NAME ________________________________

BIO/CMPSC 300
Chapter 9 Genome Annotation I
Fall 2017

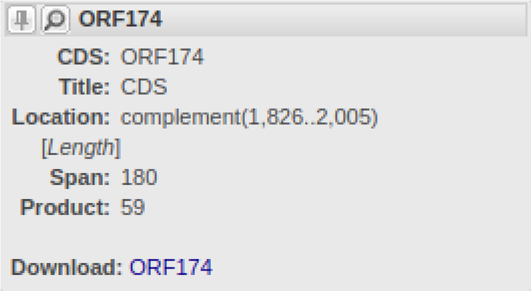
1. Download the *Enterococcus faecium* resistance plasmid sequence file from the shared repository.

2. Use the ORF Finder tool at NCBI (https://www.ncbi.nlm.nih.gov/orffinder/) to annotate the plasmid.

3. **Briefly** (3 sentences) describe what the ORF Finder tool does.

4. What does changing the "Minimal ORF length" option do?

5. When we click on the ORF174 (for example) we get the box that is displayed in Figure 1. Describe is the CDS, Title Location and Product fields contain?



ORF174

CDS: ORF174
Title: CDS
Location: complement(1,826..2,005)
[Length]
Span: 180
Product: 59

Download: ORF174

Figure 1: After alicking on the ORF174 annotation, we see this box open.

6. Below is a generic pattern matching algorithm. Explain specifically how each step of the algorithm is being used by the NCBI ORF Finder to find ORFs in the *Enterococcus faecium* resistance plasmid sequence file.

Pattern-Matching Algorithm

a) Initialize parameters of algorithm:
`pattern`: search pattern = ?

`searchedText:` text that will be searched for pattern = ?

`start:` start location of search (assumes first character is position 1)= ?

`stop:` stop location of search (this represents last location to search from) = ?

`increment` = incrementing value (negative number for upstream search, positive number for downstream search) =?

`threshold` = minimum percentage match required =?

b) Compare `pattern` to characters of `searchedText` starting at position `start`. If percentage of matching characters is `>=threshold`, output `start` position and end algorithm. If not, add `increment` to `start` and continue to step 3. In your own words, what is this step doing?

c) If `increment` is positive and `start` is `<=stop`, repeat step 2. If not, pattern was not found, end algorithm. In your own words, what is this step doing?

5. Once a start codon is found, how could you modify the algorithm above to find an open reading frame beginning with an identified start codon and ending with a stop codon? Hint: the modification involves changing just two parameters.

6. The algorithm above would find an ATG start codon in one of three reading frames by reading a sequence in the 5' to 3' direction, but really we should consider all *six* possible reading frames: three from the DNA as it was entered and three more on the complementary strand. What changes need to be made to the algorithm above to search for ORFs in the complementary strand?

Example of all six reading frames:

5' -TGTCATAGGATAAGCACC -3'

1. TGT CAT AGG ATA AGC ACC
2. GTC ATA GGA TAA GCA
3. TCA TAG GAT AAG CAC

4. GGT GCT TAT CCT ATG ACA
5. GTG CTT ATC CTA TGA
6. TGC TTA TCC TAT GAC