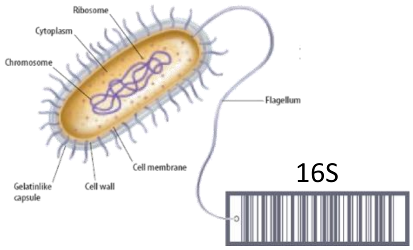
Homework assignment #6-1

Due in class on Wednesday 10/17

An introduction to 16S sequencing (10 points)

Jenny collected a sample. Looking at it through a microscope reveals some cool-looking bacteria. But which bacteria are these? And is this a single species, or are there several that look similar?

To find out, Jenny applies a technique called "16S sequencing". All bacteria carry a certain gene (called "16S ribosomal subunit"). Parts of it are highly conserved, but other parts are very variable, and can be used to identify different species. It's as if each bacterium carried a unique barcode!



Briefly, the method consists in:
1. extracting all DNA from a sample;
2. selectively amplifying just the 16S gene, discarding the rest;
3. sequencing the amplified DNA, e.g. with an Illumina sequencer (the most common platform used for 16S today);
4. identifying the bacteria by comparing the results to databases.

The sequencing results are in, and Jenny asks for your help!

The file is in FASTQ format (https://en.wikipedia.org/wiki/FASTQ_format). This will typically contain a lot of sequences, but for this homework, we will use a demo file with only ten.

1. [Warm-up] In our example, the format is Illumina 1.8. Download the data file `demo.fastq`. To open it, use any plain-text editor. You may change extension to *.txt if it makes life easier. Use a **`fixed-width font`**, so the sequences and the matching quality scores are aligned on the screen. A few sequences look noisier than others; which ones?

2. Consider the sequence `Seq_01`. Let's find out which bacterium we're dealing with! For this, we will use a tool named *Basic Local Alignment Search Tool* (BLAST) to search for similar sequences in the NCBI database (National Center for Biotechnology Information).
   a. Navigate to https://blast.ncbi.nlm.nih.gov/Blast.cgi.
   b. Copy the sequence `Seq_01` into "Enter query sequence" box.
   c. Which bacterium is it?

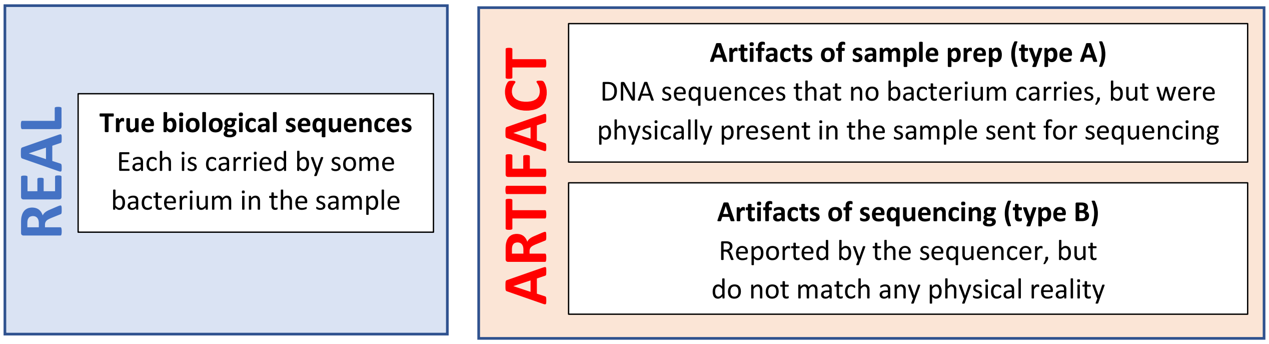
3. All 10 sequences look similar, but there are some subtle differences. Taking `Seq_01` as reference, identify all differences. Use your favorite text editor to highlight in color the differing nucleotides, and the matching quality scores. (Again, make sure you're using a fixed-width font.)

Do these differences ("substitution errors") reflect true biological variability, or are they just noise? How can we tell? We will now try to make and justify a guess as to which of these differences are, in fact, "real".

4. Drawing on the assigned reading (or other resources), complete the sentence:

   The two major sources of substitution errors in 16S data are: the Illumina sequencer itself, and a molecular machine called ________, used in a reaction called ________, to selectively amplify the 16S-encoding DNA prior to its sequencing.

5. Note that these two types of errors have an important difference. Either will lead to erroneous sequences in the output file, but in one case these are completely fictitious, and in the other – the corresponding DNA sequence did in fact exist in the sample sent for sequencing:



REAL

True biological sequences
Each is carried by some bacterium in the sample

ARTIFACT

Artifacts of sample prep (type A)
DNA sequences that no bacterium carries, but were physically present in the sample sent for sequencing

Artifacts of sequencing (type B)
Reported by the sequencer, but do not match any physical reality

   A low quality score in the sequence file can help spot only one of these error types. Which one?

6. Going back to the demo file, and using the result of question 3, suggest a plausible assignment of those 10 sequences to the three categories above: “real biological”, “artifact of sample prep”, “artifact of sequencing”. Describe your reasoning. Can this classification be made with certainty? Are there possibilities you cannot eliminate?

7. Jenny does not like uncertainty. If she had a “technical replicate” (derived from the same sample, but prepared and sequenced independently), would that help? What outcome would confirm or refute the classification you proposed in 6?

Some advice:

- To avoid doing question 3 manually, use your favorite programming environment to help you (Matlab, Python etc.).
- Both Matlab and Python have dedicated packages / toolboxes to work with sequencing data. **But I am not asking you to learn them or use them – in fact, I would prefer you didn’t.** This assignment does not require anything more than basic string comparison.
- The Wikipedia page on FASTQ format has a lot of information. For this assignment, you will most likely need only the basic information about the format, plus sections 2.1 and 2.2 describing “quality scores” – what they are, and how they are encoded.