# Bi/Ge105: Evolution
# Homework 2
# Due Date: Wednesday, January 22, 2020

"I am losing precious days. I am degenerating into a machine for making money. I am learning nothing in this trivial world... I must break away and get out into the mountains to learn the news." - John Muir

**1. A feeling for the numbers in evolution continued**

The processes of evolution take place at many different scales in both space and time. Like last week, the goal of this first problem is nothing more than to "play" with some of the characteristic scales associated with a broad range of processes in evolution ranging from the very small (e.g. number of mutations per cell in a bacterium after one round of replication) to the very large (e.g. the age of the Grand Canyon). These estimates are intended to be done using simple arithmetic of the "one-few-ten" variety (i.e. few times few is ten) and to give an order-of-magnitude picture of the phenomenon of interest. Take pride in your results and state and justify (with citations) the assumptions you make carefully and give a simple, intuitive description of how you came to your results. Please don't report rough estimates with long lists of "significant" figures.

(a) Genomes are one of the most interesting features of "living matter". One question of interest is the extent to which the "space" of possible genomes and gene products has been explored over the history of life. In a very interesting article by Whitman *et al.* called "Prokaryotes: the unseen majority", we learn of the vast numbers of bacteria on Earth, with the current estimate coming in at something like $10^{30}$ bacteria. The number of viruses is even greater with the so-called "virus to bacterium ratio" having a value of roughly 10, implying something on the order of $10^{31}$ phages on earth, suggesting that these viruses are the largest genomic reservoir on our planet. If we assume that over more than 3 billion years, these viruses have been steadily replicating in their cycle of infection and lysis, how many total viral genomes have there been in the history of life? (Obviously, this is a very rough estimate). Now, compare this number to the number of *possible* viral genomes, assuming that each viral genome is 50,000 bp in length.

What does this estimate tell you about the extent to which sequence space has been explored? Note: do the approximations, errors and uncertainties in our estimates have any bearing on our conclusions here?

(b) Let's continue with the theme of the enormity of sequence space in biology. As a warm up exercise in that regard, in this problem we are going to think about the space of possible proteins. In a 2001 Bioengineering seminar, Professor Frances Arnold made a startling remark regarding the astronomical number of possible protein sequences. In this problem we would like you to generate some intuition for the astronomical number of ways of choosing amino acid sequences. To drive home this point, Prof. Arnold noted that if we consider a protein with 300 amino acids, there will be a huge number of different possible peptide sequences. Compute how many different sequences there are for a 300 amino acid protein?

While interesting, this wasn't the shocking part. Prof. Arnold's provocative remark was that if we took only one molecule of each of these different possible proteins, it would take a volume equal to five of our universes to contain all of these different *distinct* molecules. Estimate the physical size of a protein with 300 amino acids. Justify your result, but remember it is an estimate. Next, find an estimate of the size of the universe and figure out whether Prof. Arnold made a reasonable statement. Was her estimate of the volume of all of these possible proteins too large or too small?

(c) Mutations are thought to be one of the main genomic ingredients of evolution. Given that the bacterial mutation rate is of order $10^{-9}$ per bp per replication, how many single base pair mutations do you expect to see in a 5 mL tube of bacteria that is saturated after an overnight culture? First, given a roughly 20 minute doubling time, figure out how many cells you expect in such a culture after 12 hours given that you started with only a single cell. Then, use the replication error rate quoted above, and make an estimate of how many times each possible point mutation in the bacterial genome will be found in that culture.

(d) We talked about the diversity of the living world in class, but with a distinctly macroscopic perspective that focused on organisms such as insects and animals. What about our knowledge of the diversity of the microbial world? Every time an electron microscope is used to take an image it cor-

responds to roughly a $1\mu$m $\times$ $1\mu$m area. The electron microscope is used to explore the structure of the nanometer scale world of cells, for example. Biology is a subject characterized by great naturalist voyages in which figures such as Humboldt, Darwin, Wallace, Huxley and Hooker traveled around the world to try and collect data on biological diversity. The point of this problem is to get a sense of the *microscopic* diversity explored. Make an estimate of the total area looked at in biological samples using electron microscopes in the history of science. How does this correspond to the area of the Earth? What do you conclude about the extent to which we have "explored" the microbial diversity on the planet?

(e) One of the early methods that was tried out for dating the age of the Earth was to estimate the rate of sedimentation and to figure out how long certain geological features would have taken to form. (This strategy didn't end up working for making reliable estimates - see G. Brent Dalrymple, "Ancient Earth, Ancient Skies: The Age of Earth and its Cosmic Surroundings" for an excellent history of the attempt to date the Earth). Using the attached figure of the Grand Canyon (Figure 2), estimate the rate of sedimentation during each of the periods / subperiods listed in the time scale (Permian, Pennsylvanian, Mississippian, Cambrian, and Proterozoic to the base of the Grand Canyon Supergroup). What might have caused the variation you calculate?

(f) Of course, in making an estimate like this, we have made a variety of simplifying assumptions which could substantially alter our estimates. What has been left out that you think might be important? What information would you need in order to improve your estimate of sedimentation rate, and how might you collect such information in the field or in the lab?

(g) One of the great controversies in the history of the development of our understanding of evolution had to do with the question of whether or not the Earth was old enough to accommodate the "slowness" of evolution. However, it is not at all clear how people knew how to assign any numbers to the debate. Lord Kelvin was able to make an estimate for the age of the earth and found an answer in the millions of years which was claimed to be too short. Let's examine the timing question by making some estimates about one of the key case studies from this course which is the evolution of whales. For simplicity, assume that at the time of the extinction of the dinosaurs (65

## CHAPTER X.

### ON THE IMPERFECTION OF THE GEOLOGICAL RECORD.

On the absence of intermediate varieties at the present day — On the nature of extinct intermediate varieties; on their number — On the lapse of time, as inferred from the rate of denudation and of deposition — On the lapse of time as estimated by years — On the poorness of our palæontological collections — On the intermittence of geological formations — On the denudation of granitic areas — On the absence of intermediate varieties in any one formation — On the sudden appearance of groups of species — On their sudden appearance in the lowest known fossiliferous strata — Antiquity of the habitable earth.

IN the sixth chapter I enumerated the chief objections which might be justly urged against the views maintained in this volume. Most of them have now been discussed. One, namely the distinctness of specific forms, and their not being blended together by innumerable transitional links, is a very obvious difficulty. I assigned reasons why such links do not commonly occur at the present day under the circumstances apparently most favourable for their presence, namely on an extensive and continuous area with graduated physical conditions. I endeavoured to show, that the life of each species depends in a more important manner on the presence of other already defined organic forms, than on climate; and, therefore, that the really governing conditions of life do not graduate away quite insensibly like heat or moisture. I endeavoured, also, to show that intermediate varieties, from existing in lesser numbers than the forms which they connect, will generally be beaten out and exterminated during the course of further modification and improvement. The main cause, however, of innumerable intermediate links not now occurring everywhere throughout nature, depends on the very process of natural selection, through which new varieties continually take the places of and supplant their parent-forms. But just in proportion as this process of extermination has acted on an enormous scale, so must the number of intermediate varieties, which have formerly existed, be truly enormous. Why then is not every geological formation and every stratum full of



Figure 1: Chapter X of Darwin's great book. One of the main stylistic approaches of Darwin's writings was that he always himself highlighted the things that he found were potential weaknesses in his work, and the sparsity of the fossil record was one such problem. That is the subject of our current problem.

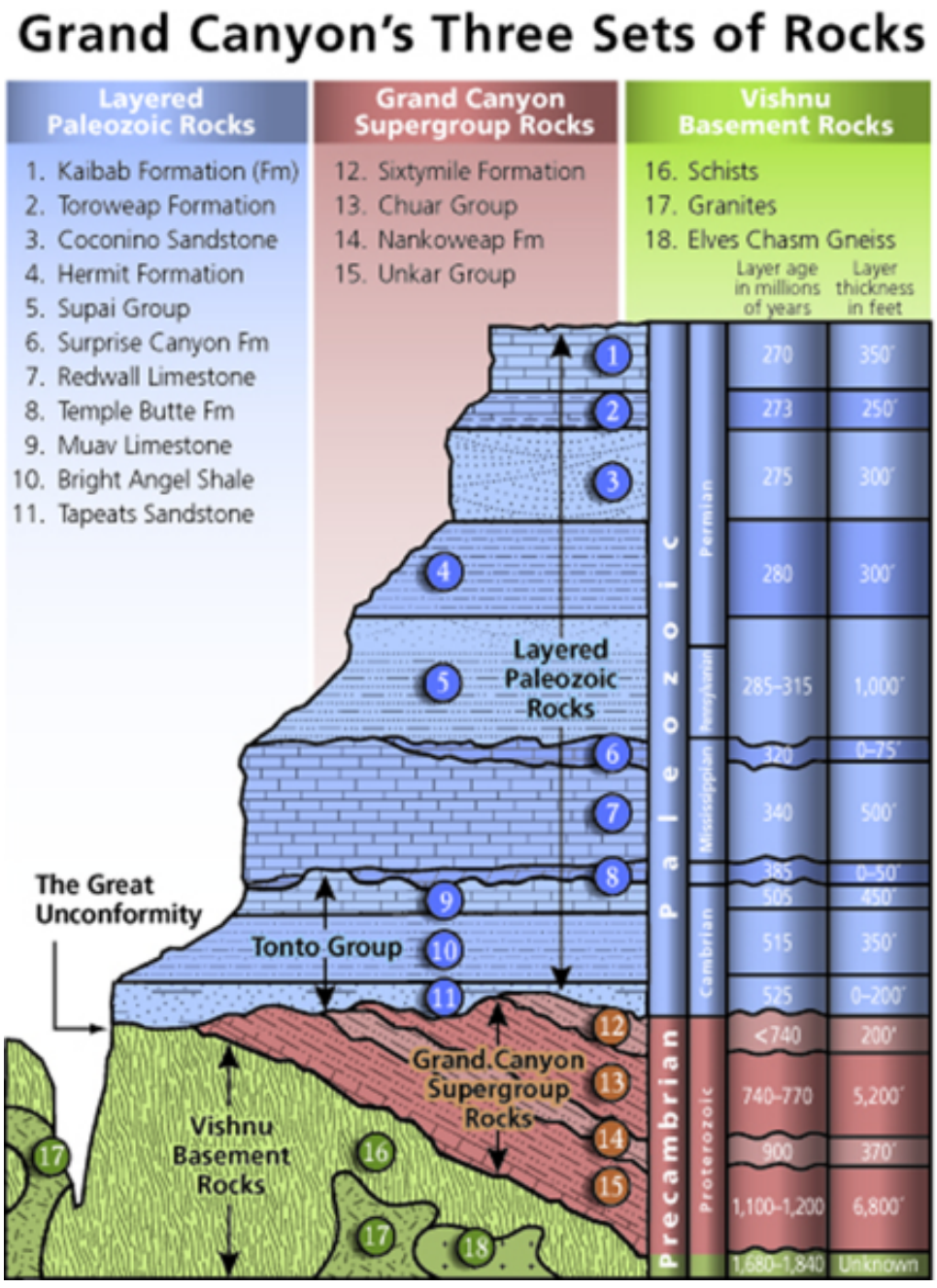


Figure 2: The Grand Canyon over time.

Ma), the mammalian ancestor of whales was $\approx 10$ cm in length. Using *Rodhocetus* and *Basilosaurus* as examples, figure out how much change in length there was per generation in the overall body plan in going from the post-dinosaur ancestor to these early whales. The logic of your estimate should involve figuring out when these whale ancestors lived, how big they were and estimating the typical generation time. It is not entirely clear that this estimate provides any insight into how whales actually evolved, but the numbers provide an interesting sense of how little it would take on a generation by generation basis to result in enormous structural changes over geological time scales. Also, it raises the question again of how and why scientists felt that the time scale proposed by Kelvin for the age of the earth was "too short" for evolution to have produced the world we see.

**2. How Did Frogs Get to São Tomé?**

One of the most important topics in evolution is biogeography, central to the thinking of both Wallace and Darwin as the idea of evolution took root in their minds. Given our own upcoming trip to the famed oceanic islands of the Galapagos, the question of how such islands are initially populated by animals of different kinds is a central question. In this problem we will explore the fascinating example of the frogs of São Tomé as a compelling story in biogeography. Specifically, as with the laboratory exercise you are doing on DNA from birds, we will explore in more detail the way that DNA sequence was used as a window into the dispersal of frogs onto the oceanic islands of the Gulf of Guinea.

São Tomé is an island located 255 km off the west coast of Africa. Volcanic activity formed this island roughly 13 million years ago, and continued to shape the landmass until as recently as the last hundred thousand years. Nevertheless, due to their considerable distance from the African coast and how recently they emerged from beneath the surface of the water, the islands in the Gulf of Guinea are a clear example of biodiversity due to dispersal. While birds may have flown to the island and seeds may have dispersed via birds or storms carrying them, the question of how amphibians traversed such far distances is harder to resolve for reasons having to do with their low saline tolerance. To understand just how challenging this journey is, in this problem we will compare the *Ptychadena newtoni* species to other *Ptychadena* species to determine the São Tomé inhabitant's origin.

**Enter the Sequence Revolution?**

As illustrated in lab, DNA sequencing is a powerful tool to determine the phylogenetic relationship between similarly related species, but in order to generate precise results, the DNA region(s) to sequence must be carefully chosen. Highly conserved regions of the genome such as the molecules associated with the central dogma often serve as excellent molecular fossils. In the problem posed here, we will use the popularly-chosen 16S ribosomal RNA region on mitochondrial DNA.

The seemingly endless array of sequences openly available through various databases make it possible to access sequences of all kinds. With such a vast number of sequences, there is a need to organize them so that they can be easily manipulated, leading to a variety of standard formats. With this homework, you have been given sequence files relevant to the different *Ptychadena* species in a well known format known as FASTA. For this assignment you will have two .txt files provided with the homework. You will see that each sequence in a given file is composed of a line (beginning with a ">" symbol) containing information about the sequence, i.e. the species name, the ID number for obtaining the sequence from a particular database and, as we have provided here, the location of the species. The subsequent lines before the next ">" contain the sequence. We have already aligned the sequences by placing gaps ('-') in each of them, making it easy to compare each sequence directly.

While one of the files contains 16S mitochondrial DNA sequences from 26 different species scattered throughout mainland Africa, the other file contains the sequences of three amphibians of the same species on São Tomé. Because there may be some variation in the sequence of DNA across individuals within the same population, it is often useful to collect samples from multiple individuals of the same species to provide stronger evidence for the relationships of the species with others. In this assignment, you should find that, not surprisingly, the three *Ptychadena newtoni* on São Tomé agree well with each other in their relationships to the *Ptychadena* species across mainland Africa.

**Comparing Frog Sequences.**

**Question 2a:** Using what you learned in the computational tutorial for this week, write a function that directly compares two sequences and assigns a score. There are a variety of scoring systems for comparing sequences, so for this problem, create a system where the score is the number of matches between two sequences divided by the number of positions compared. If at any position, either one of the sequences has a gap '-', ignore that position in the scoring.

Once you have written your function, compare each São Tomé sample's sequence to that of each mainland African species and identify the best three matches, verifying that the three São Tomé samples agree in their top three matches. Locate the regions of Africa of these three frog species.

You should only need BioPython's SeqIO and maybe NumPy's zeros function for this problem. **Refer to Tutorial 1 for additional guidance.**

**Can "Rarely" Over Short Time Scales Lead to "Frequently" Over Long Time Scales?**

One of the pieces in the evolution puzzle is the challenge of trying to make sense of what Alfred Russel Wallace discovered about the distribution of different organisms in both space (biogeography) and time (fossil record). In this part of the problem, we will apply our street-fighting mathematics skills to acquaint ourselves with some of the arguments that have been made for the dispersal hypothesis.

Dispersal biogeography has been pejoratively referred to as "a science of the improbable, the rare, the mysterious, and the miraculous." Our goal in this problem is to see if we agree with that assessment or if George Gaylord Simpson had it right when he argued that people have little intuition for accumulated weight of rare events that play out over very long time scales. Concretely, we will try to estimate how often amphibians would successfully colonize the islands in the Gulf of Guinea.

Here, we advise you make your estimates for the probability of a successful colonization event by using what Sanjoy Mahajan in his great book *Street*


Figure 3: Map of Africa with São Tomé and Príncipe in the red circle.

*Fighting Mathematics* refers to as "divide and conquer". What this means is that you take the overarching problem and then divide it into ever smaller sub problems each of which you can figure out. For example here, we need to figure out how many frogs end up in the Gulf of Guinea from the Congo River. But to know that, we have to in turn figure out how much of the land adjacent to rivers such as the Congo River gets flooded during the biggest flooding events. Then, we might want to estimate the frequency with which trees end up in rivers that might serve as rafts, etc. Useful resources could include the map in Figure 3, Google Maps and Earth Nullschool.

**Question 2b:** Based on your results from the DNA sequences, from which part of Africa would you conclude the *Ptychadena newtoni* originated? If we accept that proposition, let's now try to understand the challenges of such a colonization event. Apply the street-fighting mathematics that you used in the previous problem to see how many groups of amphibians from these parts of mainland Africa will make it to São Tomé over the 13 million years of the island's existence.

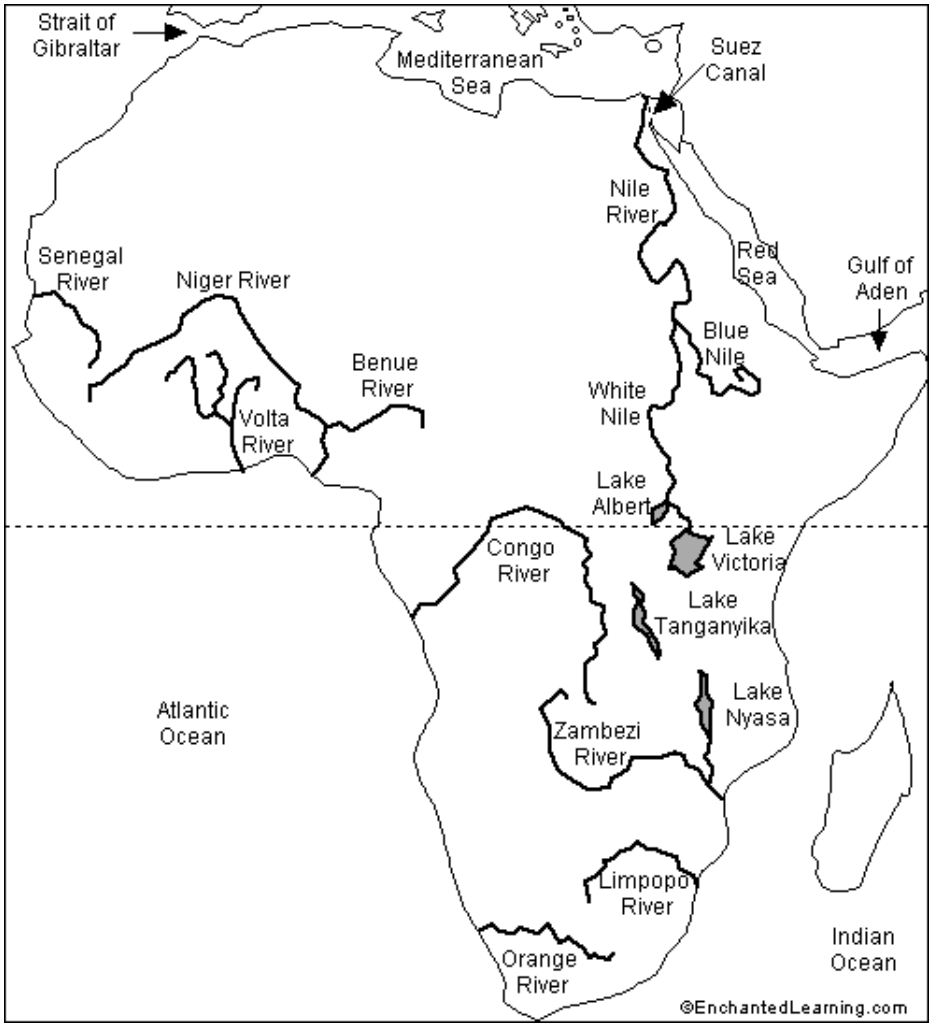


Figure 4: Map of Africa and water sources.