**MIP 280A4: Computational Microbiology**

**BLAST, In-class exercise questions**

**Monkeypox virus DNA polymerase**

1. Download into Geneious the Monkeypox virus genome from RefSeq with accession NC\_063383. We will be using this virus’s DNA polymerase gene as a BLAST query. This gene is at position 54,438 -> 51,418 of the genome, and is named OPG071.

Extract this gene sequence from the full-length genome and use it as a query in a nucleotide BLAST search (blastn) on the NCBI website, using default parameters.

Before beginning the search, answer these questions. Hint: expand the Algorithm Parameters section at the bottom of the page. (1 pt each)

* 1. What is the maximum E-value cutoff for results that will be reported?
  2. What scoring scheme will BLAST use for alignments?

Match reward:

Mismatch penalty:

Gap cost:

Now run the search and answer the following questions about the results (1 pt each)

* 1. What is the alignment score (Max Score) of the top scoring hit?
  2. What is the pairwise percent identity of the top scoring alignment?
  3. What is the E-value of the top hit?
  4. What does a BLAST E-value indicate?
  5. What is the accession of the top hit?
  6. Why wasn’t this sequence (NC\_063383) its own top hit?
  7. By default, BLAST shows the 100 top hits. Are there any non-monkeypox virus sequences in the top 100 hits?

1. OK, in the previous question, we learned that you can find a lot of other monkeypox virus sequences in NCBI when you search with a monkeypox query sequence. Because of the recent monkeypox virus outbreak, it is a highly sequenced virus.

But what if we wanted to ask a different question with BLAST: what *other viruses* are most closely related to monkeypox virus? Let’s approach this by restricting results to virus sequences from the RefSeq genomes database.

Do this by clicking the “Edit Search” button on the BLAST results page to return to the search setup screen. Change the database to “RefSeq Genome Database (refseq\_genomes)” and in the Organism field, enter “Viruses”, which should autocomplete to “Viruses (taxid: 10239)”. Then re-run the BLAST search.

Answer the following questions about the blast hits from this new search:

* 1. What is the accession of the top scoring hit now?
  2. How many hits are there now?
  3. What is the most closely related virus to monkeypox virus that’s not monkeypox virus (according to this particular polymerase gene)?
  4. Smallpox is caused by variola virus (which is now eradicated!) What is the pairwise identity between monkeypox virus DNA polymerase and variola virus DNA polymerase?
  5. Given the previous answers does it make sense that vaccines derived from vaccinia virus, originally created for smallpox, are now being used to protect against monkeypox?
  6. What information does the “Query Cover” column provide? Hint: it may help to look at the “Graphic Summary” tab.
  7. Click on the Alignments tab. Select different options from the “Alignment View” tab. What is your favorite alignment view?

**James’s Notebooks**

Perform the paper-based exercise to find a particular phrase given smaller or larger words.

1. What is the impact of an increased word size on database search speed? (1 pt)
2. What is the impact of a larger word size on the ability of a search to find a similar but not exactly identical match? (1 pt)

**Word size and BLAST searches**

1. Repeat the BLASTN search from question #2 above, again restricting results to virus sequences in the RefSeq genomes database, but this time, change the algorithm from “Highly similar sequences (megablast)” to “Somewhat similar sequences (blastn)”.
   1. What word sizes do the megablast and blastn algorithms use by default? Hint: expand the Algorithm Parameters section at the bottom of the page. (2 pts)

megablast default word size:\_\_\_\_\_\_\_\_\_

blastn default word size: \_\_\_\_\_\_\_\_\_\_\_

* 1. How many hits did you have this time, when using blastn instead of megablast? (1 pt)
  2. What is the most distantly related poxvirus that produced an alignment with query coverage > 90%? (1 pt)
  3. What is the E-value of the overall lowest scoring hit in your results? Where did this sequence come from? (1 pt)
  4. Do you find this lowest scoring hit plausible? In other words, do you believe that it represents a legitimate homolog of the monkeypox virus DNA polymerase? Why or why not? (1 pt)

**Cellular homologs of viral genes**

1. Sometimes viral genes are evolutionarily related to cellular genes. Imagine you wanted to look for (possibly distant) homologs of the monkeypox virus DNA polymerase in cellular organisms. To do this, you would want to perform a BLAST search against protein sequence databases using the monkeypox virus DNA polymerase sequence.

What are two variants of BLAST that you could use to accomplish this? (2 pts)

1. OK, let’s use BLASTP to search for cellular homologs of the monkeypox DNA polymerase. To do this, translate the polymerase gene in Geneious and use this translated protein sequence as a query in a BLASTP search on the NCBI website. Restrict the results by selecting:
   * For database, select “Reference proteins (refseq\_protein)”.
   * For Organism, enter “cellular organisms (taxid:131567)”

Before beginning the search, answer these questions. Hint: expand the Algorithm Parameters section at the bottom of the page. (1 pt each)

* 1. What is the word size for this BLASTP search?
  2. What substitution matrix will be used to score amino acid matches or mismatches?
  3. What is the gap penalty scheme?

Now run the search and answer the following questions about the results (1 pt each)

* 1. What is the E-value of the top scoring hit?
  2. From what organism does the top-scoring hit derive?
  3. What is the pairwise % identity of the top-scoring hit?
  4. What is the pairwise % similarity (positive-scoring amino acids)?
  5. Does the top-scoring alignment involve the entire query sequencing?
  6. Do you believe that the top scoring sequencing is a legitimate homolog of the monkeypox virus DNA polymerase? Why or why not?

**Using BLAST to align two or more sequences.**

BLAST can also be used more flexibly by providing both query and subject sequence(s) on the NCBI website. To do this, you click the “Align two or more sequences” checkbox and enter query sequence(s) in the top box and subject (database) sequence(s) in the bottom box.

Let’s use this feature to investigate the location of rRNA genes in the *E. coli* genome. Navigate to the blastn website and turn on the “Align two or more sequences” checkbox. In the top box, enter the 16S sequence provided on Canvas as a query. In the bottom box, enter the accession of the *E. coli* K12 RefSeq genome (NC\_000913). (Note that you can enter NCBI accessions in the NCBI blast webpage instead of sequences!). Run the BLASTN search.

1. How many times is the 16S rRNA gene present in this E. coli genome? (1 pt)
2. Why do you think it’s necessary to have multiple copies of this gene in the genome? (1 pt)
3. Are all the 16S gene sequence copies identical? (Hint: switch to the Alignment tab of the results and select an alternative results view). (1 pt)
4. Are all the 16S gene sequence copies in the same orientation in the genome? (1 pt)