Classifying Lepidoptera Family and Genus by DNA Barcodes

Yihan Shi, Olivia Fan, Jeremiah Hodges, Chris Yang

1 Introduction

1.1 Background Information

A great challenge in the field of modern biology is developing reliable and accurate technologies for the screening of DNA sequences. Over the past decade, a variety of DNA-based approaches have surfaced and been applied to multiple fields of investigation, from epidemiology to taxonomy (Pereira, Carneiro, and Amorim 2008). In particular, rapid species identification based on DNA sequencing has been fruitful in terms of identifying existing organisms in an ecosystem, diagnosis of pathogens, and discovery of new species. Genome information from short DNA barcode sequences, usually 400-800 base pairs, is often sufficient to identify almost all mixed samples on earth (Kress and Erickson 2008). As comprehensive and powerful as it seems, DNA barcode classification faces the difficulties of identifying appropriate loci and addressing missing nucleotides,

Our goal in this study is to develop strategies to address the above challenges and build effective prediction models based on DNA information. Based on a historical dataset of 40000 annotated DNA sequences with established annotations, we built several predictive classification models to annotate the sequence at the Family and Genus level. The main goals of this analysis include the following: predicting family and genus of Lepidoptera based on DNA barcode, determining the role of a whole DNA sequence in classification, and identifying important loci along sequences for classification.

1.2 Data Description & Pre-Processing

The original training data consists of 40000 annotated DNA sequences, where each observation represents an organism of known taxonomy. The test data consists of 7000 aligned DNA sequences, where each observation represents a single Finnish butterfly belonging to the order Lepidoptera.

K-mers are defined to be short recurring elements in the genomes of all living species. These DNA sequence's subsequences of length k, prove to be an effective approach to preserve sequence information (Matyas Cserhati 2019). Since these elements are conserved and diverged across species owing to their functional significance, kmers are ideal signatures for insect species identification (Matyas Cserhati 2019). Therefore, we split each DNA sequence from the training data into k-mers, where k is a fixed length between 1 and 6. After obtaining substrings of length k, we gathered the total count of each substring. In order to maximize the information obtained underlying the data generating mechanism, we combined the different k-mers and their respective count to obtain the covariates used in the model.

1.3 Exploratory Data Analysis

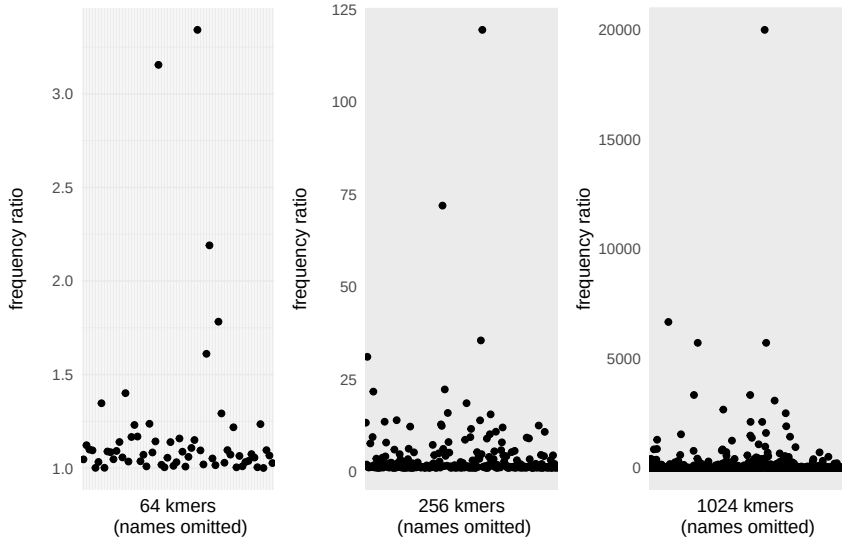

Figure 1: Ratio of frequencies for the most common value over the second most common value of kmer

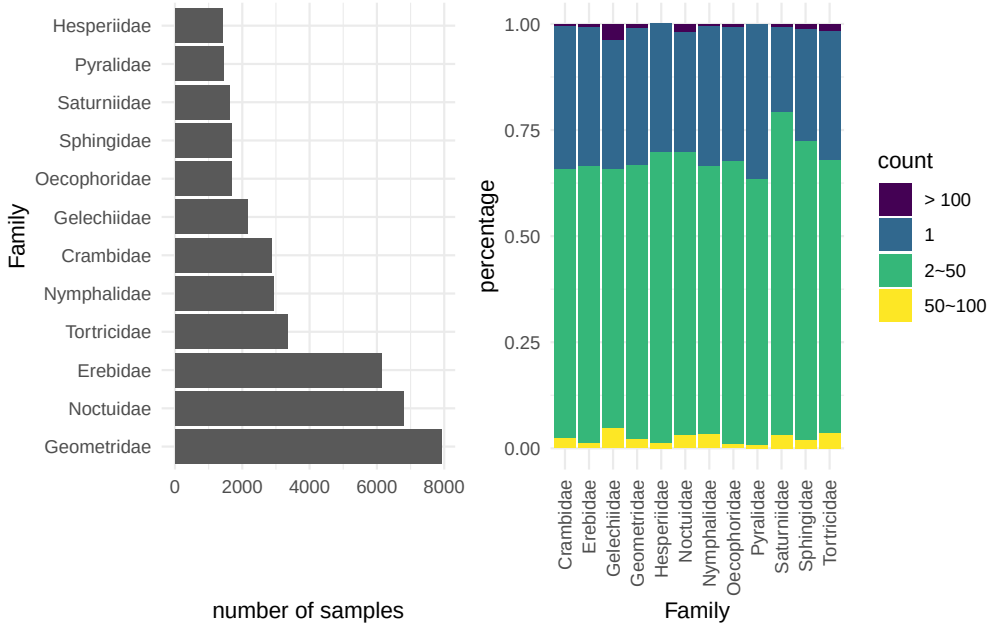

Figure 2: (a) Genus sample size for each family; (b) Percentage of genus sample count within each family

The number of distinct k-mers grow exponentially as their length increase (4^n where n is the number of nucleotides). While we want to maximally incorporate DNA information by including k-mers of various lengths as predictors, many k-mers have similar counts across 12 families and provide little information for classification. We decided that if a k-mer has very few unique values relative to the number of samples and if the ratio of the frequency of the most common value to the frequency of the second most common value is large, the k-mer should be excluded due to its low variance. The cutoff for the percentage of distinct values out of the number of total samples is set at 0.1, and the cutoff for the ratio of frequency is set at 95/5. If a kmer's frequency of its most prevalent value over the second most frequent value is above 95/5 and its percentage of unique values is below 10%, it is flagged as a near zero variance predictor to be excluded. A few k-mers have relatively large frequency ratio within their groups (Fig. 1) and low percent of unique values (not shown) for exclusion. Different

cutoff values will be tested to examine important kmers and the model's performance in sensitivity analysis.

To classify Genus within each Family of Lepidoptera, we first explored the distribution of sample count of each Genus, which ranges from 1 to 432. Most of the Genus has 2~50 sample counts, and a few has more than 50. However, almost all families have at least 25% of the genus with only 1 sample. Various methods will be explored in the following sections to address this issue.

Based on the distribution of the sample sizes across families, we see that the families are inherently unbalanced, with Geometridae the family having the most number of samples having over 4 times the sample size of families such as Sphingidae, Pyralidae and Hesperiidae. This inbalance could translate later on to the differences in predictive confidence in the family model due to the disparaging levels of information.

2 Methodology

2.0.1 Model Selection

Prior to fitting the model, we conducted model selection through identification of near zero variance predictors, which either (1) have one unique value, or (2) have very few unique values relative to the number of samples and the ratio of the frequency of the most common value to the frequency of the second most common value is large. We removed these variables to achieve improved model performance because predictors with low variance have little or no predictive power and do not influence the outcome classification significantly, which allowed us to reduce the number of features from 5460 to 2352.

2.1 Predict Lepidoptera families

In order to achieve maximum performance on the multiclass classification problem, using the counts of the combined kmers ($k = 1, 2, 3, 4, 5$) as predictor variables, we fitted four models and assessed their accuracy on test set via a 80:20 split between training and test data.

Table 1: Model accuracy comparison

LDA	QDA	Naive Bayes	KNN (k = $\sqrt{N}/2 = 89$)
94.4%	93.0%	79.8%	82.4%

Before fitting the LDA model, we applied Principal Component Analysis (PCA) on the features (with threshold of 95%) in order to (1) further reduce the dimensionality, (2) more importantly, to avoid applying LDA directly on the count data since applying PCA before LDA ensures that we remove the effects of overdispersion. The accuracy rates obtained on the test set reveals that LDA proves to be the optimal model. To further refine this model and find the optimal interval for the kmers, we trained the model on combined kmers from 1 through 5, from 1 through 6 and from 1 to 7 respectively and obtained the following out-of-sample accuracy rates which we used as the model selection criteria:

Table 2: LDA Results

Kmer	Accuracy	Total predictors
k=1,2,. . . ,5	94.4%	1364
k=1,2,. . . ,6	**95.6%**	**5460**
k=1,2,. . . ,7	95.7%	21844

Although the LDA model improves by a whole percent from using the 1-5 to 1-6 kmers as features, it does not continue to improve significantly beyond the 1-6 kmers. Adding data from 7 kmers for the model to train on

only improves the accuracy by 95.7%. Therefore, we decide to train our final model using 1-6 kmers, since it strikes an optimal balance between accuracy and computational efficiency. Ultimately we derived an in-sample accuracy of 96.8%, and an out-of-sample accuracy of 95.6% on this final LDA model using 1-6 kmers.

Table 3: Accuracy of various frequency cut values

	10	40	80
Accuracy of Family Model	96.2	96.3	**96.4**

We additionally tuned the parameter frequency cut for the nearZeroVariance function. The parameter represents the cutoff for the ratio of the most common value to the second most common value. Larger values keep more columns, whereas smaller values cut out more columns. The default value is 19. From the table we can see changing the value from 19 to 80 results in an increase of accuracy of a mere 0.1% from the un-tuned final model. Therefore, we decided to keep this final LDA model using 1-6 kmers with freqCut of 19.

Furthermore, we tuned the PCA threshold hyperparameter to an optimal 0.999, which improved the accuracy from 95.6% to 96.3%. We accessed our data imputation approach with the missing loci, and found that applying the ambiguous base imputation approach improved the accuracy by 0.1%. Albeit not significant, our imputation approach proves a valid approach without extrapolation or bias on the existing data.

2.2 Predict Lepidoptera genusus

Due to the effective performance of the family model, we continued to use LDA as our final model to predict genuses, after tuning the PCA threshold (the minimum level of variance or information that must be retained when reducing the dimensionality) for optimal accuracy. Similar to the family model, we used kmers where k equal 1 through 6 as predictors, after removing the near zero variance predictors, then we used PCA to reduce dimesionality as well as the dispersion effect on the count data. For each of the 12 families, we created the genus predictor comparing the performance of LDA at various PCA thresholds via 80-20% split and obtained the following accuracy rates. The bolded accuracy represents the highest accuracy for the given genus, therefore the final model or our choice for the genus.

Table 4: Out-of-sample Accuracy for LDA under different PCA thresholds

Threshold	0.75	0.90	0.95	0.99
Noctuidae	90.8%	91.5%	92.0%	**92.1%**
Geometridae	80.0%	80.9%	81.9%	**82.2%**
Gelechiidae	87.8%	92.5%	**93.2%**	92.0%
Sphingidae	90.5%	**91.7%**	91.7%	91.7%
Tortricidae	85.6%	88.1%	**89.6%**	88.7%
Crambidae	77.4%	80.0%	80.5%	**81.7%**
Erebidae	81.9%	**82.4%**	82.2%	81.4%
Oecophoridae	78.3%	**81.0%**	80.8%	76.2%
Hesperiidae	84.5%	85.5%	**85.9%**	83.1%
Pyralidae	78.3%	80.0%	79.7%	**80.7%**
Nymphalidae	80.0%	**89.1%**	88.5%	88.7%
Saturniidae	**93.4%**	92.7%	92.2%	91.9%

We also assessed the performance of the in-sample accuracy through the training data:

Figure 3: In-sample accuracy for LDA family model

2.3 Predict New Species

We predicted new species by setting a threshold of 0.8; that is, (1) we classify a DNA sequence as new family if the predicted probability from the family model is below 0.8, and (2) for a sequence predicted from an observed family (if the predicted probability from the family model is above 0.8), we classify it as a new genus if the predicted probability from the corresponding genus model is below 0.8. As a result, we classified 138 out of 7000 samples in the unlabeled data as new species. We found 12 families and another new family category, along with 1606 genera along with the new genus category (Figure 6(a)).

2.4 Model Assumptions

To ensure appropriateness of the LDA model, we assessed conditions for independence and normality. The individual sequences are independent, therefore the samples meet LDA's assumptions that the features are independent of one another. Additionally, the principal components are normalized; therefore the data meets the LDA assumption that the distribution of each feature within each class is normal.

3 Results

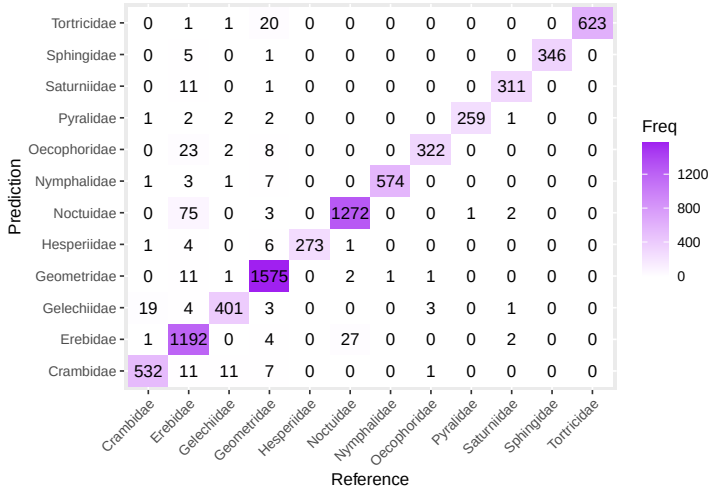

Figure 4: Confusion matrix for family model

We assessed the performance of our family model using a confusion matrix, through which we discover that Erebidae, Noctuidae and Geometridae are most frequently classified correctly, mainly due to their large sample

sizes which provide ample information to the model. On other other hand, there is 75 samples, a noticeable number of observations predicted to be Noctuidae when it is actually Erebidae. Similarly, there is 27 samples predicted to be Erebidae when it is actually Noctuidae. A literature search suggests that these two families are indeed similar to each other; in fact, some subfamilies of the Noctuidae, such as the Herminiinae, were moved as a whole to Erebidae due to their similarities (Donald Lafontaine 2010). This corroborates our results that these two families tend to be confused by the model, despite high performance on other families.

According to Table 2 above, the genus models have generally moderate to strong performance, but vary in terms of out-of-sample accuracy. For example, Pyralidae has significantly lower accuracy at 80.7%, which could be mainly explained by the presence of singletons within the genus according to the EDA plot (Fig. 2b). Simiarly, other families with large presence of singletons such as Crambidae, Erebidae and Geometridae all have out-of-sample accuracies at around 80%, which suggests that the influence that singletons have in undermining prediction accuracy.

Furthermore, due to the nature of the multi-class classification problem, we used the macro-average measure to assess the performance of our final models. Macro-average AUC is computed by first calculating the AUC for each individual class, and then taking the average of these AUC values, allowing us to evaluate the performance of a classifier on each class separately, and avoid the issue of class imbalance affecting the overall score.

Table 5: Macro AUC Value for Genus Models

Family	AUC
Noctuidae	0.9689
Geometridae	0.9689
Gelechiidae	0.9809
Sphingidae	0.9695
Tortricidae	0.9609
Crambidae	0.9460
Erebidae	0.9632
Oecophoridae	0.9533
Hesperiidae	0.9755
Pyralidae	0.9760
Nymphalidae	0.9658
Saturniidae	0.9776

Based on the reported macro AUC values, we observe that the family classification models have strong performance in terms of their abilities to differentiate between the classes. Most of the family models have high AUC values above 0.95, which suggests that the models are performing well in distinguishing between the classes.

We graphed the distribution of the predicted probabilities for the genus models over correct and incorrect classifications separately. As shown in the histogram below showing the predicted probabilities for correct predictions (left), most of the genuses have predicted probabilities near to 1 without noticeable deviance. Therefore, we are fairly confident in our predictions, even though a minority of genuses such as Erebidae, Tortricidae, and Pyralidae have a scant amount of probabilities in the lower spectrum mainly due to the small sample size. As shown in the predicted probabilities for incorrect predictions (right), genus with small sample sizes such as Saturniidae have noticeably lower probabilities and thus lower confidence.

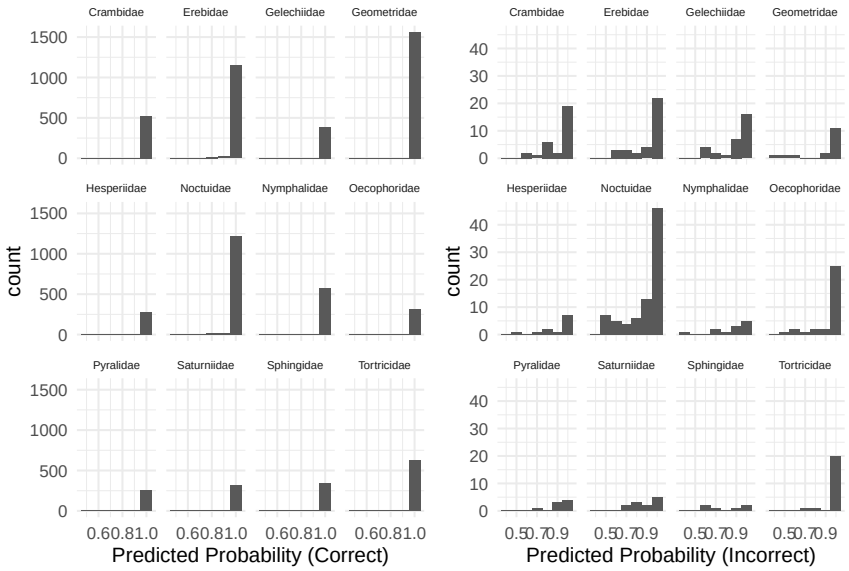

Figure 5: Predicted probabilities for family model

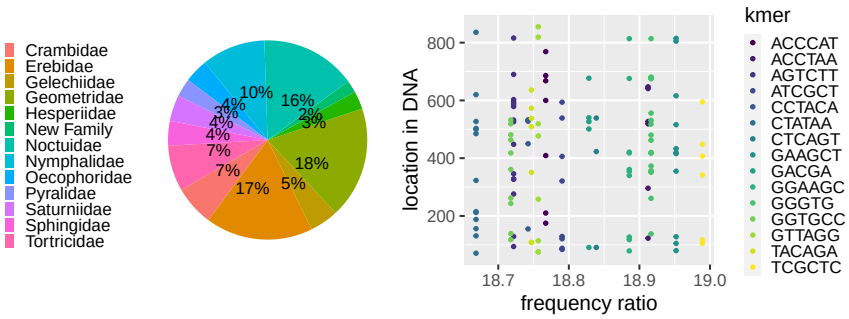

Figure 6: (a) Family composition of unlabelled data; (b) Loci of the 15 most varied kmers

We investigated the locations where the 15 most varied (defined by high frequency ratio and percentage of unique values) kmers are located in whole DNA sequences. 12 important kmers are 6-nucleotide long, and 3 important kmers are 5-nucleotide long. Although it is difficult to inspect which specific location is important because the loci of these kmers are spread throughout whole DNA sequence, loci 400-600 have the most instances of important kmers (Fig. 6b).

4 Conclusion

Research Question

In this study, we explored different strategies for classifying Lepidoptera families and their corresponding genuses. During the data processing stage, we found that extracting Kmers of different lengths from whole DNA sequences is an effective method to compensate for missing nucleotides at various loci. Among 4 classification models — Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Naive Bayes, and K-nearest neighbors, the LDA model with kmers of length 1-6 as predictors presents the highest out-of-sample accuracy of 95.6% to predict 12 families and corresponding genuses. Our confidence that each family/genus belongs to its predicted category is also high (Fig. 5). Therefore, as long as the local sequence patterns can be extracted, neither the whole DNA sequence nor any individual loci are important for classification. At the family level, the former is supported by the evidence that the number of features (kmers) is reduced greatly while still producing a high accuracy and strong classification ability (AUC). The latter is validated by our kmer approach which

obtained such high out-of-sample accuracy despite compressing information about locations in the sequences. While individual loci can be difficult to identify, loci 400-600 contain the highest number of important kmers. Kmers is sufficient to capture local patterns that are functionally important (Rahman et al. 2018).

Sensitivity analysis

We varied the number and lengths of kmers used as predictors in the LDA model to compare how sensitive the family prediction accuracy is to different numbers and lengths of kmers (Table 1). Although the number of predictors ranges from 1364 to 21844, the family prediction accuracy did not fluctuate much. We observe that while using k from 1 through 6 versus from 1 through 5 improves accuracy from 94.4% to 95.6%, on the other hand, increasing to k from 1 through 7 improves accuracy by a mere 0.1% and is computational efficient. Some genus predictions are sensitive to Principle component analysis thresholds such as Oecophoridae, Gelechiidae, and Crambidae, while others are more consistent regardless of the thresholds (Table 2). In genus classification, we applied different PCA thresholds on LDA model to compare how they affect the prediction accuracy.

Table 6: Accuracy of various frequency cut values

	10	40	80
Noctuidae	98.8%	**98.9%**	98.6%
Geometridae	81.7%	**82.3%**	**82.3%**
Gelechiidae	92.7%	93.2%	93.2%
Sphingidae	92.0%	92.0%	**92.9%**
Tortricidae	88.0%	88.8%	89.6%
Crambidae	80.5%	81.0%	**81.8%**
Erebidae	81.2%	81.7%	82.0 %
Oecophoridae	72.7%	73.0%	73.9%
Hesperiidae	84.2%	83.4%	83.4%
Pyralidae	80.0%	80.0%	80.3%
Nymphalidae	87.8%	88.3%	88.8%
Saturniidae	91.8%	92.2%	92.8%

We tuned the same value of the nearZeroVariance function for all of the genus family. Overall the validation set accuracy of 4 models improved after tuning this paramater, indicated by the bold.

5 Limitations & Future Directions

Although we achieved high prediction accuracy for the Lepidoptera family and genus prediction, there are several limitations. First, almost all families have at least 25% of the genus with only 1 sample. During the model training stage, there is a certain chance that we lose information about these genus because they are not included in model-fitting models. Second, the loci information we lack might be especially important to classify specific families. When several families are genetically similar to each other, kmer might not be sufficient to differentiate them. For example, we found that Erebidae and Noctuidae are almost entirely mistaken for each other. It's possible that a few nucleotide differences at specific locations are key in this case. If more computing power is available, increasing the length of kmers to 31 on 64-bit machines could preserve more sequence information and improve accuracy further. If loci information is proven to be central in a certain scenario, labeling nucleotides along with their ordering in a whole sequence as predictors are also worth considering.

6 References

Donald Lafontaine, Christian Schmidt. 2010. "Annotated Check List of the Noctuoidea (Insecta, Lepidoptera) of North America North of Mexico." https://doi.org/https://doi.org/10.3897/zookeys.40.414.

Kress, W. John, and David L. Erickson. 2008. "DNA Barcodes: Genes, Genomics, and Bioinformatics." *Proceedings of the National Academy of Sciences* 105 (8): 2761–62. https://doi.org/10.1073/pnas.0800476105.

Matyas Cserhati, Chittibabu Guda, Peng Xiao. 2019. "CK-Mer-Based Motif Analysis in Insect Species Across Anopheles, Drosophila, and Glossina Genera and Its Application to Species Classification." *Computational and Mathematical Methods in Medicine* 2019. https://doi.org/https://doi.org/10.1155/2019/4259479.

Pereira, Filipe, Joao Carneiro, and Antonio Amorim. 2008. "Identification of Species with DNA-Based Technology: Current Progress and Challenges." *Recent Patents on DNA & Gene Sequences* 2 (3): 187–200. https://doi.org/10.2174/187221508786241738.

Rahman, Atif, Ingileif Hallgrímsdóttir, Michael Eisen, and Lior Pachter. 2018. "Association Mapping from Sequencing Reads Using k-Mers." Edited by Jonathan Flint. *eLife* 7 (June): e32920. https://doi.org/10.7554/eLife.32920.