

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

GNVDRSYMEDTMERDASWRRHFHGHMLHMNTVMRRVVRQDRASKYPHQAYVENMGHDDMD

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?

The screenshot shows the NCBI BLAST web interface. At the top, there are tabs for different BLAST programs: blastn, blastp, blastx, tblastn, and tblastx. The 'blastp' tab is selected. Below the tabs, there is a section for 'Enter Query Sequence' with a text input field containing the protein sequence: GNVDRSYMEDTMERDASWRRHFHGHMLHMNTVMRRVVRQDRASKYPHQAYVENMGHDDMD. There are also fields for 'Or, upload file' and 'Job Title'. Below this is the 'Choose Search Set' section, which includes 'Databases' (Standard databases selected), 'Compare' (Select to compare standard and experimental), and 'Standard' (Database: RefSeq Select proteins (refseq_select)). The 'Program Selection' section shows 'Algorithm' (blastp (protein-protein BLAST) selected). At the bottom, there is a 'BLAST' button and a checkbox for 'Show results in a new window'. Two red arrows point to the 'Database' dropdown and the 'blastp' algorithm selection.

Figure 1

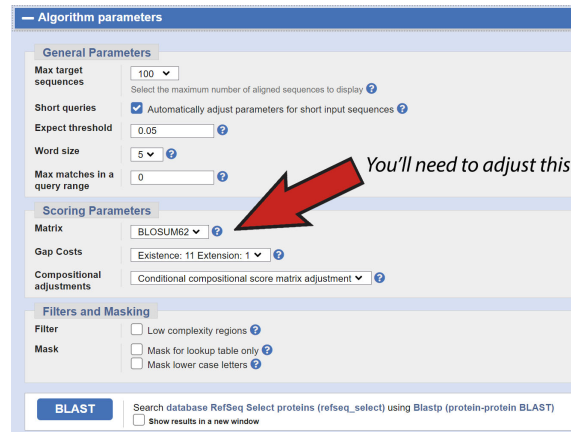


Figure 2

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?

```
TTTGATCTGGCTCAGGATGAACGCTGGCGGTGGCCTAACACATGCAAGTCGAACGCTCCCCTCGGGGA
GAGTGGCGGACGGGTGAGTAACGCGTGAGAATCTACCTTCAGGTCTGGGACAACCACTGGAAACGGTGGC
TAATACCGGATGTGCCTACGGGTGAAAGATTTATTGCTGAAGAAGAGCTCGCGTCTGATTAGCTAGTTG
GTGGGTAAAAGCCTACCAAGGCGGCGATCAGTAGTGGTCTGAGAGGACGATCAGCCACTGGGACTG
AGACACGGCCAGACTCCTACGGGAGGCAGCAGTGGGGAATTTCCGCAATGGGCGAAAAGCCTGACGGAG
CCAGACCGCTGAGGGAGGAAGGCCCTTGGGTTGAAAACCTCTTTTGTGAGGGAAGAAAAAATGACGGT
ACCTGACGAATCAGCCTCGGCTAACTCCGTGCCAGCAGCCGCGTAATACGGAGGAGCAAGCGTTATCC
GGAATTATTGGGCGTAAAGCGTCCGACAGTGGCTCTCAAGTCTGCTGTCAAATCCGGTAGCTCAACTAC
CGTCCGGCAGTGGAACTGAAAAGCTAGAGAGTCGTAGGGGTAGAGGGAATCCCGGTGTAGCGGTGAAA
TGCCTAGAGATCGGGAAGAACATCGGTGGCGAAGGCGCTCTACTGGACGACATCTGACACTCAGGGACGA
AAGCTAGGGGAGCGAATGGGATTAGATACCCAGTAGTCTAGCTGTAAACGATGGATACTAGGTGTAGC
TTGTATCGACCCGAGCTGTGCCGAAGCTAACCGTTAAGTATCCGCGCTGGGAGTACGCACGCAAGTGT
GAAACTCAAAGGAATTGACGGGGCCCGCACAAAGCGGTGGAGTATGTGGTTTAAATTCGATGCAACGCGAA
GAACCTTACCAGGCTTGACATGTCGCAATCTTGATGAAAAGTTGAGAGTGCCTTCGGGAGCGCAACAC
AGGTGGTGCATGGCTGTCGTCAGCTCGTGTGCTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCT
CGTTTTTAGTTGCCAATATTAAGTTAGGCACTTAGAGAGACTGCCGGTGACAAACCGGAGGAAGGTGGG
GATGACGTCAAGTCAGCATGCCCTTACGTCTGGGCTACACACGTAACAATGGGGGGGACAGAGGT
CGTAGCTCGCGAGAGTCTGTAATCCCAAAAACCTCTCTCAGTTCAGATTGCAGGCTGCAACTCGCCT
GCATGAAGGAGGAATCGCTAGTAATCGCCGTCAGCATAACGGCGGTGAATCCGTTCCCGGCCTTGTACA
CACCGCCCGTCACACCATGGAAGCTGGCCACGCCGAAGTCGTTACCCTAACCCGCAAGGGAGGGGGATG
CCGAAGGCAGGTTGGTACTGGGTGAAGTCGTAACAAGGTAGCCGTACCGAAGGTGTGGCTGGATCA
CCTCC
```

The screenshot shows the NCBI BLAST search interface. The 'Enter Query Sequence' field contains the 16S ribosomal RNA sequence. The 'Database' section has 'rRNA/ITS databases' selected, and '16S ribosomal RNA sequences (Bacteria and Archaea)' is chosen from the dropdown. The 'Program Selection' section has 'Somewhat similar sequences (blastn)' selected. Red arrows point to the 'rRNA/ITS databases' radio button, the dropdown menu, and the 'blastn' radio button.

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?

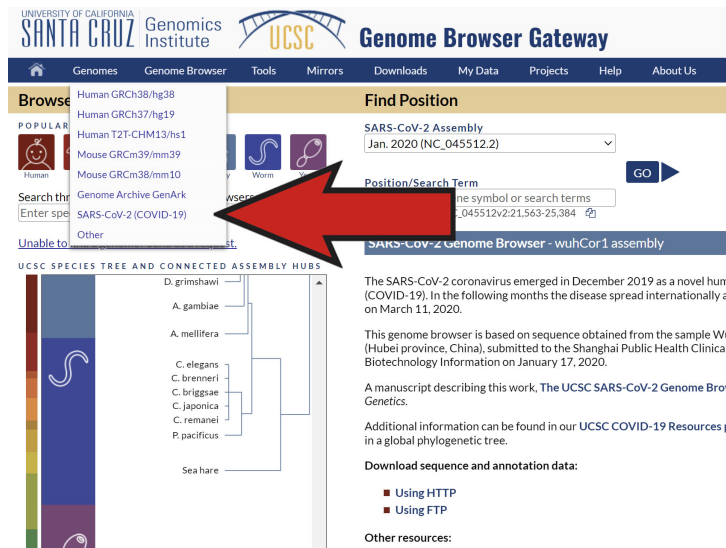


Figure 4

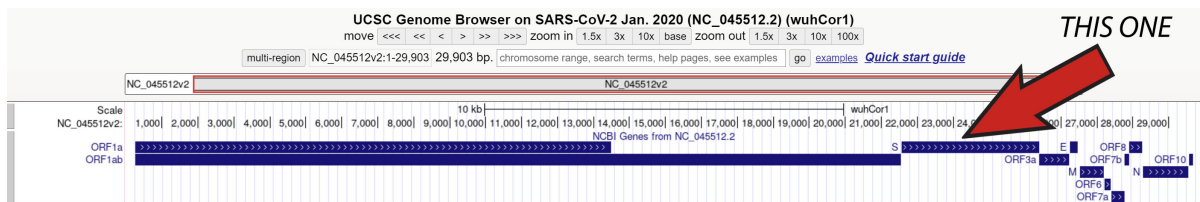


Figure 5

Algorithm parameters

General Parameters

Max target sequences: 100

Expect threshold: 100

Word size: 5

Max matches in a query range: 0

Scoring Parameters

Matrix: PAM250

Gap Costs: Existence: 14 Extension: 2

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only
 Mask lower case letters

BLAST Search database refseq_protein using Blastx (search protein databases using a translated nucleotide query)
 Show results in a new window

Figure 6

blastn blastp **blastx** tblastn tblastx

BLASTX search protein databases using a translated nucleotide query, more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange?

GCTAGT... (FASTA sequence)

Or, upload file: Choose File No file chosen

Genetic code: Standard (1)

Job Title: wuhCor1_dna range=NC_045512v2.1563-25384...

Choose Search Set

Databases: Standard databases (nr etc.) Experimental databases

Database: Reference proteins (refseq_protein)

Organism: Homo sapiens (taxid:9606)

Exclude: Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

BLAST Search database refseq_protein using Blastx (search protein databases using a translated nucleotide query)
 Show results in a new window

Figure 7

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein databases using a protein query

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange?

NASV... (FASTA sequence)

Or, upload file: Choose File No file chosen

Job Title: YP_009724390.1 surface glycoprotein [Severe...]

Choose Search Set

Databases: Standard databases (nr etc.) Experimental databases

Database: Reference proteins (refseq_protein)

Organism: Homo sapiens (taxid:9606)

Exclude: Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Program Selection

Algorithm: blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

BLAST Search database refseq_protein using Blastp (protein-protein BLAST)
 Show results in a new window

Figure 8