Bi 1: The Great Ideas of Biology
Homework 2
Due Date: Thursday, April 20, 2023

> “Our doubts are traitors and make us lose the good we oft might win by fearing to attempt.”
>
> William Shakespeare, *Measure for Measure*

1. Genomic Data Storage: A Street-Fighting Mathematics Approach

As before, we will exercise our creativity and make estimates. Learning to do simple order-of-magnitude calculations is an incredibly valuable skill. They allow you to check whether your initial ideas about a problem make sense, much like you might check that the units on both sides of an equation match. More importantly, they unlock a hidden superpower latent in all of us: the ability to roughly “figure out” the answers to complex problems using nothing but our brains! We can even start to delay the typical impulse to search the internet for everything, discovering that we often already know most of what we need to start exploring the natural world.

Initially, this style of estimation can feel intimidating. There are often multiple valid approaches to a problem, which might lead you to worry about whether you’re doing things the “right” way. But this is precisely the point: we are not seeking exact answers so much as we are interested in seeing your logical *approach* to a challenge. (This makes it harder for us to grade your homework, and easier for you to get full marks.) Remember to show your work, and to ensure that your solutions are justified with quick sanity checks. If you look something up, cite it; but in general, you will probably not need to. These are *estimates*, not precise answers; hence no long lists of “significant digits”. Be brave.

What's in a Genome?

As we will discuss throughout the term, the genome is like a modern Rosetta Stone, providing clues to the evolutionary history of life on Earth. Beyond our understanding of the central dogma — the process whereby DNA is transcribed into mRNA which is subsequently translated into the many proteins that are the basis of the structure and function of an organism — the genome has many wonderful secrets. These mysteries range from the various regulatory mechanisms that determine when certain genes are expressed, to the exquisite molecular dance that creates an entire *you* from just a single zygote (sperm and egg). As the course continues and we come across more of the genome's secrets, you may eventually come to ask: just how much information is in a genome? We will get a feel for the answer now with an estimate.

To give you some ideas for how to tackle this problem, assume that each printed character is 8 bits (1 byte) while a base pair is 2 bits. Remember that you can come up with four possible unique 2-bit words, corresponding to the four nucleotides, and that, because we are working with DNA in this problem, each nucleotide is complemented with another one, so we assume no additional information is encoded on the opposite strand of DNA. For simple conversions, let 1 kilobyte be 10^3 bytes, etc. Finally, note that the *E. coli* genome is 4.6 million base pairs (across one chromosome) and the human genome is 6.5 billion base pairs (across 23 pairs of chromosomes).

Question 1a

Use street-fighting estimates to determine how much data (in bytes) is stored among all of the books in the Sherman Fairchild library (SFL). Next, make an estimate of the number of bytes in the *E. coli* genome. Finally, estimate the amount of genomic data passed from one human parent to his or her offspring.

Hint: To make the library estimate, you could think about the number of floors in the library, the area of each floor, the number of shelves and books on each floor, the number of words on each page, etc.

Another compelling example of information storage in genomes is that of viruses. For example, influenza virus (responsible for the flu) carries a 14,000 nucleotide genome inside a viral particle only 100 nm in diameter.

Question 1b

How many bytes of information are carried in the influenza genome, and what is the virus's information density in units of bytes/volume? Make an estimate of the information density in a typical computer storage device (such as a thumb drive), and then scale up the biological information storage density of the influenza virus and compare it to modern hardware.

How Much Sequence Information is Available?

You are alive during an incredible moment in biological progress. DNA sequencing technology has improved explosively, far outpacing Moore's Law. The result is a rapidly growing number of new sequences that are read and recorded every day, stored in massive databases around the world. The Sequence Read Archive (SRA), the raw sequence database of the National Institutes of Health, is the source of many of the sequences we'll use in problems and tutorials this term. As of January 2020, it contains over 6 petabytes of publicly-accessible sequence data!

Question 1c

Harry Potter is the best-selling fantasy series of all time, with around 500 million books sold. If you were to treat each base in the SRA as an English character, how many copies of the entire series (that is, treating all seven books as one) could you write? Compare to the number of copies of the whole series sold. *Do not look up the number of words or pages in each book!* Be bold, be brave, and estimate!

2. How Did Frogs Get to São Tomé?

In class, we discussed the fascinating biogeographical story of the frogs of São Tomé. In this problem, we will explore in more detail the way that DNA sequencing was used as a window into the dispersal of these frogs onto their oceanic islands.

São Tomé is an island located 255 km off the west coast of Africa. Volcanic activity formed it roughly 13 million years ago, and continued to shape the landmass until as recently as a hundred thousand years ago. Due to its considerable distance from the African coast, and how recently (geographically speaking) it emerged from beneath the surface of the ocean, the island offers an exciting window into biodiversity due to dispersal. While birds or storms may have carried seeds to the island, the question of how amphibians traversed such great distances is much harder to explain. This is because amphibians do not survive long in salty water. And yet frogs exist on São Tomé! To understand where they came from, we'll turn to sequence analysis.

Comparing Frog Sequences

As illustrated in class, DNA sequence analysis is a powerful tool for determining the phylogenetic relationship between similarly related species. For the results to be useful, however, the regions of the genome compared must be carefully chosen. Here, we'll use part of the mitochondrial genome known as 16S ribosomal RNA. The idea here is that all living organisms have ribosomes, the macromolecular complexes that perform the process of translation (i.e. turning mRNA sequences into corresponding amino acid sequences of proteins). As a result, the ribosome serves as a useful molecular record of the evolutionary history of a given organism.

Your dataset consists of two files. One contains sequences from 26 different species of frog from the genus *Ptychadena* that were collected on mainland Africa. The other contains sequences (from the same part of the genome) from 3 individuals[1] of a single frog species found on São Tomé. By comparing the similarity of these sequences, we'll attempt to draw conclusions about where the island frogs came from.

These files are in a common DNA sequence format known as FASTA. Each sequence starts with a header line (beginning with ">") containing basic metadata, like the species name, or an ID number for obtaining the sequence from a particular database. In our case, we've included the geographic location of the species. The subsequent lines (up until the next ">") contain the sequence. As discussed in tutorial, we have already aligned the sequences by placing gaps ('-') in each one, making it easy to compare them directly.

[1] Because there is some variation in sequence between individuals in a population, it is often useful to sample multiple individuals in order to draw stronger conclusions.

Question 2a

Write a Python function that directly compares two DNA sequences and assigns a similarity score. There are a variety of scoring systems for comparing sequences. For this problem, create a system where the score is equal to 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 a São Tomé frog's sequence to that of each mainland African species and identify the closest three matches. Repeat this process for all three São Tomé samples. They should all have the same top three matches. Locate the regions of Africa where these three matching mainland species originate from.

As always, refer to the tutorial for additional guidance.

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

With the sequence comparison complete, it's time to connect your findings to the greater geographical picture. In lecture, we often emphasize that one of the key goals of the course is to reflect on and examine the great *principles* of biology. An obvious contender for most important principle of all is that of the theory of evolution. Yet for all its successes, some aspects of evolution remain puzzling even today. One challenge is making sense of the distribution of different organisms in both space (biogeography) and time (the fossil record). Here, we'll apply our street-fighting skills to acquaint ourselves with some of the arguments that have been made for an idea known as dispersal.

One dispersal hypothesis for the origin of São Tomé's frogs comes from a highly entertaining 2006 paper. It is well-known that rainstorms can wash logs, brush, and other debris into waterways, forming "natural rafts" that can float for surprising distances. Some of these conglomerations have been known to traverse the Atlantic in as little as two weeks. Much like a dandelion seed gets carried across a field on a stiff breeze, what if frogs from the mainland hitched a ride down the Congo River and eventually made it to São Tomé? The authors perform a careful analysis in support of this theory, but like all good street fighters, our first instinct should be to make an estimate: is this theory even plausible?

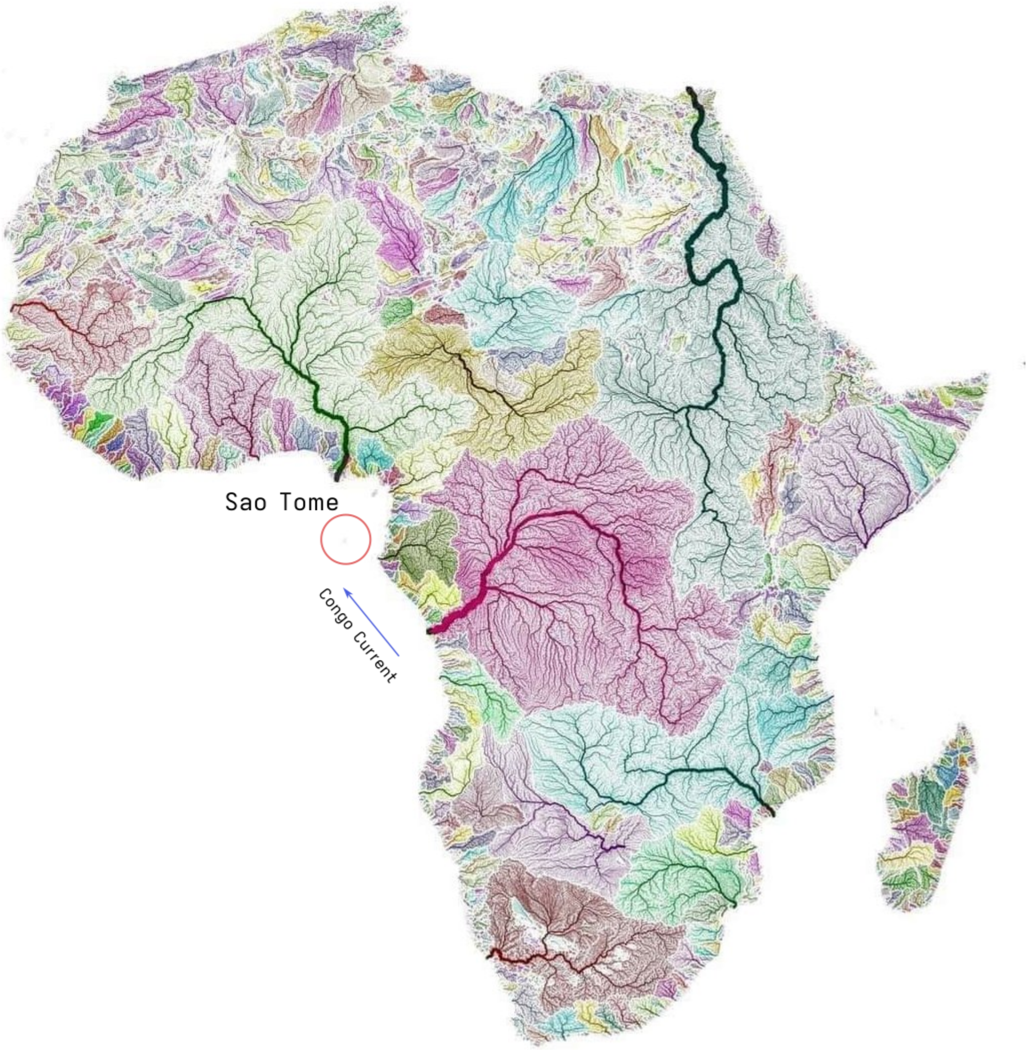

Figure 1: Map of Africa's river basins. The islands of São Tomé and Príncipe are circled in red. The Congo River basin is the prominent magenta one shown in the center. The sea-surface Congo Current is marked by a blue arrow. Adapted from Grasshopper Geography.

After all, dispersal biogeography has been (pejoratively) referred to as "a science of the improbable, the rare, the mysterious, and the miraculous." Our goal, then, is to put that assessment to the test. Concretely, we'll try to estimate how often amphibians could successfully colonize the islands in the Gulf of Guinea, where São Tomé is located, in accordance with the natural raft hypothesis. Perhaps George Gaylord Simpson had it right when he argued that people have poor intuition for the accumulated weight of rare events when time scales are very long.

As always, we advise you to “divide and conquer”. To estimate the probability of a successful colonization event, we need to figure out how many frogs end up in the Gulf of Guinea from the Congo River. To know that, we have to determine how many natural rafts form, and how often; this might depend on the area of land adjacent to the river, the frequency with which that land floods, the number of frogs and trees in that area, etc. Useful resources might include the map in Figure 1, Wikipedia, and AmphibiaWeb’s species distribution mapper.

Question 2b

Based on the results of your sequence analysis, walk us through an estimate of how many groups of amphibians from these parts of mainland Africa have made it to São Tomé over the 13 million years of the island’s existence, in accordance with the natural raft hypothesis.

3. Sequencing in the Time of Cholera

“King Cholera”

Cholera is a vicious infection caused by the bacterium *Vibrio cholerae*. Typically spread by contaminated drinking water, once ingested it colonizes the small intestine and causes severe fluid loss. Throughout history, it has claimed tens of millions of lives and caused seven global pandemics, with the seventh ongoing since 1961. Although modern sanitation has largely eradicated it in the developed world, cholera remains endemic and deadly in other regions. Without medical treatment, mortality is around 50% — a coin flip between life and death.

Shortly after a devastating earthquake in 2010, Haiti experienced its first outbreak of cholera in over 100 years, resulting in over 800,000 cases. But where did the disease suddenly come from? In the weeks following the outbreak, experts around the world struggled to find an answer.

In this problem, we will retrace the story that helped health organizations and scientists determine the source of the outbreak. Much like how John Snow famously traced London's 1854 outbreak to the Broad Street water pump, you will follow the Haitian team's investigation — starting with the first reports of cholera in Mirebalais — to find out where it originated. But unlike Snow, we also have DNA sequencing technology in our arsenal!

The Geographical Spread of the Outbreak

As shown in Figure 2, the investigation team tracked cholera cases up the Artibonite River to the city of Mirebalais, where it branches into several tributaries. Just two miles upstream was a camp of UN peacekeepers who had been dispatched from Nepal to assist in earthquake relief efforts. Could this just be a coincidence? At the time, the rapid salvo of health and safety issues affecting local Haitians was (understandably!) generating significant unrest. Some prominent scientists dismissed claims that the peacekeepers were to blame for the outbreak, instead suggesting that climate conditions had simply released dormant bacteria from the environment. Media outlets and UN officials amplified these claims.

The investigation, however, quickly found that the UN camp had been regularly dumping its raw sewage into an open septic pit, which would overflow into the river after rainfall. All signs now pointed to the peacekeepers.

The DNA Detectives

To test their suspicions, a separate contingent of scientists next turned to DNA sequencing. Their goal was to determine whether the cholera had indeed come from Nepal, by comparing the similarity of the strains infecting people in Haiti to those known to be endemic to Southeast Asia. The researchers obtained isolated samples of *V. cholerae* from patients in Nepal, and sequenced them alongside isolates from Haiti. They also included samples from Bangladesh and Peru, which served as helpful points of comparison in case the Haitian strains had indeed originated elsewhere.[2]

Building a Phylogenetic Tree

With the sequences in hand, how should they go about comparing them? As we know (e.g. from Question 2a), mutations provide a clever basis for these analyses. If two DNA sequences A and B have fewer differences between each other than they do with sequence C, then A and B must be more closely

[2] For an unabridged description of their methodology and analysis, see the original paper.

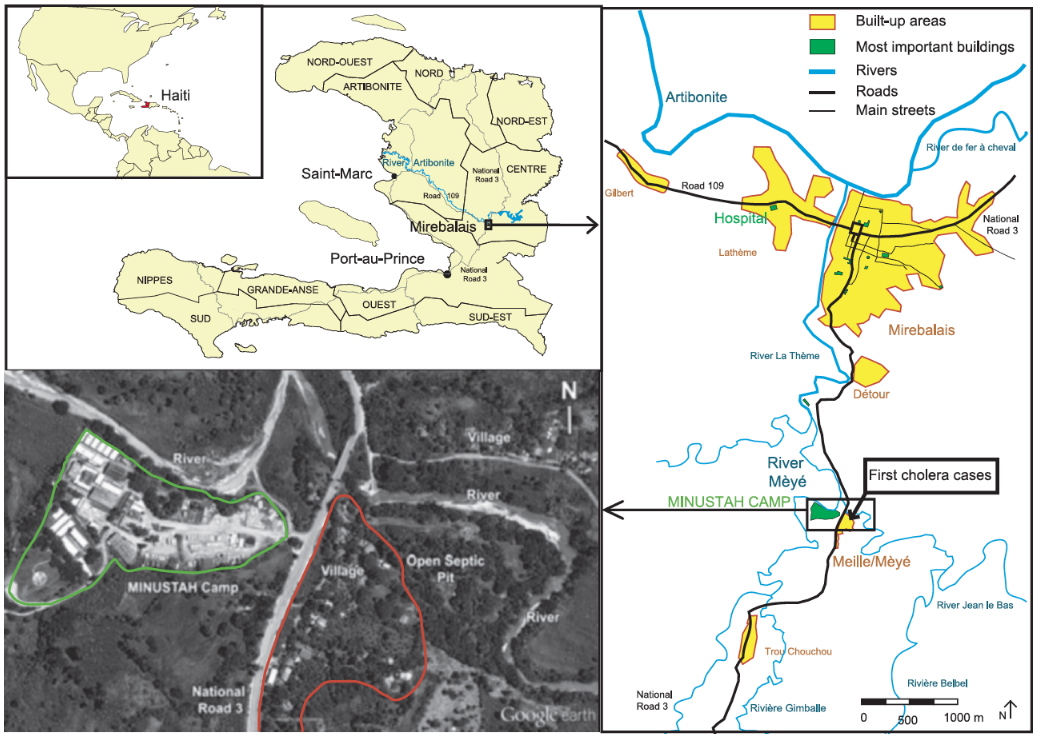

Figure 2: Map of region of Haiti that reported cholera diagnoses. The MINUSTAH camp is shown in green in the right and lower left panels. Note the proximity of the septic pit to the village and tributary.

related to each other than either is to C. The branch of biology concerned with studying these kinds of mutational histories is called *phylogenetics*. We often represent these relationships graphically in a data structure known as a phylogenetic tree, which clusters samples according to their genetic similarity. Figure 3 shows a toy example for the aforementioned case with A, B, and C.

As you might already have guessed, there are many possible trees for a set of samples depending on the criteria you choose for grouping them. For the following exercises, we will judge differences based on *single-nucleotide polymorphisms*, or SNPs. A SNP (pronounced "snip") is just a mutation at a single position in the genome, like an adenine (A) turning into a guanine (G). In practice, we also have to deal with insertions, deletions, and all manner of rearrangements in sequences, but we will ignore those nuances for now. So, if two sequences have fewer SNPs between them than another pair, they are more closely related. (This is a lot like the scoring you did in Question 2a!)

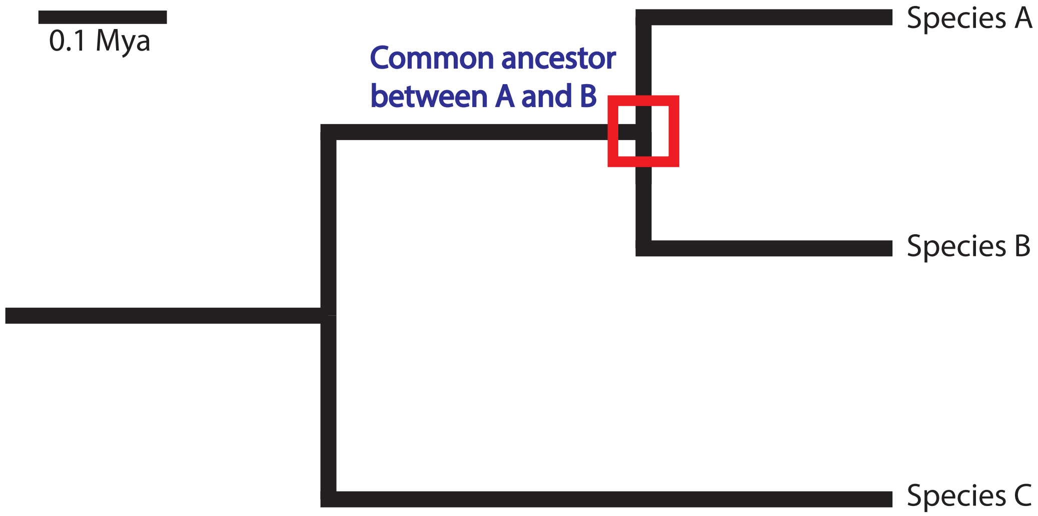

Figure 3: A simple example of a phylogenetic tree for three species. Time is always implicit from the root to the tips (here that's the positive x-direction), up to present day. A and B are more closely related to each other than either is to C. Note that this tree uses branch length to denote time quantitatively; this information couldn't come from sequence data alone, but would require additional time-resolved information (like a fossil record).

Strain	**Date collected**
Haiti	November 2010
Nepal - Isolate 2	July 2010
Nepal - Isolate 13	August 2010
Nepal - Isolate 14	September 2010
Bangladesh	2002

Table 1: Summary of *V. cholerae* strains to be analyzed.

Question 3a

In Table 2, we have determined the number of genome-wide SNPs between each pairwise combination of five cholera strains by analyzing their sequences. With this information, create a phylogenetic tree by hand (i.e. don't use computational tools to do the clustering, although you may choose to draw the result digitally) that represents these relationships. Hence, propose a possible relationship between the Haitian strain and the strains from Southeast Asia. What does this mean for the hypothesis that the Nepalese peacekeepers were the source of the outbreak?

	Haiti	Nepal 2	Nepal 13	Nepal 14	Bangladesh
Haiti	0	84	84	2	31
Nepal 2	84	0	0	71	60
Nepal 13	84	0	0	71	60
Nepal 14	2	71	71	0	33
Bangladesh	31	60	60	33	0

Table 2: Number of SNPs between each strain, compared pairwise. Note the symmetry across the diagonal; the concept of a SNP only makes sense if we define the mutation with respect to some baseline or reference sequence, such as via pairwise comparisons.

Estimating Mutation Rates

An implicit idea in the sequence comparisons we are making here — which was also part of our thinking about São Tomé — is that of the *molecular clock*. The basic concept is that, as organisms with some common ancestor diverge over time, changes accumulate in their genomes at a (mostly) steady rate. In the coming weeks, we will describe a related set of ideas known as the neutral theory of evolution, which argues that the rate at which these mutations become established in a population is equal to the mutation rate itself.

Question 3b

Suppose *V. cholerae* divide every 40 minutes and have a genome size of approximately 4 million base pairs. Make an estimate of the mutation rate of the Haitian strain using the SNPs between this strain and the most closely related strain. Your units (which should be explicitly written in your answer) should be in mutations per base pair in the genome per replication, or mutations per base pair in the genome per generation.