Lab 3: Multiple Sequence Alignment
Due: Monday, February 24th – at the start of class

The purpose of this lab is to use the Clustal Omega multiple sequence alignment tool to compare three or more sequences. This lab is to be done individually.

Clustal Omega: Multiple Sequence Alignment Algorithm

Since using a dynamic programming approach to multiple sequence alignment would look like an extension of Needleman-Wunsch and would therefore have a run-time equal to $O(n^s)$, where s is the number of sequences being compared, and n is the length of each sequence, we instead use heuristics to manage the complexity of the multiple sequence alignment problem.

Clustal Omega (used to be called ClustalW) is one of the most popular multiple sequence alignment tools; it uses a progressive alignment algorithm in which the order of adding new sequences to the alignment is determined by first calculating a rough phylogenetic tree called a **guide tree** (see Figure 1B below). The guide tree is generated by first doing pairwise alignments and then using the score or percent similarity from those alignments to draw a tree showing which sequences are more or less closely related. Starting with the two most closely related sequences (*B. xylanisolvens* and *B. fragilis* in Figure 1), Clustal then does global, pairwise alignments to align each new sequence with those already aligned, in order of decreasing relatedness. Note that although this is an efficient way to produce a multiple alignment, the fact that it is based on global alignment means Clustal may not correctly align sequences that share regions of similarity if the sequences are not very similar overall.

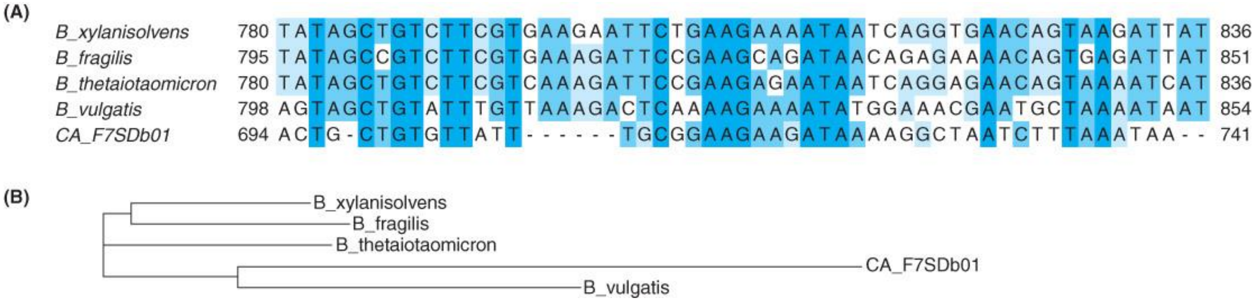
(A)
B_xylanisolvens 780 TATAGCTGTCTTCGTGAAGAATTCTGAAGAAAATAATCAGGTGAACAGTAAGATTAT 836
B_fragilis 795 TATAGCCGTCTTCGTGAAAGATTCCGAAGCAGATAACAGAGAAAACAGTGAGATTAT 851
B_thetaiotaomicron 780 TATAGCTGTCTTCGTCAAAGATTCCGAAGAGAATAATCAGGAGAACAGTAAAATCAT 836
B_vulgatis 798 AGTAGCTGTATTTGTTAAAGACTCAAAAGAAAATATGGAAACGAATGCTAAAATAAT 854
CA_F7SDb01 694 ACTG-CTGTGTTATT------TGCGGAAGAAGATAAAAGGCTAATCTTTAAATAA-- 741
(B)
B_xylanisolvens
B_fragilis
B_thetaiotaomicron
CA_F7SDb01
B_vulgatis

Figure 1: (A) Segment of a multiple sequence alignment for the coding region of a penicillin-resistance gene from five different species. Darker shading indicates nucleotides that are conserved among more of the five sequences. (B) Guide tree used by ClustalW to produce this multiple alignment. [Data from EBI ClustalW][Ref: Exploring Bioinformatics, Figure 4.3]

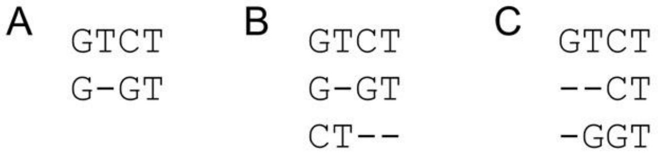

Figure 2: Multiple sequence alignment is complex because the order of adding sequences to the alignment can affect the alignment results.[Ref: Exploring Bioinformatics, Figure 4.4]

What to do:

1. Download the file ermBdiverseSequencesForClustalanalysis.txt from Moodle (Lab 3 files). This file contains erythromycin-resistance (ermB) gene sequences taken from 9 different bacteria.

2. Go to http://www.ebi.ac.uk/Tools/msa/clustalo/ - change PROTEIN to DNA in the drop-down box. Choose to upload a file and use the file you downloaded from Moodle. Leave OUTPUT FORMAT as Clustal with character counts. Click Submit.

3. Answer the following questions:
   a. The output is on multiple tabs. Use the Alignments tab to answer the following.
   Understanding the Results:
   The numbers listed after each block of results are the base pair indices of the last nucleotide in the sequence. The stars under the letters are only there if every sequence had the exact same nucleotide aligned at that base pair location.
      1. How many nucleotides were in each sequence you aligned?

      2. How many differences exist between the sequences you aligned?

      3. What kinds of differences can you see among these genes? Do substitutions (mismatches) outnumber indels (gaps), or vice-versa?

   b. Use the Phylogenetic tree tab to answer the following. Which *ermB*-like genes are the most similar? Which are less similar? Draw the phylogenetic tree created by Clustal Omega.

4. Open the ermBdiverseSequencesForClustalanalysis.txt file you downloaded from Moodle and take a look at the format of it. It contains multiple FASTA files in one .txt file. Create your own input text file with at least 4 sequences in it; each sequence should have at least 14 nucleotides in it. Write down your sequences below. Which sequences do you think are most similar?

5. Run Clustal Omega on your new input file. Does it agree with you in terms of which sequences are most similar? Write down your results below.