate the 7000 sequences at the family and genus levels and to introduce a measure of uncertainty in our predictions. This study showcases the power of DNA barcoding in classifying species and highlights the importance of using machine learning techniques to improve the accuracy of these classifications.

To achieve this goal, we will explore various classification methods, such as using k-mer sequences with LASSO regression and random forests. We are investigating the importance of the entire DNA sequence for classification and identify the loci that are particularly relevant to classification. The results of this project have practical implications for biodiversity research, as accurate taxonomic classification is essential for understanding and managing biodiversity.

2 Methods

2.1 Data Wrangling

The data library contained 40,000 instances of four variables that were used in this study. It contained information on the families, genera, species associated with a specific DNA sequence. Our first step to tidy this data was to assign each observation a unique ID based on the initial ordering of the data for easier identification and future data manipulation. We then split the 'DNA' sequence on the individual nucleotide level and replace characters that are not G, T, A, C, U, or a dash ("-") with a dash to remove other non-standard characters that might be in the dataset in preparation for counting k-mer sequences.

2.2 Exploratory Data Analysis

Figure 1 reviews the distribution of each genus' frequency in the Lepidoptera library. The overall number of observations for each family are reported above each boxplot. For more interpretable y-axis scaling, this visualization filtered out any genus that occurred more than 250 times in the library. There were relatively few outliers of this sort, only Eupithecia (Geometridae family) with 432 instances, Scoparia (Crambidae) with 362 instances, Euxoa (Noctuidae family) with 289 instances, and Catocala (Erebidae) with 256 instances. We see that for most families, the upper quartile of the most frequently observed genera still contain less than 25 instances in the Lepidoptera library. All families have dozens of genera that only appear once in the library, which could pose a challenge when classifying genus.

Table 1: Most Frequent Genus By Family

Family	Genus	Genus Count	Proportion of Family
Sphingidae	Xylophanes	223	13.35%
Crambidae	Scoparia	362	12.62%
Saturniidae	Automeris	200	12.36%
Gelechiidae	Chionodes	233	10.85%
Oecophoridae	Philobota	174	10.3%
Tortricidae	Acleris	230	6.86%
Geometridae	Eupithecia	432	5.45%
Hesperiidae	Astraptes	74	5.2%
Noctuidae	Euxoa	289	4.26%
Erebidae	Catocala	256	4.18%
Pyralidae	Ephestiodes	57	3.93%
Nymphalidae	Erebia	105	3.58%

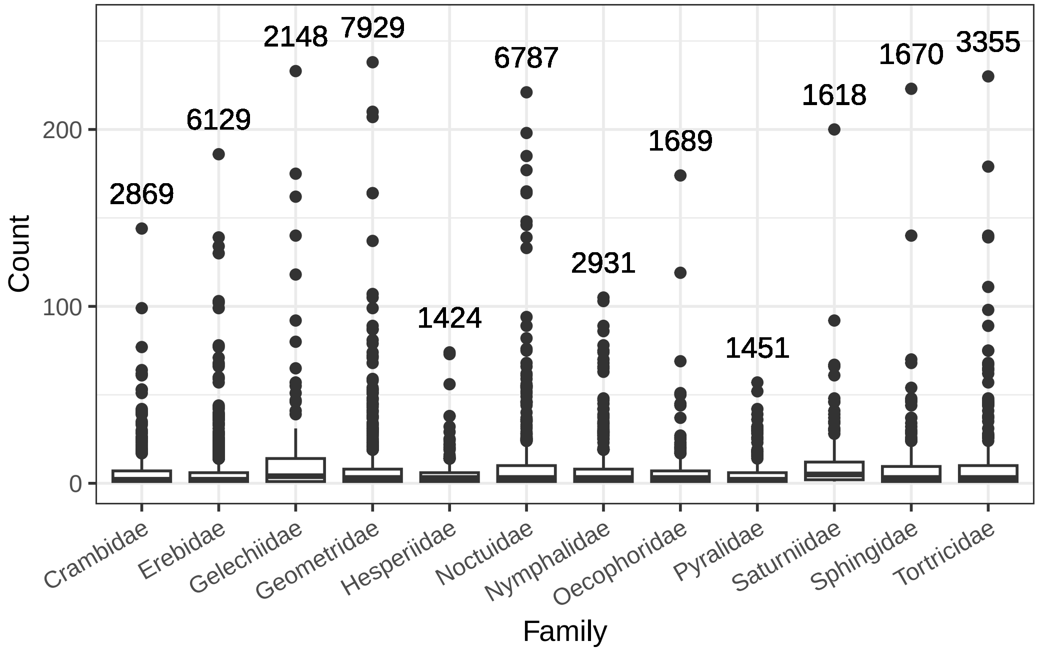


Figure 1: Number of Observations of Genera Within Families

Table 1 displays the most frequent genus of each family, as well as the proportion of said genus within that family. We see that the number of observations in the Lepidoptera library belonging to each family is quite balanced, and, amongst each family, there is not one genus that dominates the makeup in the dataset. We also see in Figure 1 that every family has a comparable number of instances in the dataset. Therefore, we can conclude that there isn't a class imbalance, so no further procedures of under or over-sampling is needed to ensure the integrity of later models.

2.3 Modeling

2.3.1 Family Model

One of the goals of our analysis is to create a model to predict family of a DNA strand based on the loci. To fit the family model, the Lepidoptera library was randomly sampled to be split into a training and test set that contained 70% and 30% of the observations, respectively. We then tried both a k-mer penalized regression model with k = 3,4,5, as well as a model where we let the loci themselves be covariates in a penalized LASSO regression (Loci model). The Loci model had the highest average accuracy predicting families (97.5%), so we will be proceeding with this model in our analysis of sequences.

Table 2: In-Sample Loci Model Output

Class	Sensitivity	Specificity	Precision	Recall
Crambidae	0.9561091	0.9976182	0.9687500	0.9561091
Erebidae	0.9552823	0.9897693	0.9436775	0.9552823
Gelechiidae	0.9683544	0.9974836	0.9562500	0.9683544
Geometridae	0.9884220	0.9973480	0.9892704	0.9884220
Hesperiidae	0.9975610	0.9998238	0.9951338	0.9975610
Noctuidae	0.9593456	0.9914802	0.9588702	0.9593456
Nymphalidae	0.9929825	0.9996332	0.9953107	0.9929825
Oecophoridae	0.9593496	0.9984024	0.9632653	0.9593496
Pyralidae	0.9736842	0.9999118	0.9975490	0.9736842
Saturniidae	0.9874477	0.9992022	0.9812890	0.9874477
Sphingidae	0.9939516	0.9999112	0.9979757	0.9939516
Tortricidae	0.9929789	0.9999071	0.9989909	0.9929789

2.3.2 Modeling Loci Locations

In addition to creating a model that accurately identifies families and genera, an important question of interest was specifically which loci are important for classification of Lepidoptera. We used our Loci model described above for this task, since the importance of various loci in predicting family would allow us to analyze which locations are most important for classification. We used 5-fold cross validation to fit the model and find the optimal lambda parameters. We ultimately used this model to predict the family of the 7000 unlabeled specimens.

As seen in Table 2, the precision, recall, specificity, and sensitivity are incredibly high, indicating that the loci locations are very indicative of what family a species resides in. This suggests that the coefficients of our model for each loci will be indicative of its contribution to classification, as it has very high accuracy on the in-sample test set (97%).

We can analyze the coefficients from this model to determine which loci locations are particularly important to family classification. Specifically, LASSO sets unimportant coefficients to zero, so any coefficients that are set to zero for all families provide no benefit to classification. Those that have high coefficients across the families (in absolute value) are generally important to classification. Thus, we propose the following metric to evaluate a loci's importance:

$$\mathrm{C}(I) = \sum_j |I_j|$$

where I indicates the loci location and j indicates which family the coefficient is for, and I_j is the value of the coefficient for the Ith loci in the jth family. We will call this the C-Value of a loci.

Table 3: Top 10 Loci According to C-Value

	A	B	C	D	E	F	G	H	I	J
Loci	422.00	520.00	329.00	415.00	582.00	521.00	522.00	580.00	327.0	303.00
C-Value	7.33	6.88	6.82	6.49	5.73	5.27	5.25	5.01	4.4	4.32

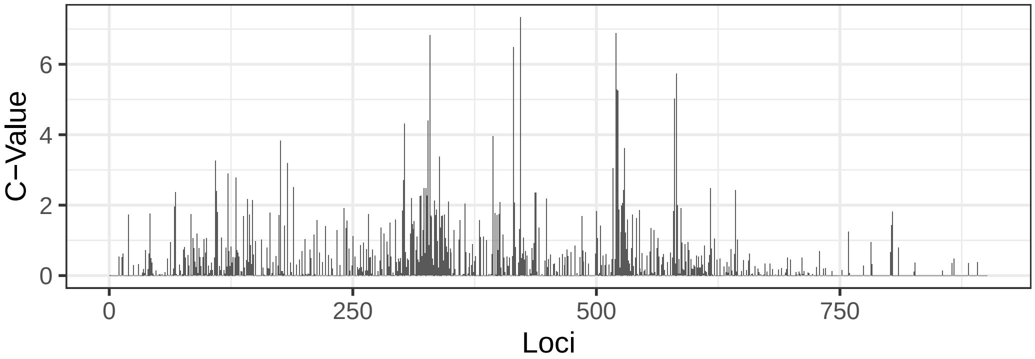


Figure 2: Importance of Loci by C-Value

In Table 3, we list the top 10 loci according to their C-Value. This shows that the top loci are largely present in the middle between 300-600, and that they are most distinctive when determining the family.

From Figure 2, we can see the distribution of C-Value across the DNA strand — this also supports our hypothesis that the middle values are the most important for determining family value, as they have the highest sum of coefficients across the families. The last 300 or so values are least distinctive, and many of the locations at the beginning and between 200-300 are not very important.

In fact, as shown by the table above, there are more than 500 loci with a C-Value of less than .25, and 376 with a value of 0; that is, the coefficient for that loci was set to zero for all values of family by LASSO.

Thus, from this analysis, we can conclude that the whole sequence is not important for classification. In fact, 376 loci are completely irrelevant for classification in our LASSO model, as their coefficients were entirely set to zero. The important loci are those largely in the middle, between 300-600, as justified by both the top 10 most important loci and our graph above.

2.3.3 Genera Modelling

To predict the in-family genera that the unlabeled sequences belonged to, we fit twelve 5000 tree random forest models, one for each family. Each random forest predicted the genus within a specific family, and used loci locations as covariates. Models were fit using a 70-30 train test split, and accuracy was assessed by performance on tsting data for a specific family. We decided to create separate models for genus prediction for each family primarily because our family predictions were shown to have very high accuracy and because it reduced the number of classes each random forest model could predict, thus making it easier for the model to identify patterns in the data. One weakness of this approach is that errors in family necessarily result in errors in genus prediction, as our genus models cannot predict a new genus not present in the family. Future work could try to address this sort of uncertainty propagation by adding a way to easily correct incorrect/low confidence family labels.

Table 4: Genus Prediction In-Sample Accuracy by Family

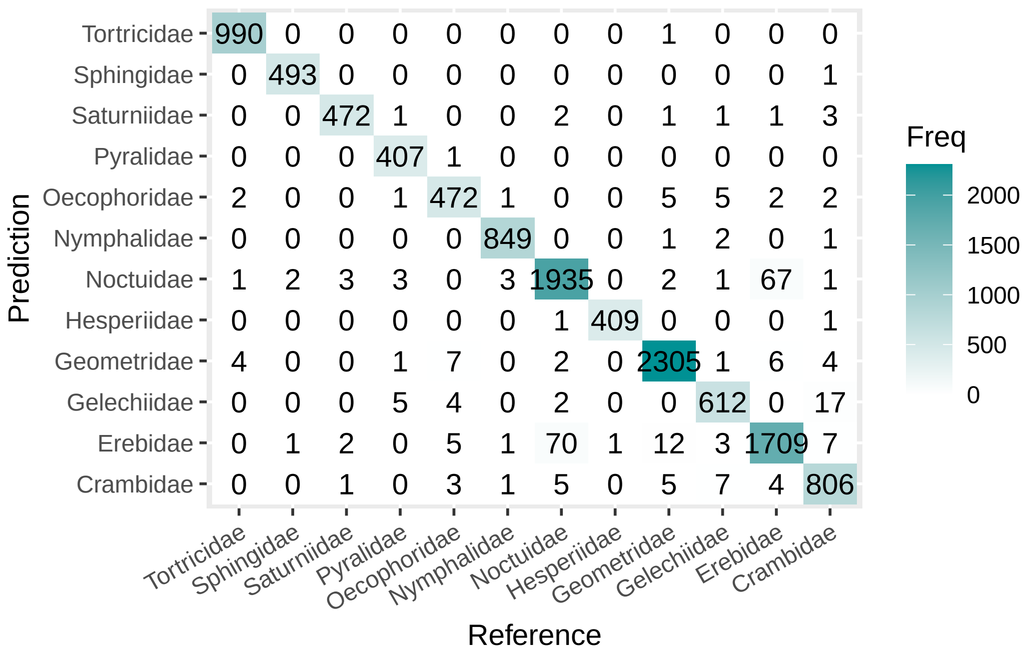
Family	Accuracy
Geometridae	0.8221942
Gelechiidae	0.8944099
Sphingidae	0.9101796
Tortricidae	0.9076465
Crambidae	0.8118467
Erebidae	0.8346928
Oecophoridae	0.7928994
Hesperiidae	0.8430913
Pyralidae	0.7839080
Nymphalidae	0.8930603
Saturniidae	0.9257732

We have an average in-sample accuracy of 0.856 across all 12 families. According to Table 4, our most accurate genus predictions are within the Saturniidae family (0.926), and our least accurate predictions are within the Pyralidae family (0.784). Saturniidae may have such success because it contains the fewest number of single-occurring genera among all families (26 single occurrences compared to an average of ~122 among the other 11 families).

3 Results

3.1 In-Sample Accuracy

In Figure 3, we report the confusion matrix for family predictions with our final model. While our results were strong overall with 97% accuracy, the model did have a tendency to label Noctuidae as Erebidae and occasionally Geometridae.



Prediction \ Reference	Tortricidae	Sphingidae	Saturniidae	Pyralidae	Oecophoridae	Nymphalidae	Noctuidae	Hesperiidae	Geometridae	Gelechiidae	Erebidae	Crambidae
Tortricidae	990	0	0	0	0	0	0	0	1	0	0	0
Sphingidae	0	493	0	0	0	0	0	0	0	0	0	1
Saturniidae	0	0	472	1	0	0	2	0	1	1	1	3
Pyralidae	0	0	0	407	1	0	0	0	0	0	0	0
Oecophoridae	2	0	0	1	472	1	0	0	5	5	2	2
Nymphalidae	0	0	0	0	0	849	0	0	1	2	0	1
Noctuidae	1	2	3	3	0	3	1935	0	2	1	67	1
Hesperiidae	0	0	0	0	0	0	1	409	0	0	0	1
Geometridae	4	0	0	1	7	0	2	0	2305	1	6	4
Gelechiidae	0	0	0	5	4	0	2	0	0	612	0	17
Erebidae	0	1	2	0	5	1	70	1	12	3	1709	7
Crambidae	0	0	1	0	3	1	5	0	5	7	4	806

Figure 3: Confusion Matrix for In-Sample Family Prediction With Loci Model

3.2 Unlabeled Sequence Composition

Table 5: Comparison of Relative Family Composition of Labeled and Unlabeled Specimens

Family	Unlabeled Specimens	Labeled Specimens
Crambidae	0.071725	0.0690000
Erebidae	0.153225	0.1572857
Gelechiidae	0.053700	0.0488571
Geometridae	0.198225	0.1792857
Hesperiidae	0.035600	0.0302857
Noctuidae	0.169675	0.1735714
Nymphalidae	0.073275	0.1055714
Oecophoridae	0.042225	0.0434286
Pyralidae	0.036275	0.0294286
Saturniidae	0.040450	0.0437143
Sphingidae	0.041750	0.0434286
Tortricidae	0.083875	0.0761429

Our final predictions for the 7000 unlabeled sequences are made using the loci model to predict family and using the respective random forest-fitted family model to predict genus. Our predictive probabilities for family range from 0.223 to 0.999, with a median of 0.998, while our predictive probabilities for genus range from 0.016 to 1.000, with a median of 0.681. Additionally, as shown in Table 5, the relative proportions of families within the labeled and unlabeled specimens are quite similar.

4 Discussion

In conclusion, this study provided concrete insights into classification of DNA sequences and drew conclusions on the importance of the location of loci in DNA modeling. Through LASSO penalization of unimportant loci, we realized that the entirety of the DNA sequence is not needed for prediction, and that the most important loci positions for these classification problems were largely in the middle of the sequence. A significant implication of this finding for the larger field of computational genomics and biology is that future studies are able to reduce required computing power to process a same amount of data, or that similar investigations have greater scalability. Both the loci and genera models using a LASSO and random forest approach produced high in-sample accuracies.

One major weakness of this study was that it did not account for the presence of a novel family or genus in the test dataset. We cannot conclude that our model is able to identify new instances of Lepidoptera, since our model can only currently predict categories present in the training dataset. Future work should explore ad-hoc ways to handle new families and genera, likely by outputting “New” or “Unknown” when the probability of any existing class is too low. Future work can also investigate using loci importance to inform a better classifier.

Lastly, for our prediction of the unlabeled sequences, the average probability for family prediction is 95.1%, whereas genus prediction only sits at 58.3%. Moreover, the median family prediction probability is a quite impressive 99.8%, contrasting with the 68.2% value for genera. In order to enhance the accuracy of our predictions, we can address the issue of imbalanced data by collecting additional samples for genera with limited data points. By doing so, we can balance our dataset and avoid any biases towards predicting genera that have more data points. Another interesting observation we noticed from the predicted probabilities is that there is no correlation between prediction probabilities of families and its respective genera, meaning that better prediction at the family level does not necessarily translate to the genus level for a given family.

References

Hebert, P. D., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings. Biological sciences*, *270*(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Winterton, Shaun L., Brian M. Wiegmann, and Evert I. Schlinger. 2007. “Phylogeny and Bayesian Divergence Time Estimations of Small-Headed Flies (Diptera: Acroceridae) Using Multiple Molecular Markers.” *Molecular Phylogenetics and Evolution* 43 (3): 808–32. https://doi.org/10.1016/j.ympev.2006.08.015.