University of Pennsylvania
BIOL4536 Fall 2023

HW#6
(BLAST)

Assigned October 11th
Due October 18th, 3:30pm

Question (1) Go to the NCBI BLAST page (`https://blast.ncbi.nlm.nih.gov/`) and select "Nucleotide BLAST" and under "Program Selection" choose "blastn" (should be the third one). We learned in class that you can align DNA with a scoring scheme that scores mismatches between pyrimadines differently from mismatches between purines. Expand the "Algorithm parameters" section and examine the "Scoring parameters". Is it possible to configure BLAST to score mismatches between pyrimadines differently from mismatches between purines? If so, explain how.

Question (2) Next we'll look at an interesting example where (upwards of) every single amino acid has changed, yet there's still a significant alignment.

Use Protein BLAST (configured as described below) to identify the following protein sequence.

```
>Query
GNVDRSYMEDTMERDASWRRHFHHGMLHMNTVMRRVVRQDRASKYPHQAYVENMGHDDMD
```

NCBI BLAST can be found here: `https://blast.ncbi.nlm.nih.gov/`
Under "Database" choose the "Reference Select proteins" and under "Program Selection" choose "blastp" (see Figure 1). You will need to play around with the substitution matrix to find a hit (see Figure 2). What matrix works best? Based on the matrix that works, what do you conclude about the evolutionary distance between this sequence and its closest homologs in the database. What gene is it? What species is the closest hit? Show the best alignment. How many amino acids are unchanged? How many are "positives"? What does "positive" mean here? What's the *E*-value?

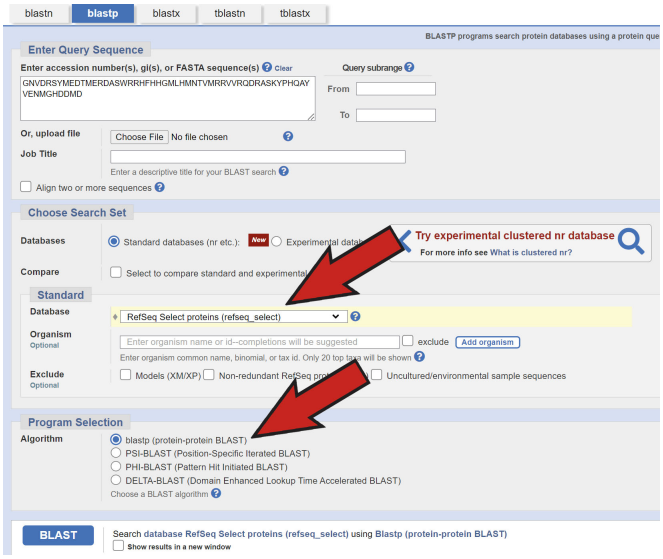

Figure 1

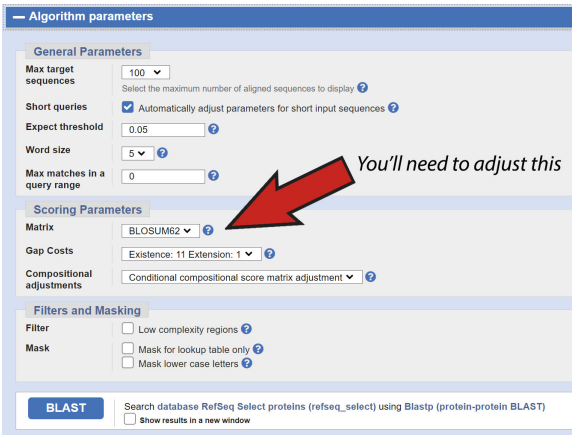

Figure 2

Question (3) The following page has a "contig" of 13,728 bases of the genome of an unknown microorganism sequenced from a sample of Mediterranean seawater.

`https://www.ncbi.nlm.nih.gov/nuccore/MIZB01000007.1`

First we will use BLAST to annotate this raw DNA sequence with protein coding genes. Click "Run BLAST" on the ncbi page (it's on the right side). Select the "blastx" tab, which will translate the sequence into protein in all (six) possible ways.
Under "Database" select "Model Organisms (landmark)". This database is relatively small and non-redundant from a wide range of taxonomies, thus the search is quick and the result will be concise.
Under "Organism" start to type "archaea" and select "Archaea (taxid:2157)".
Hit "BLAST" and wait.
How many results were returned? How many species are involved in the hits?
Click on "Graphic Summary". Supply a screen shot of the graphic. Hover over the colored bars to figure out how many genes there appear to be on this contig.

Go back and run BLAST again but this time change database to "Non-redundant protein sequences (nr)" (again restrict to Archaea). Does this reveal another possible gene on the contig? If so, what's the gene's name?

Question (4) Metagenomics. An organism's gut and other locations contain billions of microorganisms. This is known as the organism's microbiome. The microbiome is investigated by sequencing a variable stretch of the ribosomal gene 16S. From these sequences the species can be determined by BLAST¿ The following is the piece of 16S from an unidentified organism. Paste it into the search box of the "blastn" page. Under "Database" select (the radio button) "rRNA/ITS databases". Make sure "16S ribosomal RNA sequences (Bacteria and Archaea)" is selected from the pulldown menu. Make sure the species box is left empty and "blastn" is selected under "Program Selection". Hit "BLAST". (See Figure 3). Based on the top hit, what is the species?

TTTGATCCTGGCTCAGGATGAACGCTGGCGGTCGGCCTAACACATGCAAGTCGAACGCTCCCCTCGGGGA
GAGTGGCGGACGGGTGAGTAACGCGTGAGAATCTACCTTCAGGTCTGGGACAACCACTGGAAACGGTGGC
TAATACCGGATGTGCCTACGGGTGAAAGATTTATTGCCTGAAGAAGAGCTCGCGTCTGATTAGCTAGTTG
GTGGGGTAAAAGCCTACCAAGGCGGCGATCAGTAGCTGGTCTGAGAGGACGATCAGCCACACTGGGACTG
AGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATTTTCCGCAATGGGCGAAAGCCTGACGGAG
CCAGACCGCGTGAGGGAGGAAGGCCCTTGGGTTGTAAACCTCTTTTGTCAGGGAAGAAAAAAATGACGGT
ACCTGACGAATCAGCCTCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGAGGCAAGCGTTATCC
GGAATTATTGGGCGTAAAGCGTCCGCAGGTGGCTCTTCAAGTCTGCTGTCAAATCCGGTAGCTCAACTAC
CGTCCGGCAGTGGAAACTGAAAAGCTAGAGAGTCGTAGGGGTAGAGGGAATTCCCGGTGTAGCGGTGAAA
TGCGTAGAGATCGGGAAGAACATCGGTGGCGAAGGCGCTCTACTGGACGACATCTGACACTCAGGGACGA
AAGCTAGGGGAGCGAATGGGATTAGATACCCCAGTAGTCCTAGCTGTAAACGATGGATACTAGGTGTAGC
TTGTATCGACCCGAGCTGTGCCGAAGCTAACGCGTTAAGTATCCCGCCTGGGGAGTACGCACGCAAGTGT
GAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGTATGTGGTTTAATTCGATGCAACGCGAA
GAACCTTACCAGGGCTTGACATGTCGCGAATCTTGATGAAAGTTGAGAGTGCCTTCGGGAGCGCGAACAC
AGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCT
CGTTTTTAGTTGCCAATATTAAGTTAGGCACTCTAGAGAGACTGCCGGTGACAAACCGGAGGAAGGTGGG
GATGACGTCAAGTCAGCATGCCCCTTACGTCCTGGGCTACACACGTACTACAATGGGGGGGACAGAGGGT
CGCTAGCTCGCGAGAGTCTGCTAATCCCAAAAACCTCTCCTCAGTTCAGATTGCAGGCTGCAACTCGCCT
GCATGAAGGAGGAATCGCTAGTAATCGCCGGTCAGCATACGGCGGTGAATCCGTTCCCGGGCCTTGTACA
CACCGCCCGTCACACCATGGAAGCTGGCCACGCCCGAAGTCGTTACCCTAACCCGCAAGGGAGGGGGATG
CCGAAGGCAGGGTTGGTGACTGGGGTGAAGTCGTAACAAGGTAGCCGTACCGGAAGGTGTGGCTGGATCA
CCTCC

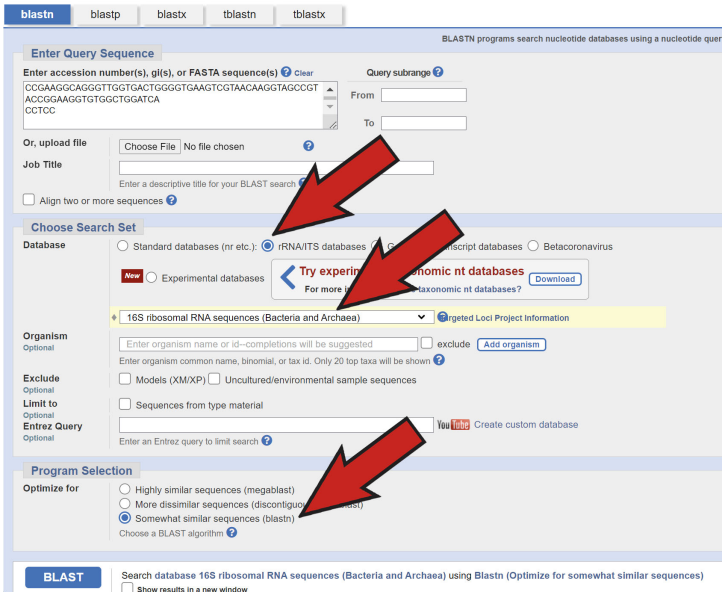

Figure 3

Question (5) Suppose we want to create a drug that targets the COVID spike protein. We want to make sure it's not similar to any human genes to avoid adverse side-effects. Select COVID from the genome browser (see Figure 4). The spike protein (Figure 5) has the following coordinates:
`NC_045512v2:21,563-25,384`
BLAST the DNA of this gene against human genes to see if there is anything similar. Use blastX with PAM250 and set the "Expect threshold" to be 100 (Figure 6) and configured as in Figure 7. How many genes are returned (give a screen shot of the Descriptions). What's the *E*-value of the top hit? Show the actual alignment.

Now go back and get the protein sequence of this gene and do it again with blastp. To get the protein click on the gene in the genome browser to pull up its info page then click on Protein Product, that will take you to its genbank page. From there select the "fastA" link to get the sequence in the proper format. Use again PAM250 and Expect threshold of 100. Configure as in Figure 8.
What is the top hit and *E*-value now? Draw a conclusion - in other words, should we worry when targeting the spike protein about off target side-effects due to homology to human proteins?

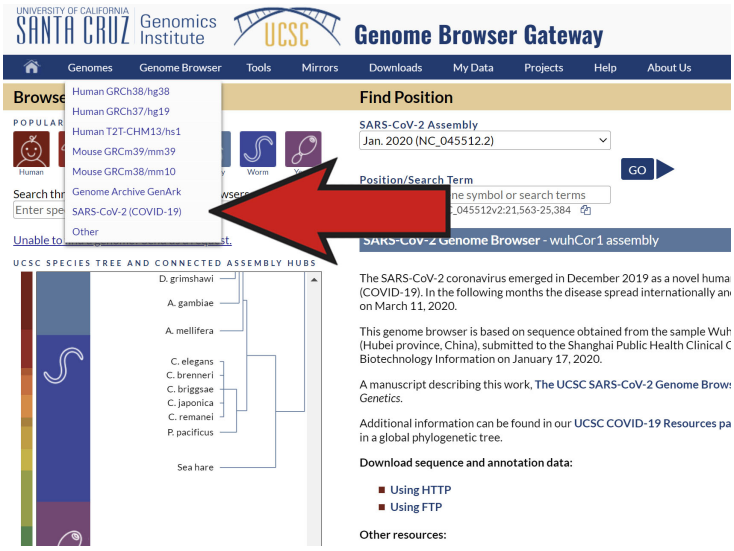

Figure 4

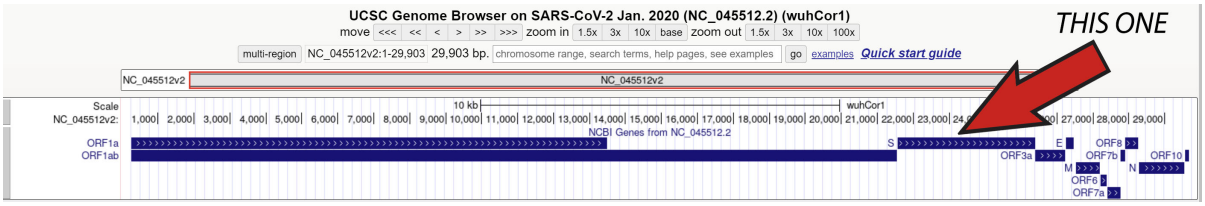

Figure 5

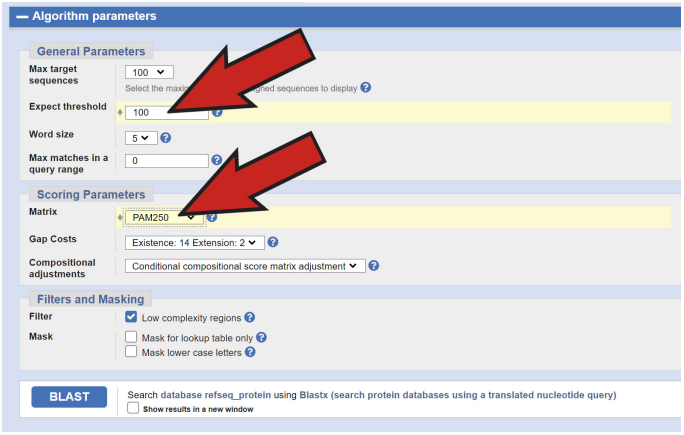

Figure 6

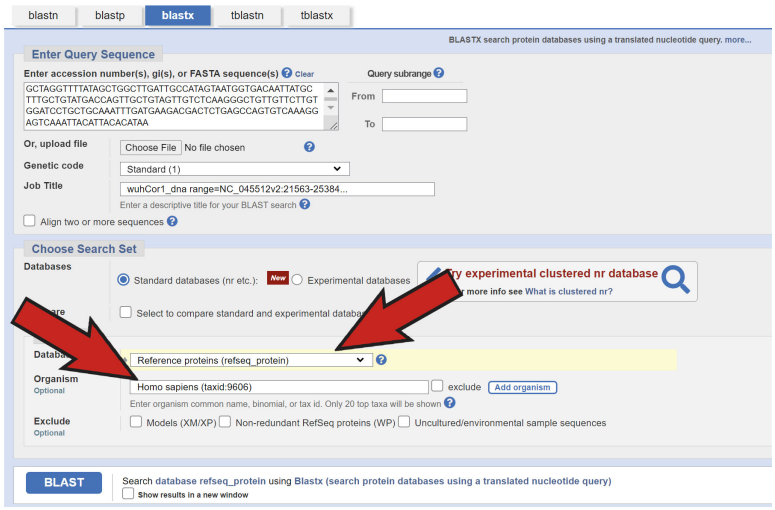

Figure 7

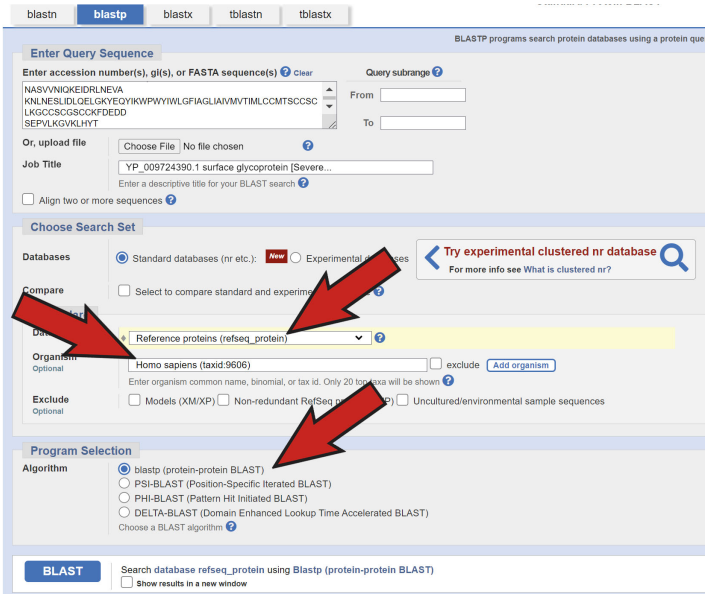

Figure 8