Bi1: The Great Ideas of Biology
Homework 1
Due Date: Thursday, April 13, 2017

“From this proposition it will follow, when arithmetical addition has been defined, that $1 + 1 = 2$.”
- Page 379 of Bertrand Russell and Alfred North Whitehead’s Principia Mathematica.

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

Throughout the term, you will have to exercise your creativity in making estimates. Doing simple order-of-magnitude estimates to see whether your ideas about a given problem make sense can be as powerful as the device of checking that units on both sides of an equation match. When doing the simple estimates required in a number of our assignments, we will try to provide you with enough relevant information so that you do not need to spend a bunch of time browsing online. These values may include the size of a eukaryotic cell, the amount of time a stork can survive without eating food or the age or distance of a volcanic island from the nearest mainland. From these numbers, you should be able to estimate the number of cells in your body, the distance a stork can fly continuously and, as in problem 2, the number of amphibians reaching a far-away island. **In submitting your work, be sure to justify your estimates fully. Remember that you are providing an estimate, not an exact answer (hence no long lists of “significant digits”). We are most interested in seeing your approach.** The idea is to develop creative but logically compelling analyses and estimates on the way to your solution.

How Much Data in the 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 still has many wonderful secrets. These mysteries range from the various regulatory mechanisms that determine when certain genes are expressed to the genome-wide rearrangement mechanisms that allow jawed vertebrates to use a stunningly diverse collection of antibodies to neutralize the vast number of infectious viruses and bacteria that we encounter daily. As the course continues and we run into more of the genome’s secrets, eventually you may come to ask: just how much information is in a genome? We will get a feel for this question with an estimate now.

To give you some ideas for how to tackle the first part of the 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 other strand of DNA).

**Question 1a:** Use street fighting estimates to determine how much data is stored among all of the books in the Millikan library (express your answer in byte units (i.e. Mb, Tb, etc.)). Next, make an estimate of the number of bytes in the *E. coli* genome. Finally, turn to humans and estimate the amount of genomic data passed from one human parent to his or her offspring. If you are stuck, refer to the hints below.

To make the library estimate you don’t need to look anything up. Rather, think about the number of floors in the library, the area of each floor, the number of books on each shelf, the number of pages in a book and the number of words on a page. For simple conversions, a kilobyte is 1000 bytes; a megabyte is 1000 kilobytes; etc. Finally, note that the *E. coli* genome is 4.6 million base pairs and the human genome is 6.5 billion base pairs across 23 pairs of chromosomes. Reminder: do not provide answers with unjustified numbers of significant digits - these are meant to be order of magnitude style estimates.

Another compelling example of information storage in genomes is that of viruses. Influenza virus is amazing! There are so many things to admire about these tiny viruses. One particularly intriguing feature is that each virus
carries within it eight distinct RNA molecules and all of them are needed for a viable infection. The 14,000 nucleotide genome is divided amongst these 8 negative-sense RNA molecules contained within the viral particle roughly 100 nm in diameter.

**Question 1b:** How many bytes of information are carried in the influenza genome and what is the 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 our computer technologies.

DNA sequencing is moving forward at an explosive pace. The exponentially increasing number of sequences obtained from various genomes has resulted in the proliferation of DNA sequence databases across the world. As of February 2017, the Sequence Read Archive (SRA), the raw sequence database of the National Institutes of Health and from which we obtained many of our sequences for the numerous problems and tutorials that you will face this term, contains an astounding 9.8 quadrillion ($10^{15}$) bases in sequences for people to freely access.

**Question 1c:** If you were to treat each base from the SRA as equivalent to a character in a piece of text, estimate the number of copies of the entire Harry Potter series (that is, treating all seven books as one) that could be written with these bases. Compare to the number of copies of the whole series sold (70 million). Do not look up the number of words or the number of pages in each book! This is part of the estimate.

2. How Did Frogs Get to São Tomé?

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

São Tomé is an island located 255 km off the west coast of Africa. Vol-
canic 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 class, 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. 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?**

In class, we argued that one of the key points of the class is to talk about the great principles of biology. Obviously, a contender for most important principle of all is that of the theory of evolution. 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 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 1, 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.

3. Sequencing in the Time of Cholera

**Introduction**

Cholera is an infection where the *Vibrio cholera* bacterium induces drastic water loss in the small intestine, resulting in dehydration in the victim. This bacterium often spreads to communities by contaminated drinking water. Today, cholera remains largely problematic in Southeast Asia and Africa.

Shortly after a devastating earthquake in 2010, Haiti experienced its first cholera outbreak, an epidemic that led to over 400,000 cases. In this problem, we will develop the story that helped health organizations and scientists determine the source of the outbreak. Much in the same way that John Snow traced the 1854 cholera outbreak in London to the water systems, you will follow the Haitian team’s investigation of the first reports of cholera in Mirebalais to figure out where the cholera originated.
The Geographical Spread of the Outbreak

As shown in figure 3, the investigation team tracked the cholera along the Meille River to a camp just two miles south of the city of Mirebalais. This camp consisted of a Nepalese contingent of UN peacekeepers that had disposed of their sewage into the river. Could the cholera have come from
Nepal? As Nepal had previously experienced cholera cases, the investigators thought so. They used sequencing methods to examine how related the cholera found in Haiti was to that from various parts of Southeast Asia.

**Obtaining Sequences**

Next, the team needed to determine which sequences to compare. In order to verify the suspected connection between the Haitian cholera outbreak and the arrival of the Nepalese UN workers, an independent research group obtained isolated samples of *V. cholerae* from various patients in Nepal and sequenced them for comparison with the strains in Haiti. They also included an isolate from Bangladesh, which serves as a meaningful additional point of comparison to see if there happen to be strains of cholera other than those collected in Nepal that may be more closely related to the Haiti strains.

![Figure 3: Map of region of Haiti that reported cholera diagnoses. The MINUSTAH camp is shown in green in the right and lower left panels.](image)
Building the Tree

With the sequences in hand, we want to show that the strain in Haiti is similar to a strain of cholera from one of the samples obtained in Nepal. As we know, mutations provide a clever means for making these comparisons. If species A and B have more similarities in their sequences with each other than they do with species C, then A and B are more closely related to each other than to C.

A common way to visualize the similarities of all of these sequences is with a phylogenetic tree. In these diagrams, individual samples are grouped into different branches based on their genetic similarity. A fork in the branch represents the point of divergence between two groups. In even more complex trees, the branches will have varying lengths to denote a sense of time. Figure 4 shows a way to visualize the relationship between species A, B and C.

As you might imagine, depending on your criteria for grouping based on specific features of species (size, weight, color) or through DNA sequencing, the same set of species can be clumped in different ways, yielding different trees. For this exercise, we will judge differences based on single-nucleotide polymorphisms, or SNPs. A SNP consists of a single nucleotide change in a region of the genome, such as a sequence originally having an adenine (A) nucleotide at a certain position being changed to a guanine (G) nucleotide (so basically, a point mutation). Here, we will ignore instances where a nucleotide has been inserted or deleted from the sequence. In this case, we can draw a phylogenetic tree where fewer SNPs between two sequences denote a closer relation. We will tackle a simpler version of this study.

Table 1: Date when each strain was collected.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Date of Sample Collection (MM/DD/YY)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haiti</td>
<td>~ 11/01/10</td>
</tr>
<tr>
<td>Nepal - Strain 2</td>
<td>07/30/10</td>
</tr>
<tr>
<td>Nepal - Strain 13</td>
<td>08/31/10</td>
</tr>
<tr>
<td>Nepal - Strain 14</td>
<td>08/01/10</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>2002</td>
</tr>
</tbody>
</table>
Table 2: Number of single-nucleotide polymorphisms across the genome of each of the five samples of cholera relative to one another. Note the symmetry across the diagonal.

<table>
<thead>
<tr>
<th></th>
<th>Haiti</th>
<th>Strain 2</th>
<th>Strain 13</th>
<th>Strain 14</th>
<th>Bangladesh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haiti</td>
<td>0</td>
<td>69</td>
<td>69</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>Strain 2</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>71</td>
<td>60</td>
</tr>
<tr>
<td>Strain 13</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>71</td>
<td>60</td>
</tr>
<tr>
<td>Strain 14</td>
<td>2</td>
<td>71</td>
<td>71</td>
<td>0</td>
<td>33</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>31</td>
<td>60</td>
<td>60</td>
<td>33</td>
<td>0</td>
</tr>
</tbody>
</table>

**Question 3a:** In Table 2, we have determined the number of genome-wide SNPs between each pairwise combination of five cholera strains. Aside from one strain being collected in Haiti, three of these strains come from Nepal and one from Bangladesh. With this information, create a phylogenetic tree by hand (i.e. no computational algorithms) that shows a possible relationship between the Haiti cholera strain and the strains in Southeast Asia.

**Figure 4:** A basic outline of a phylogenetic tree. Here, species A is more closely related to species B, as their common ancestor existed approximately 200,000 years ago.
An implicit idea in the sequence comparisons we are making here and in the problem about the biogeography of oceanic islands is that of the “molecular clock”. The idea of the molecular clock is that changes accumulate in genomes over time when organisms with some common ancestor diverge over time. What we actually see are the accumulated changes (known as “substitutions”) in genomes over time. A very deep and interesting set of ideas from the so-called neutral theory of evolution, which we will discuss in more detail in a few weeks, argues that the substitution rate (i.e. the rate at which mutations are fixed in a population) is equal to the mutation rate itself, an assumption we use now to try and understand the relatedness of the different strains of cholera.

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.

4. Biogeography of New Zealand

Introduction

Vicariance and dispersal are the two contrasting means of explaining the colonization of once-connected landmasses that now have significant barriers between them. On the one hand, vicariance happens as a result of a major geographical split between two regions, such as when tectonic plate shifts break up a supercontinent or when a mountain range rises up to divide a region. In contrast, organisms may successfully disperse to new regions generally inaccessible to them with the use of naturally formed rafts or by humans who either intentionally or accidentally transport these organisms to new regions.

In order to determine whether a species came to a region by vicariance or by dispersal, biogeographers must place the approximate time when the
species’ diverged on the geological timeline of the landmasses of interest. If the species diverged more recently than when the landmasses became separated, the species will likely have reached their current homes by dispersal. If the species diverged from its sister before the landmasses became separate, a compelling case can be made for vicariance as the means of species diversity.

The biogeography of Australia and New Zealand serves as a wonderful case study for examining the differences between vicariance and dispersal. In particular, we will take part in the discussion over whether the southern beech trees *Nothofagus* arrived in Australia and New Zealand by vicariance or by dispersal. *Nothofagus* has a rich fossil record placing it well back to the time when New Zealand and Australia were both a part of the Gondwana supercontinent, but it was not until DNA sequencing analysis of the southern beech tree and its subsequent placement on the geological timeline that a strong argument for one or the other mechanisms of establishment of biodiversity can be made. To come to our conclusion, we will need to determine two things: (1) the geographical history of the islands in Oceania and (2) the sequence relatedness between trees on the islands.

**Understanding Cladograms**

So far we have seen phylogenetic trees which show the relatedness between different sequences, but we can also use trees to display the geographical history of landmasses over time. Recall that all the present-day continents were once all physically connected as one supercontinent, Gondwanaland. Over time tectonic plates have moved, causing Gondwanaland to separate into many of the continents and islands with which we are familiar today. This process of landmasses splitting can be displayed in a cladogram, where each branch on the tree corresponds to one landmass splitting into two.

**Question 4a:** Based on the cladogram shown in figure 5, state the order of events that led to the present-day islands in Oceania. That is, which landmasses split from which, at what time, and in what order?

**Establishing a vicariance-only hypothesis**

If *Nothofagus* has diverged purely from vicariance events, we can make predictions of how related trees on the various islands should be. For the
purposes of our estimate, we will use a mutation rate of 7 mutations per billion base pairs per generation (a rate from the well-studied plant *Arabidopsis thaliana*), a generation time of 20 years, and a genome length of 500 million base pairs. Note that we are making a lot of assumptions here: we are only considering basepair substitution and no larger scale rearrangements of the genetic sequence, and we are assuming that the rate of mutations has been consistent across time and across the genome. Despite these limitations, we can use this estimate as a first pass for how similar we expect the two species to be given a vicariance hypothesis.

**Question 4b:** Given the timing of landmass splitting you determined in 4a, determine how many mutations you expect to have accumulated between the New Zealand and Australia *Nothofagus* species if they did in fact diverge as a result of vicariance.

**Testing our hypothesis**

We can now test our vicariance hypothesis by comparing the ”true” number of mutations between the two *Nothofagus* trees to the prediction you obtained in 4b.
Question 4c: If we go out and sequence the whole genomes of *Nothofagus* on both New Zealand and Australia and conclude that there are a total of 9 million mutations between them, how long ago would you conclude that they diverged? Once again, use the numbers provided in 4b to answer this question.

What do you conclude about the vicariance vs. dispersal hypotheses in this case?

Figure 6: Map of Oceania.