**Bio 4990/6990: Computational Activity—Random Forests**

1. Describe the model we are fitting in this activity:
   1. Features:
   2. Response:
   3. Training Data Source:
   4. Algorithm:
2. Trees in the Forest!

**\*6990 make sure you used a for loop to explore\***

* 1. How does changing the number of trees in the classifier impact performance?
  2. How many trees do you think are sufficient for the classifier?

1. Performance of the classifier:
   1. What is the out-of-bag score for the classifier (using the number of trees you selected above)?
   2. What is the accuracy on the training dataset (overall, and per class)?
   3. What is the accuracy on the testing dataset (overall, and per class)?
   4. Is the out-of-bag accuracy more similar to the training or testing accuracy, and why?
   5. Based on these metrics, do you think your model is struggling with bias or variance?
2. Variable Importance:
   1. What are the most important variables based on GINI impurity, and how might you interpret this biologically?
   2. What are the most important variables based on permutation importance (MDA), and how might you interpret this biologically?
3. Explain how this model could be used to make predictions.
4. Report on the results from the model you trained to make predictions across all categories (instead of just LC vs non-LC). Report on the following:
   1. Out-of-bag accuracy.
   2. Confusion matrix.
   3. Variable importance (impurity).
   4. Variable importance (permutation).
   5. Do you favor this model or the original model for downstream use?
   6. Report the accuracy of the selected model on the validation dataset.