**NAME: _____**

**BIO/CMPSC 300**
**Activity 3**
**Chapter 9 Genome Annotation I**
**Fall 2019**

1. Locate the *Enterococcus faecium* resistance plasmid sequence file included fasta file.
   Note: The fasta file (*Enterococcus_faecium_resistance_plasmid.fasta*) has been included in the directory of this activity file.

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?

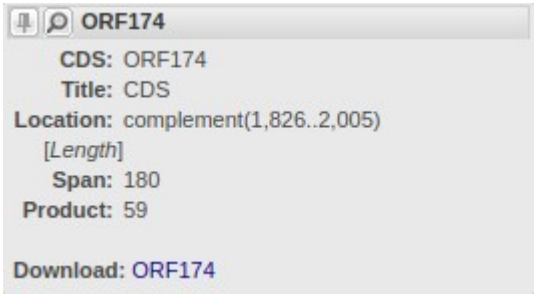


*Figure 1: After clicking on the ORF174 annotation, we see this box open.*

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:

- **`Organism`**: What type of organism are we studying here?
- **`Pattern`**: What search pattern do we use?
- **`SearchedText:`** What type of genome to be searched for patterns?
- **`Start:`** Where is the start location for our search (assumes first character is position 1)
- **`Stop:`** What position is the stop (last) location of our sequence in which to search for patterns? (this represents last location to search from)
- **`Increment:`** What is the incrementing value to be used to jump to a new reading frame? (Hint: a negative number for an upstream search and a positive number for a downstream search)
- **`Threshold:`** What is the 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?

7. For the below sequence, five of the six reading frames have been listed. What is the sixth reading frame that has been omitted? Explain why this is a reading frame.

```
5'  -TGTCATAGGATAAGCACC -3'

  1.TGTCATAGGATAAGCACC
  2. GTCATAGGATAAGCA
  3.  TCATAGGATAAGCAC

  4.GGTGCTTATCCTATGACA
  5. GTGCTTATCCTATGA
  6.   ?
```