9.1 BLAST – Tutorial

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

- Understand the different types of homologs
- Install a package using Conda
- Run a basic BLAST command
- Understand the output of an alignment
- Apply options with arguments to blast commands

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/Week9/9.1.BLAST/Tutorial.9.1`

Data Sources:
CDS sequences are downloaded from Ensembl (https://ensembl.org/)

9.1.1: Sequence Alignment

Question 1 (SELECT ALL THAT APPLY)

Which of the following decrease the score of a nucleotide sequence alignment?

a. Nucleotides that match between the two sequences
b. Nucleotides that do not match between the two sequences
c. Gaps in the alignment between the two sequences
d. If one of the sequences is much longer than the other

Question 2

What does a substitution matrix do in amino acid sequence alignment scoring?

a. Substitutes all the amino acids with codons
b. Creates a conservation score for the alignment
c. Substitutes all the amino acids letter representations with numeric representations
d. Scores mismatches based on the properties of the amino acids

Question 3 (SELECT ALL THAT APPLY)

Gene X in *Felis catus* (domestic cat) and Gene Y in *Canis familiaris* (domestic dog) are derived from the same common ancestral gene. This means that Gene X and Gene Y are:

a. Homologs
b. Analogs
c. Orthologs
d. Paralogs
e. Specologs

9.1.2: Installing Packages in Terminal

Question 4

Install the BLAST package. What command did you use to install the package?

9.1.3: Basics of BLAST

Question 5

If you are not currently there, change your directory so that the PWD is `/home/jovyan/Week9/9.1.BLAST/Tutorial.9.1`.
Look at the files in the directory `question_5`. Each file stores two amino acid sequences. Which file is a correctly formatted FASTA file?

a. `fasta1.fa`
b. `fasta2.fa`
c. `fasta3.fa`
d. `fasta4.fa`

Question 6

If you are not currently there, change your directory so that the PWD is `/home/jovyan/Week.9/9.1.BLAST/Tutorial.9.1`.
The human gene MAP2K1 codes for a protein that is a component of the MAP kinase signal transduction pathway. The role of MAP2K1 in the cascade is to phosphorylate residues on the kinases MAPK3 and MAPK1. The orthologous gene in mice Map2k1 performs the same function.
Use BLAST to align the human MAP2K1 CDS (`Hs_MAP2K1_CDS.fa`) and the mouse MAP2K1 CDS (`Mm_Map2k1_CDS.fa`). Use the human gene as the query sequence.
What command did you use?

Question 7

What is the percent identity of the human MAP2K1 CDS and the mouse Map2k1 CDS from question 6? How many matching bases were there in the alignment?

a. 95 %, 1172
b. 67 %, 1065
c. 88 %, 1172
d. 91 %, 1065

Question 8

Using the alignment of the human MAP2K1 CDS and the mouse Map2k1 CDS from question 6, identify the length of each sequence and examine the alignment. Which bases (and how many) of each sequence are not part of the alignment?

a. The last 10 bases of each sequence are not part of the alignment.
b. 117 bases in each sequence are not part of the alignment. The bases are distributed randomly throughout the sequence.
c. As there are no gaps in this alignment, 0 bases in each sequence are not part of the alignment. (All bases are included in the alignment.)
d. The first 117 bases of each sequence are not part of the alignment.

Question 9

The human gene MAP2K2 also codes for a protein that is a component of the MAP kinase signal transduction pathway. MAP2K2 is a paralog of MAP2K1 and the genes share many of the same functions in the cell.

Use BLAST to align the human MAP2K1 CDS (`Hs_MAP2K1_CDS.fa`) and the human MAP2K2 CDS (`Hs_MAP2K2_CDS.fa`). Compare the results of E-value of this alignment to the alignment of human MAP2K1 and mouse Map2k1. Which alignment has a smaller E-value and what does it mean?

a. Human MAP2K1 & **mouse Map2k1** have a **smaller** E-value, thus the alignment is more significant.
b. Human MAP2K1 & **human MAP2K2** have a **smaller** E-value, thus the alignment is more significant.
c. Human MAP2K1 & **mouse Map2k1** have a **larger** E-value, thus the alignment is more significant.
d. Human MAP2K1 & **human MAP2K2** have a **larger** E-value, thus the alignment is more significant.

Question 10

Examine the alignment of the human MAP2K1 CDS and the human MAP2K2 CDS from the previous question. Which bases in each sequence are part of the alignment?

a. MAP2K1: 91 – 1160, MAP2K2: 79 – 1184
b. MAP2K1: 79 – 1160, MAP2K2: 91 – 1184
c. MAP2K1: 91 – 1184, MAP2K2: 79 – 1160
d. It is impossible to tell without counting the individual gaps and mismatches

---

9.1.4: BLAST Arguments

Question 11

The human genes PCBP1 and PCBP3 are paralogs. Each gene encodes an RNA binding protein. Each protein has 3 KH domains required for this RNA-binding activity.

Use BLAST to align the human PCBP1 CDS (`Hs_PCBP1_CDS.fa`) and the human PCBP3 CDS (`Hs_PCBP3_CDS.fa`). You should have 0 hits. Decrease the word size until you find a hit. What is the longest sequence of bases that matches between the two CDS sequences? (**Hint**: You may find it easier to view the results if you use the tab delimited output format!)

a. 13
b. 17
c. 21
d. 25

Question 12 (SELECT ALL THAT APPLY)
Decrease the word size in your blast command even further to 11, and only return hits that have an E-value less 0.01.
The regions of the PCBP1 CDS that code for the 3 KH domains are: 39-225, 291-486, 837-1029
The regions of the PCBP3 CDS that code for the 3 KH domains are: 135-285, 387-546, and 879-1071
Compare these regions to the portions of the CDS that align. Select the statements that are true.

a. There are two regions of alignment between the two CDSs
b. There are four regions of alignment between the two CDSs
c. All three of the KH domains in PCBP1 are (mostly) contained within the regions of alignment
d. All three of the KH domains in PCBP3 are (mostly) contained within the regions of alignment
e. The region between the 2nd and 3rd KH domain in both sequences is well conserved between the paralogs
f. The region between the 1st and 2nd KH domain in both sequences is well conserved between the paralogs

Question 13
Using either the `-help` command or the online BLAST manual (https://www.ncbi.nlm.nih.gov/books/NBK279684/), determine what the argument `-ungapped` does and how to use it. What is the length of the ungapped alignment of the human MAP2K1 CDS (`Hs_MAP2K1_CDS.fa`) and the human MAP2K2 CDS (`Hs_MAP2K2_CDS.fa`)?

a. 40
b. 273
c. 1096
d. 383

Question 14
Using either the `-help` command or the online BLAST manual (https://www.ncbi.nlm.nih.gov/books/NBK279684/), determine what the argument `-perc_identity` does and how to use it. What is the length of the longest **ungapped** alignment of the human MAP2K1 CDS (`Hs_MAP2K1_CDS.fa`) and the human MAP2K2 CDS (`Hs_MAP2K2_CDS.fa`) with a word size of 11 and a percent identity of at least 90 %?

a. 11
b. 273
c. 383
d. 21

Question 15 (SELECT ALL THAT APPLY)

What are the reasons for aligning CDSs instead of full gene sequences?

a. CDS alignments allow us to specifically examine the sequences that end up in the gene's encoded protein
b. Genes are too long for alignment
c. Introns and UTRs are less conserved causing breaks in the alignment
d. CDSs have a higher proportion of Cs and Gs than the full gene making it easier to align
e. Genes have intergenic sequences that cannot be aligned
f. Short exons will not show up in alignment results unless the word size is reduced