Amphibian Mitochondria- Who’s Related to Whom?

Final Paper for BIOL792- Data Science for Biology II

Ellen Wilson

May 5, 2021

Last semester, I was in Dr. Parchman’s Data Science for Biology I class, and I really enjoyed it. For that project, I found and visualized the genetic diversity between populations of pika (Ochotona princeps) with data collected by Dr. Kelly Klingler. This helped prepare me for my own Columbia spotted frog (Rana luteiventris) project. I created a working directory in Ponderosa, and used cleaned pika nuclear DNA data. I used the dDocent pipline that filtered out the raw data, and it returned files for me to analyze. I used VCF tools to continue filtering, and ultimately created 3 final files. For the final step, I used R to create a graph that visually depicted the populations of pika and the genetic diversity among them. I found that the populations that shared a mountain range shared a higher similarity of genetic information than those populations on another mountain range. I found this very interesting.

This semester for my project, I wanted to do something similar to this concept. However, instead of entire nuclear DNA genomes, I focused on mitochondrial DNA of 11 different species. The species included: Siberian wood frog (Rana amurensis), wood frog (Lithobates sylvaticus), California red-legged frog (Rana draytonii), Dybowski’s frog (Rana dybowskii), European common frog (Rana temporaria), gray treefrog (Dryophytes versicolor), axolotl (Ambystoma Mexicanum), African clawed frog (Xenopus laevis), tiger salamander (Ambystoma tigrinum), Asiatic toad (Bufo gargarizans), and human (Homo sapiens). Mitochondrial DNA (mtDNA) fascinates me because the only changes in nucleotides are from mutations, as there is no recombination, and is only passed maternally. I was curious to see which species were more closely related to whom, and where humans fall in the ancestry. I had three hypotheses: All of the frog and toad species’ mtDNA will be closely related to each other, the salamander species’ mtDNA will be related to one another but farther distanced from the other species, and human mtDNA will be far-removed from all other species.

I first downloaded FASTA files to my desktop of each of these species’ mtDNA from the National Center for Biotechnology Information (NCBI) website. Using my Unix terminal skills, I created a directory for my amphibian project and moved all of these files to my directory. I then turned them into text files so they could be easily manipulated. Next, I went to the Clustal Omega Multiple Sequence Alignment tool. I decided to use three species as reference species to compare the others against. Those were the Asiatic toad (Bufo gargarizans), the European common frog (Rana temporaria), and the axolotl (Ambystoma Mexicanum). These three species are of different genera, and could give more insight as to which species are more closely related. To use the Multiple Sequence Alignment tool, I copied and pasted each species’ mtDNA, and ran them against the reference species’ mtDNA. The output of each alignment showed how many nucleotides were the same and different. I found how many nucleotides matched per pair and entered them into an excel sheet for analysis. The excel sheet included the taxa names, the number of total nucleotides for each mtDNA, the total number of nucleotides that matched each of the reference species, the continent the species lived on and their ICUN status of concern. For analyses of this data, I first went to Jupyter Notebook and played with some python code. I found that I could organize the matching nucleotides by each reference species in order, so I could find who was most closely matching with whom. I also used the code to find each of the species’ ICUN status. For data visualization, I used ggplot in RStudio to graph the data. I didn’t have a lot of data to play with, but I created three bar graphs that showed each species’ relatedness to the reference species in regards to their location. For the final step, I plugged all of the mtDNA reference genomes back into the Clustal Omega Multiple Sequence Alignment tool all at once to obtain a phylogenetic tree.

After my analyses, I found quite a few interesting things. Looking at the Jupyter Notebook results and the ggplot graphs, I could easily see who shared the most nucleotides with the reference species. The first was the easiest to hypothesize- the tiger salamander and the axolotl shared 15,624 nucleotides, and were highly related. The Asiatic toad shared the most nucleotides with the gray tree frog with 13,024 nucleotides, and the European common frog shared the most nucleotides with Dybowski’s frog with 13,361 nucleotides. All of these findings were not surprising. However, I thought that it was fascinating that those species who matched the most were not always from the same continent. For example, the Asiatic toad is from Asia and the gray treefrog is from North America. I was also blown away by how many nucleotides the human mtDNA matched with the reference species. Humans matched 10,055 nucleotides for the Asiatic toad, 9,963 for the axolotl and 9,855 for the European common frog. I was not expecting there to be that many similarities. All 11 species together have 5,400 share the same nucleotides! Looking at the final phylogenetic tree, my hypothesis for the salamander species was correct. They were closely related to each other, but farther removed from the other species. However, my other two hypotheses were not completely correct. Most of the frog and toad species were related and near each other on the tree, but the outlier was the African clawed frog. It was far removed from the other species, and more closely related to the salamanders. I was also incorrect about the human mtDNA. From my findings, we were not drastically removed from all other species, but instead in between the African clawed toads and the salamanders. Obviously, this was a small sample size and small study, so if I were to do this project again, I would incorporate a wider variety of species to see exactly where human mtDNA lies in reference to amphibious species. However, I had fun playing around with python code, ggplots and genomic databases, and I feel more prepared for future bioinformatic problems.

Literature Cited:

Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Research. 2019 Jul;47(W1):W636-W641. DOI: 10.1093/nar/gkz268.

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2021 May 05]. Available from: https://www.ncbi.nlm.nih.gov/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>.