Bioinformatics

Assignment #2

Investigations using web BLAST

For this assignment, you will use the BLAST web site to investigate some genes/proteins. Please take the following steps, and in a Word document answer the questions.

1. a) Enter this chain of amino acids into a BLAST tool as the query in order to identify what sequence it is. Search the ref-seq database, which contains only curated sequences of high quality. Since you want an exact match, choose parameters for your search that are appropriate for very closely related species:

```
  1 mtpmrkinpl mklinhsfid lptpsnisaw wnfgsllgac lilqittglf lamhyspdas
 61 tafssiahit rdvnygwiir ylhangasmf ficlflhigr glyygsflys ktwnigiill
121 latmatafmg yvlpwgqmsf wgatvitnll saipyigtdl vqwiwggysv dsptltrfft
181 fhfilpfiia alaalhllfl hetgsnnplg itshsdkitf hpyytikdal glflfllslm
241 tltllspdll gdpdnytlan plntpphikp ewyflfayti lrsvpnklgg vlalllsili
301 lamipilhvs kqqsmmfrpl sqslywllaa dlliltwigg qpvsypfiii gqvasvlyft
361 tililmptis lienkmlkwa
```

b) Which of the 5 BLAST tools did you use? Which algorithm of that BLAST tool? Which database? What scoring matrix was used? What was the gap penalty? How long is the query sequence?

c) Use the best hit that is an actual sequence (not a predicted sequence) to identify this query sequence. Click on the accession number to obtain more information. Provide the name of gene or protein, name of organism, and the accession numbers of both protein and gene from which it is translated.

d) For your best hit, provide values for each of the following: accession number, max score, total score, query coverage, E-value and identity. Did you get an exact match? How do you know?

2. a) Employing a BLAST search, use the above query sequence to find a ~~*reference*~~ ***nucleotide*** sequence ***from only the species you identified above*** that corresponds with the query sequence. (Hint: read the descriptions of the 5 different BLAST searches {nucleotide blast, protein blast, blastx, tblastn, tblastx} and choose the appropriate one. Also choose the reference nucleotide database will match your query given the tool you chose). What is the nucleotide sequence? How did you find it? (Crossed out "reference" - there seems to be no reference nucleotide sequence for this species).

b) Provide the accession number, its max score, its total score, the query coverage, E-value and max identity. What scoring matrix was used? What was the gap penalty?

3. **Skip.** (This question has been removed because the Expasy part will not give a satisfying answer if you don't have the mRNA rather than DNA sequence. This is difficult to get for our sequence.)

   ~~a) Find the amino acid translation of the nucleotide sequence from #2 for all six reading frames. Provide the first 10 amino acids of each *reading frame* (NB: **not** *open* reading frame). You may use a web tool or your own code to accomplish this. A good tool may be found here: http://web.expasy.org/translate/~~

   ~~b) Which reading frame contains the original protein sequence? Paste in the amino acid sequence of this *open reading frame* (it should be nearly identical to the original query sequence). If you had only the gene sequence but not the protein sequence how would you know from theses reading frames alone how to eliminate the 5 reading frames that do not contain the actual amino acid sequence and determine which open reading frame contains the actual amino acid sequence?~~

4. a) Using the original query sequence, conduct a BLAST search to find sequences that are homologous in a variety of species, some closely related to the original species and some not. Choose a scoring matrix that is optimized for distantly related species. Report the BLAST tool you used, the algorithm of that tool, the database you searched, the scoring matrix and the gap penalty

   b) In the results window, select 15 high-scoring hits that actual proteins (ex: not experimental, predicted, etc.) from a variety of species that you will use to build a multiple sequence alignment and a distance tree. Include human in your selected hits. After checking the boxes of the 15 different species you chose, click on the link for "Multiple alignment". Paste the first aligned section (that shows FASTA amino acid symbols), extending from position 1 to about 60 or 70, into your report.

   c) Click on the link for "Distance tree of results". Paste the distance tree into a separate page in landscape orientation. Note that tree nodes are colored according to type of species. Does the tree show the evolutionary relationships that would be expected? Briefly support your answer. (NB: the distance tree is **not** a phylogenetic tree for determining evolutionary relationships.)

5. a) Using the ***human ortholog*** to the original sequence from #1, BLAST the appropriate reference sequence database. Optimize search parameters so as to find an exact match, and confine the search to Homo sapiens (taxid:9606).

   b) What BLAST algorithm did you use? What scoring matrix and gap penalty was used? How long is the query sequence?

   c) Paste a screen shot of the Graphics Summary

   d) For your best hit that is an actual (not predicted) sequence, provide values for each of the following: accession number, max score, total score, query coverage, E-value and identity. Click to see the alignment. Comment on the alignment, particularly which parts of the query are covered.

e) For your worst hit that is an actual (not predicted) sequence, provide values for each of the following: accession number, max score, total score, query coverage, E-value and identity. Write a sentence contrasting how best and worst hits appear in the Graphics Summary. Write an additional sentence explaining the differences in the above statistics.

f) Click on the distance tree. Explain why some human hits appear on the tree as more distantly related than the hits of other species.

6. Click on the accession number of the best (human) hit to obtain more information. Specifically, what genome is this protein translated from?

Species information (optional aid in checking your tree for evolutionary relationships)

List of mammal genera (NB: genera is plural of genus)
https://en.wikipedia.org/wiki/List_of_mammal_genera>

Primate (scroll down to the phylogenetic tree)
https://en.wikipedia.org/wiki/Primate>

BLAST Resources to supplement text (optional):

Webinar: A Practical Guide to NCBI BLAST
https://www.youtube.com/watch?v=KLBE0AuH-Sk

Visualize and Interpret Alignment Data with the Multiple Sequence Alignment Viewer
https://ncbiinsights.ncbi.nlm.nih.gov/2017/01/25/visualize-and-interpret-alignment-data-with-the-multiple-sequence-alignment-viewer/

The New MSA Viewer (2 min 42 sec video)
https://www.youtube.com/watch?v=NZ1vsIp0j_8