BIOL 324 ImageJ and randomForest Tutorial

MODULE 1

*Before starting this weeks tutorial you must:*

1. Download the program for measurement, ImageJ, [here](https://imagej.nih.gov/ij/download.html)

2. Download the class google sheet in the file format .csv

[Note: if you can’t use imageJ for any reason, don’t panic. Display the digital image of the herbarium sheet on any device and use a ruler held up to the screen to make measurements]

**Using ImageJ for Measurement**

ImageJ is an image processing software that was developed by the NIH. It is used to analyze different types of images but this tutorial will cover how to measure certain morphological features from photographs of herbarium specimens.

To use ImageJ you have to import the photographs you want to work with.

1. Drag the file to the lower grey bar. As you are doing this the words <<drag and drop>> should appear. Once the image has been dropped it will open and now it is ready to analyze in ImageJ.

*There are many measurements that can be taken with ImageJ. It might be helpful to click through the different buttons at the top of the bar to explore all of the different ways ImageJ can make measurements. This tutorial will use the line button for all of the measurements.*

2. Zoom in on the ruler (Mac: Command shift =, Windows: CTRL + ). Most digitized herbarium specimens will have a ruler in the photograph. This is to help standardized it and is useful for measuring features in ImageJ.

3. Use the line button (5th button from the left) and draw a line by clicking and dragging along the bottom of the ruler so that the line goes from the 0 tick to the 5 cm tick. This needs to be very accurate because we are going to use this length to calibrate the rest of the measurements taken in ImageJ.

4. Go to the top bar and hit “analyze”. Then go to “set scale…”. Enter 5 into the Known distance box and enter cm into the Unit of Length box. Hit the box “Global”. By checking “Global” you are ensuring that this is applied to all photos in the current session of ImageJ. Then hit OK. ImageJ is now calibrated using the ruler from your herbarium photo.

Now you can go on to take measurements.

5. Use the line or the segmented line (you can change between them by right clicking on the line button) draw lines across what you want to measure. Once you have the right length hit (Mac: Command M, Windows: CTRL M) and ImageJ will measure and record the length in a convenient table for you. This table should pop up once you take your first measurement. When you have all of the measurements you need you can either download that table that ImageJ made or just copy and input the data in your own table. You will have to eventually input your data in a group google sheet so keep that in mind as you are recording data.

6. Once you have taken and recorded all of the measurements necessary put your findings into the class google spreadsheet so that other people will be able to access, download, and use

your data for the next step.

[Here](https://imagej.nih.gov/ij/docs/guide/user-guide.pdf) is a link to the ImageJ user guide. All the information you will need is in this document, but if you are interested in the program and the other features available, the user guide can provide more information.

MODULE 2 – week 2

**randomForest in RStudio**

randomForest is a package in R that uses machine learning for classification, regression, and clustering. It combines the simplicity of decision trees with flexibility which makes the result more accurate than simply using decision trees alone. There are many tutorials and videos online about randomForest but [this video](https://www.youtube.com/watch?v=J4Wdy0Wc_xQ) is useful if you want more background.

There are a few ways to use randomForest but this tutorial will cover using the package in unsupervised mode. This means that the program will have to find the structure of the data itself, without us training it first.

1. Open RStudio and set your working directory. You want the working directory to be wherever the data set you are going to input is located.

2. Import some packages for randomForest to run properly. These are randomForest, ggplot2, and caret. This only needs to be done once so if you have already downloaded any of these—skip this step. *Use the function install.packages() to do this. Type of the name of the package in the parenthesis.*

3. Set the seed for the random number generator. This can be any number. Line 5

4. Load the newly installed packages into the script. Lines 8-10

5. Load data into R. This is done with the read.csv function. This function requires your data to be in a .csv format. Line 14

6. Run the randomForest function on the data. Line 23

~*It should be noted here that running randomForest in classification or regression mode will require the forest to be “tuned” after the initial run to produce the most accurate results. This is not possible in unsupervised mode and the default values for ntree and mtry should work fine for the purposes of this exercise.*~

7. Create a distance matrix to see how samples are different from each other. Lines 31, 34

8. Calculate the values needed for the X and Y axes Line 37

8.Format the data for graphing Lines 40-44

9. Graph an MDS plot Lines 46-49

MDS (multidimensional scaling) is a way to represent a matrix of distances, distances being the dissimilarities between herbarium specimens as provided by randomForest, as a two dimensional plot. The purpose of this lab is to use MDS as a tool, and it is not our purpose here to explain it in detail. However, if you are interested your instructor can provided reading lists for further study. This plot should show two distinct groups that specimens fit into.

~*Another thing to note. randomForest has a hard time dealing with NA values in data. There is a function, rfImpute() that can help format your data. If you are curious it is at the bottom of the example code for this tutorial*~

This code is based on [this](https://www.youtube.com/watch?v=6EXPYzbfLCE&t=806s) video. This video shows a different way to use randomForest (classification instead of unsupervised mode) but the graphing component is the same. Skip to 13:16 if you want to just see the graphing.