Esther Olabisi-Adeniyi
Student ID: #1238826
Software Tools (BINF6210) – Assignment #2
10/31/22

Supervised Machine Learning: Taxonomy Classification Based on the COI gene

Advancements to the field of bioinformatics has allowed for the application of powerful tools like Machine Learning (ML). In Supervised ML, a prediction model is trained with known datasets so that when new datasets are introduced, the model is able to predict information about them based on what it learned from training data. With such a tool, a study by Meher et al. was able to distinguish between species based on COI gene kmer frequencies and they obtained higher success rates than observed in similarity-based, tree-based, and diagnostic-based approaches (2016, p. 322). Similarly, in my project, I created a Random forest (RF) classifier that uses single-nucleotide and 3mer frequencies of the COI gene in identifying taxa as either Asteroidea (starfish) or Tunicata (tunicates). I thought it would also be a good idea to compare the performance of a taxonomic classifier that uses single-nucleotide frequencies with one that uses 3mer frequencies.

The comparison is relevant as it helps to understand whether one variable type would be more efficient than the other. For instance, in some cases, single-nucleotide proportions may not differ enough for efficient separation, whereas kmers frequencies of k>1 may. Therefore, comparing their performances allows for observing such differences. I expect the 3mer-based classifier to distinguish between taxa better than the single-nucleotide one because kmers are specific in how the nucleotides are placed. I imagine that such specificity may be better for identification compared to single nucleotide proportions. This project further analyzes the RF taxonomic classifiers by evaluating variable importance. Variable importance shows the specific nucleotide-proportion and 3mer most crucial to the taxon identification process.

Discussion and Conclusion

The final sample size for this project was 2,624 where each taxon had 1,312 sequences. All sequences were refined to get relative uniformity in lengths (Fig 1 and Fig 2). Approximately 70% of the sample size (1,838 of 2,624) were used in training the RF taxonomic classifiers while the remaining 30% (393) were used to validate the classifiers. The taxonomic classifier distinguishing based on single-nucleotide proportions is the "Nucleotide-based classifier" while the one utilizing 3mers is the "Kmer-based classifier." The nucleotide-based taxonomic classifier mostly grouped taxa successfully (OOB error rate = 0.22%) while the kmer-based classifier also performed well but with a 0.33% OOB error rate. The nucleotide-based classifier correctly identified all the data entered for validation while the kmer-based classifier correctly identified almost all. Both classifiers successfully accomplished the objective of creating RF taxonomic classifiers. However, contrary to expectations, the nucleotide-based classifier performed better as inferred from the former's lower error rate. It may be that the taxa's COI sequences evolved in such a way that certain nucleotides are still grouped similarly (like in 3mers), thus making separation more complicated than with single-nucleotide proportions.

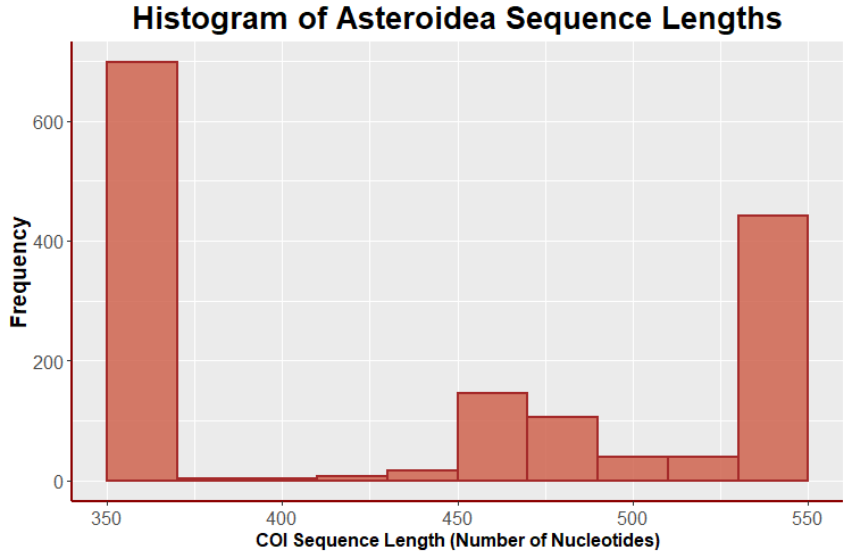

Fig 1. The distribution of COI sequence lengths in taxon Asteroidea after constraining sequence lengths, avoiding extremely short or extremely long sequences that may confound taxon identification. Although most of the sequences are around 350-375 and 525-550 nucleotides (nt) long; they are relatively spread out over the defined range of 350 to 550 rather than the 125 to 600 range found in the original datasets.

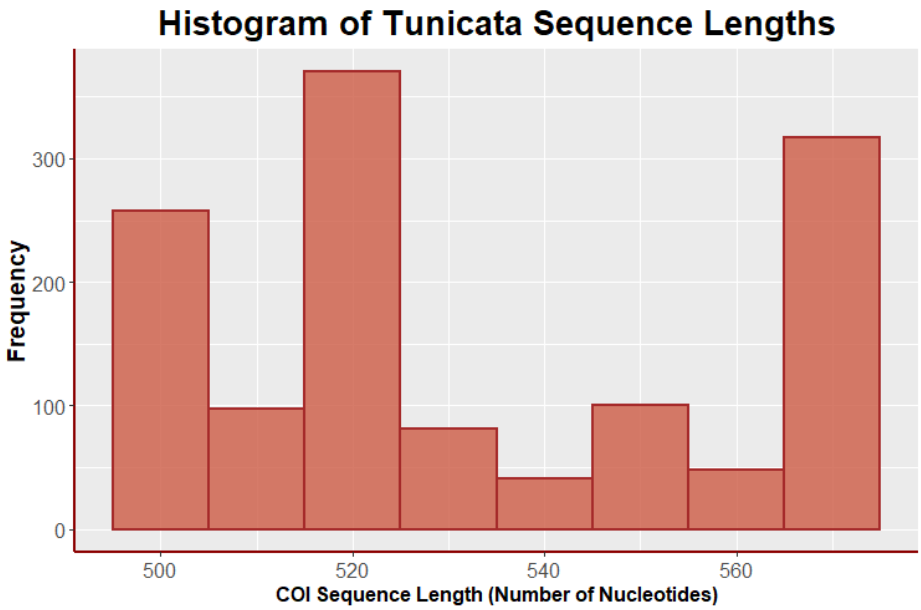

Fig 2. The distribution of COI sequence lengths in taxon Tunicata after constraining sequence lengths to avoid extremities as in Fig 1. These sequence lengths are even more relatively spread out over the defined range of ~500 to 580 nucleotides.

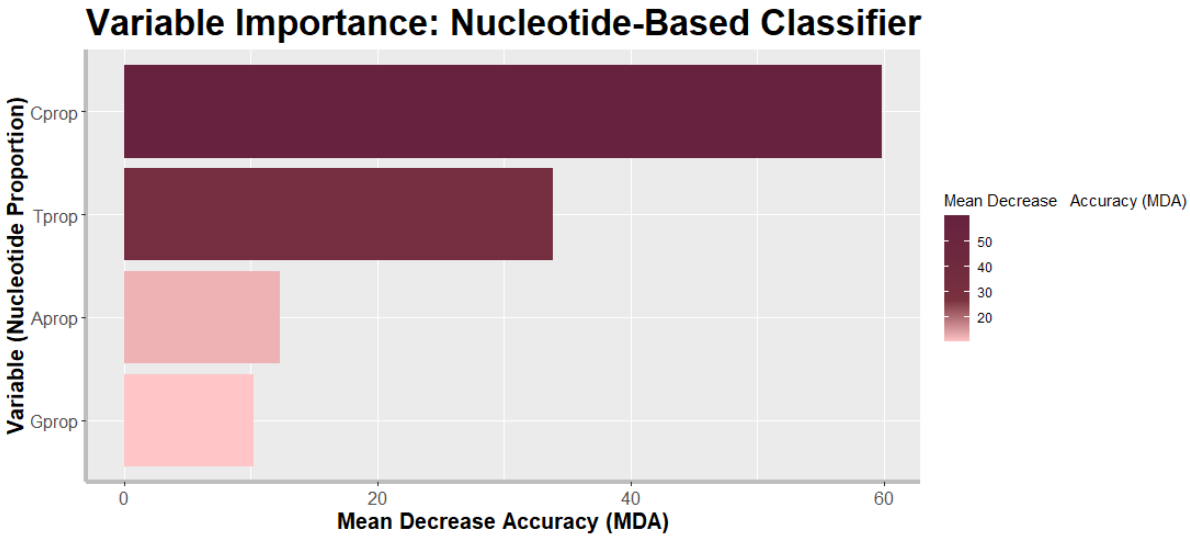

Fig 3. Variable importance plot based on mean decrease accuracy for the nucleotide-based classifier. The proportion of C is the most important variable in distinguishing between Asteroidea and Tunicata COI genes, and it is followed closely by the proportion of C nucleotides.

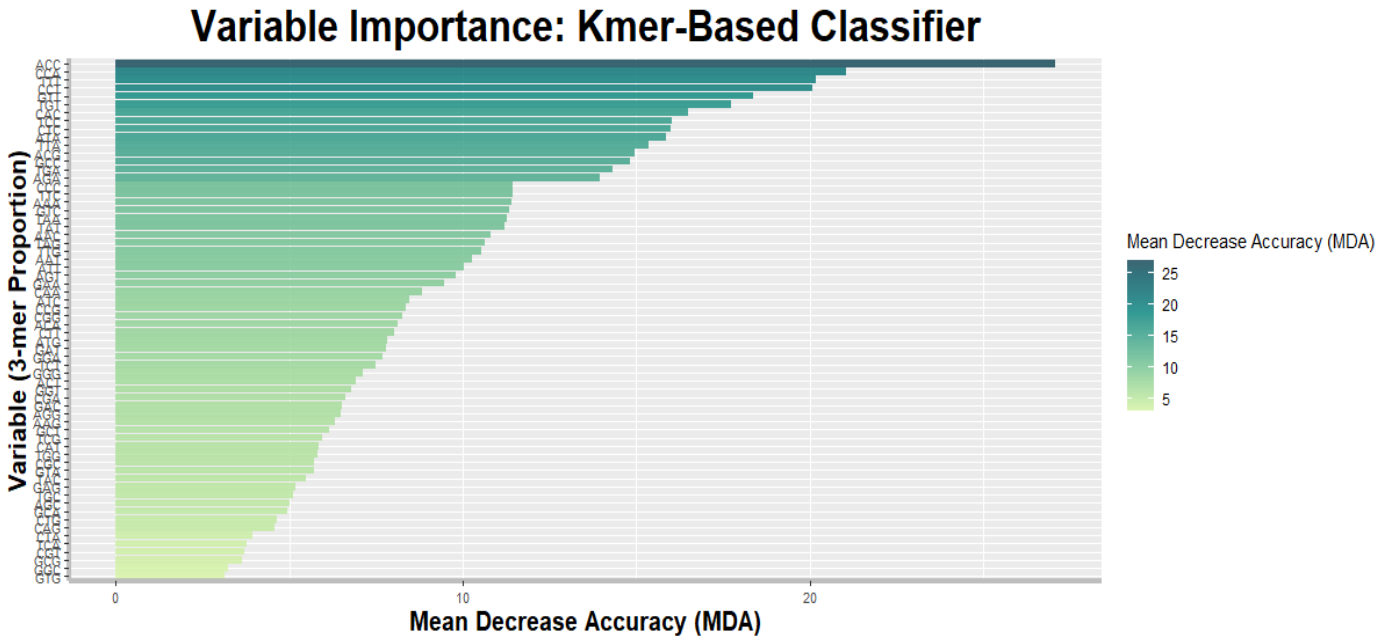

Fig 4. Variable importance plot based on mean decrease accuracy for the kmer-based classifier. The proportion of 3mer ACC is the most important variable in distinguishing between Asteroidea and Tunicata COI genes while GTG is the least important. Some pairs like TGT and GTT appear to be equally important.

Variable importance was evaluated based on mean decrease accuracy for both classifiers (Fig 3). The mean decrease accuracy value represents the amount of accuracy RF forfeits when a variable is not considered in the classification process (Martinez-Taboada & Redondo, 2020, Fig 2.). That is, the higher the MDA value, the more important a variable is. Of the four nucleotide-proportions utilized as variables in the nucleotide-based classifier, C's proportion was the most important while G's was the least important (Fig 3.). The most important variable for 3mer based identification was ACC (Fig 4.). It could be inferred that both taxa greatly differ in their composition of the ACC 3mer and the identification process would suffer if this variable were excluded. Overall, this project suggests that a 1mer-based (nucleotide-based) RF, rather than a 3mer-based RF, may be more efficient for identifying taxa by COI sequences. A future study

would involve creating 2mer and 4mer based RF classifiers to see if the single-nucleotide-based one still outperforms all.

References

Martinez-Taboada, F., & Redondo, J. I. (2020). The SIESTA (SEAAV Integrated evaluation sedation tool for anaesthesia) project: Initial development of a multifactorial sedation assessment tool for dogs. *PLOS ONE*, *15*(4), e0230799. https://doi.org/10.1371/journal.pone.0230799.g002

Meher, P. K., Sahu, T. K., & Rao, A. R. (2016). Identification of species based on DNA barcode using k-mer feature vector and Random forest classifier. *Gene*, *592*(2), 316–324. https://doi.org/10.1016/j.gene.2016.07.010

https://stackoverflow.com/a/56306139 (for coding section).

https://stackoverflow.com/a/16968999 (for coding section).

Acknowledgement

I thank the BINF6890 teaching assistant, Jessica Castellanos Labarcena, for clarifying concepts about this assignment.