9.2 BLAST Applications – Tutorial

At the end of this tutorial you should be able to:

- Use BLAST with protein sequences
- Interpret the results of protein alignments
- Make inferences about sequence evolution using BLAST results
- Create and use BLAST databases
- Predict protein functions based on BLAST alignments

How to complete this tutorial

- Go through each question in order and complete any tasks that are described in the question.
- As you complete the questions, mark your answer to each question.
- Questions will be either:
  - multiple-choice questions that require you to provide either a single answer or to select multiple answers
  - questions that require a short text answer
- Open the associated quiz on Quercus and enter your answers to each question to verify that you completed the tutorial questions correctly.
- Alternatively, open the Quercus quiz when you start the tutorial and verify your answers as you complete the tutorial. **Note that there may be some information that is in this file that is not in the Quercus quiz!**
- The answers will be released at the end of the week.

Before you begin

- Open a new terminal session from your JupyterHub (New > Terminal)
- Set the PWD to `/home/jovyan/Week9/9.2.BLAST.Applications/Tutorial.9.2`
- Install BLAST

Data Sources:
Proteomes were downloaded from UniProt (https://www.uniprot.org/)
Protein and CDS sequences are downloaded from both Ensembl (https://ensembl.org/) and UniProt (https://www.uniprot.org/)

9.2.1: Protein Sequence Alignments

Question 1

CEP78 is a 78 kDa centrosomal protein. Mutations in CEP78 cause cone-rod dystrophy and hearing loss in humans. Examine the conservation of amino acid sequence of CEP78 by aligning the human protein (`Hs_ CEP78_protein.fa`) with CEP78 orthologs (`CEP78_orthologs_protein.fa`) in the following Eukaryotic species (MYA = million years ago):

Pan troglodytes (Chimpanzee) diverged from humans 6.7 MYA*
Mus musculus (Mouse) diverged from humans 90 MYA*
Gallus gallus (Chicken) diverged from humans 312 MYA*
Drosophila melanogaster (Fruit fly) diverged from humans 797 MYA*

* Divergence times from http://www.timetree.org/
Note that *Homo sapiens, Pan troglodytes, Mus musculus,* and *Gallus gallus* are vertebrates and *Drosophila melanogaster* is an invertebrate.

What command did you run for this alignment? (Assuming you did not use the `-outfmt` option.)

Question 2 (SELECT ALL THAT APPLY)

Using your results from the alignment of the CEP78 protein sequences in question 1, determine which of the following statements are true.
(**Hint:** You will need the full alignment results. For **b.** and **c.** look at the alignment with Mus_musculus_Cep78 between bases 61 & 120)

a. The percent positive substitutions is 16% higher than the percent identity for the *Homo sapiens* to *Gallus gallus* alignment.
b. When valine (V) is substituted with isoleucine (I) it is counted as a positive substitution.
c. When phenylalanine (F) is substituted with isoleucine (I) it is counted as a positive substitution.
d. The *Homo sapiens* amino acid sequence is shorter than the amino acid sequences of the orthologs in the other 4 species.

Question 3
The gene POLR2L encodes a subunit of RNA polymerases I, II, and III. Examine the conservation of amino acid sequence of POLR2L by aligning the human protein (`Hs_POLR2L_protein.fa`) with POLR2L orthologs (`POLR2L_orthologs_protein.fa`) in the same 4 Eukaryotic species.
Which of the following is true concerning the POLR2L orthologs?

a. The amino acid sequence is the same for all 5 species
b. The amino acid sequence is the same for 4 of the 5 species
c. The amino acid sequence is the same for 3 of the 5 species
d. The amino acid sequence is different for all 5 species. It is impossible for them to be exactly the same.

9.2.2: Sequence Evolution

Question 4 (SELECT ALL THAT APPLY)
Align the CDS of human POLR2L (`Hs_POLR2L_CDS.fa`) with the CDSs of POLR2L orthologs in the 4 other Eukaryotic species (`POLR2L_orthologs_CDS.fa`) using a **word size of 15.**
Based on the results of your alignment and the results of the alignment in question 3, select the true statements:
(**Hint:** You may want to look back at the divergence times in question 1.)

a. The percent identities for the corresponding amino acid sequence alignments and CDS alignments are the same.
b. The more distantly related the species is from human, the lower the percent identity.
c. The chicken ortholog is more similar to the human CDS than the mouse ortholog.
d. The amino acid sequence is more conserved than the CDS sequence.
e. Non-synonymous mutations affect conservation of the CDS sequence, but not the amino acid sequence.

Question 5
Compare the alignment of human CEP78 to the CEP78 orthologs in other species, and the alignment of human POLR2L to the POLR2L orthologs in other species.
Select the true statement:
(**Hint:** You may want to look back at the divergence times in question 1.)

a. CEP78 has been under stronger positive selection than POLR2L
b. POLR2L is more likely to be essential than CEP78
c. *Drosophila melanogaster* sequences are the least similar to human sequences because they diverged from humans more recently than the other species
d. CEP78 is more well conserved than POLR2L in vertebrates

9.2.3: BLAST Databases

Question 6

Caenorhabditis elegans or *C. elegans* is a species of nematode worm. It is a model organism that has been used to study neural development and aging. Another commonly used model organism is *Drosophila melanogaster* or *D. melanogaster*, a species of fruit fly.

In the `proteomes` directory within the `Tutorial.9.2` directory you will find the files `C_elegans_proteome.fasta` and `D_melanogaster_proteome.fasta`.

Create a BLAST database in the `proteomes` directory for the *C. elegans* proteome with the name `C_elegans_proteome`.

Fill in the command below to match the command you used (do not include what is already there!).

```
makeblastdb -in _________ -parse_seqids -dbtype prot
```

Question 7

Create a BLAST database in the `proteomes` directory for the *D. melanogaster* proteome with the name `D_melanogaster_proteome`.

Fill in the command below to match the command you used (do not include what is already there!).

```
makeblastdb -in D_melanogaster_proteome.fasta -out
D_melanogaster_proteome _________
```

Question 8

Change your directory back to `Tutorial.9.2`.

D. melanogaster and *C. elegans* are both invertebrates. *Ciona intestinalis* or *C. intestinalis* is another invertebrate species, commonly known as the sea squirt.

The file `Ci_unknown_protein_1.fa` contains a *C. intestinalis* protein. Align this protein to the *D. melanogaster* proteome to identify the most similar protein. Use `-outfmt 7` and no other optional arguments, and remember that you should be running this command from the `Tutorial.9.2` directory.

Make a note of the protein ID (6 letters and numbers) and the percent identity of the best hit.

Which of the following is the correct command:

a. `blastp -query Ci_unknown_protein_1.fa -db D_melanogaster_proteome -outfmt 7`

b. `blastp -query Ci_unknown_protein_1.fa -db proteomes/D_melanogaster_proteome -outfmt 7`

c. `blastp -query Ci_unknown_protein_1.fa -db D_melanogaster_proteome.fasta -outfmt 7`

d. `blastp -query Ci_unknown_protein_1.fa -db proteomes/D_melanogaster_proteome.fasta -outfmt 7`

Question 9
Align the protein `Ci_unknown_protein_1.fa` to the *C. elegans* proteome to identify the most similar protein. Make a note of the percent identity of the best hit.
What is the protein ID of the most similar *C. elegans* protein? (Protein IDs are 6 letters & numbers.)

Question 10
Go to UniProt (https://www.uniprot.org/) and look up the protein ID of the best *D. melanogaster* and *C. elegans* hits for the unknown protein.
Based on the function of these proteins, what process would you predict *C. elegans* unknown protein 1 is involved in?

a. Regulation of RNA splicing
b. Creation or stabilization of the 40S ribosomal subunit
c. Initiation of mitochondrial apoptosis
d. Organization of the actin cytoskeleton

Question 11
Perform the necessary alignments to predict the function of the *C. intestinalis* protein in `Ci_unknown_protein_2.fa`. Make sure you align it to both the *D. melanogaster* and *C. elegans* proteomes and keep track of the percent identities of the best hits.
Go to UniProt (https://www.uniprot.org/) and look up the protein ID of the best *D. melanogaster* and *C. elegans* hits for the unknown protein.
Based on the function of these proteins, what process would you predict *C. elegans* unknown protein 2 is involved in?

a. Transporting ions across the cell membrane
b. Controlling the MAPK signaling cascade
c. Localizing mRNA in the cell
d. Controlling the cell cycle

Question 12
Based on the results of the alignments of the unknown *C. intestinalis* proteins to the *D. melanogaster* and *C. elegans* proteomes, which species would you predict is more closely related to *C. intestinalis*? Enter either *D. melanogaster* or *C. elegans* into the text box below.