BIOL222

Lab 8

Bioinformatics

**Objective**

* To analyze the DNA sequences that you produced

**Before Lab**

* Go to this web site <http://www.digitalworldbiology.com/BLAST/index.html> and click the green arrow to progress through the tutorial on doing BLAST.
* Use one of the test sequences at this web site to answer the questions found on the file “These are the questions you will answer about your test sequence” otherwise called Blasting through the kingdoms of life <http://www.digitalworldbiology.com/BLAST/62000sequences.html#BM10>
* If you are going to do this on your computer, you will need to download a free version of the sequence software FinchTV <http://www.geospiza.com/Products/finchtvdlrequest.shtml>
* When we get your sequence file, we will upload it into FinchTV and directly from there we can link your sequence into the BLAST bioinformatics tool. ( In Finch go to Edit and then BLAST your sequence). If you would like to play with some sequence, I have loaded a sequence file into the folder that you can upload into the FinchTV.
* Now answer the following questions about the BLASTed sequence:

**Questions:** Be sure to include the source of the information along with your answer. In

this case, the source will be the database or web page that provided the information.

1. How long is the sequence that was used to search the database?

*Hint: This sequence is called the "query" sequence because you used it to ask a*

*question (or query) of the database.*

2. What is the most likely identity of this sequence? What data supports this

conclusion?

*Hint: Refer to the slide in the BLAST tutorial that discusses the E value.*

3. What organism was the most likely source of the sequence?

*Hint: Refer to the BLAST tutorial to find an overview of the GenBank nucleotide*

*record. If more than one organism matches, look at the E values to determine the*

*most likely match.*

4. What is the common name for this organism?

*Hint: Refer to the GenBank nucleotide record. It may also help to look at the*

*Taxonomy database. The BLAST tutorial shows where to find this link.*

5. What phylum contains this organism?

*Hint: Refer to the taxonomy database. The BLAST tutorial shows how to find the*

*link.*

6. What is the accession number for the best-matching sequence?

7. Is this sequence expressed? How do you know?

*Hint: Gene expression includes the processes of transcription (making RNA) and*

*translation (making a protein). Determine if either of these molecules is described in*

*the sequence record.*

8. If your sequence is expressed, where is this gene expressed?

*Hint: The sequence record usually indicates the source of the material that was*

*sequenced. In some cases, however, you will probably need to look at other sources*

*of information. Some places to look are the title of the submission, PubMed records*

*(you might need to look at more than one record), the UniGene database, and the*

*Gene database. If your BLAST results show boxes with U's or G's*

*those are links to the UniGene and Gene databases respectively. The Expression*

*Profile in the UniGene database is linked to a set of tables that show the tissues*

*where mRNAs have been found and the developmental stage where they were*

*found.*

9. Is there a specific time during development when this gene is expressed?

*Hint: See the hint for question 8.*

10. Is anything known about factors that cause your sequence to be expressed?

*Hint: The title of the submission is a good place to start. PubMed records and the*

*Entrez Gene database are also helpful; see the hint for question 8.*

11. Estimate the number of sequences with an E value less than 0.01.

*Hint: Refer to the blast results.*

12. If possible, give the names of three different organisms with significant E values. If

organism is represented, then write down the name of that organism.

*Hint: Refer to the BLAST tutorial slide on E values for a description.*

13. Look at the first matching sequence; determine the length of the alignment and the

fraction of nucleotides that match your sequence. Draw a picture to represent the

alignment between the two sequences and include the starting and ending map

positions for the both sequences.

14. Use GenBank, PubMed, Gene, and UniGene records to find the possible function of

the protein that's specified by your DNA sequence. You may need to just Google the name

**Open Reading Frame**

When you have your sequence you will also determine the Open Reading Frame assignment on a part of your sequence.

Take one of your sequences: This is the input that you are going to provide to the BLASTx and tBLASTx programs.

Pick one of your sequences. Select a 70 base section that has good clean sequence.

Paste that sequence here(BLASTx/tBLASTx): e.g.

5’catccgcgcatggcattggctaaacggatgattcaaaacacccaagtcaatgaacggtatctggtcttgc 3’ 3’gtaggcgcgtaccgtaaccgatttgcctactaagttttgtgggttcagttacttgccatagaccagaacg 5’

On the next line type in the complementary sequence ( see red sequence above)

(practice by copying and pasting the sequence above into tBLASTx)

Now start translating this sequence. Yes – I know you do not normally translate DNA but when you don’t know where the genes are, this is how to find out.

The question is which strand do you translate and where do you start?

As we don’t know the answers to those questions we cover all the possibilities

Pick the top strand, start translating at the 5’ end using the three letter codons listed in the genetic code table below (**what do you do about the Us in the codon table?**) , at the first base and keep going till you reach the end; then start at the second base and continue till the end and then start with the third base and continue till the end.

Now do the same thing with the second strand.

Do not worry about one or two extra bases (**why did I say one or two extra bases and not three?**) at the end of each sequence

Stop codons are shown by an \*.

Amino acids coded for by each codon is shown by a single letter code (see the codes below)

At the end your sequence should look like this but completed all the way to the end.

 5’ catccgcgcatggcattggctaaacggatgattcaaaacacccaagtcaatgaacggtatctggtcttgc 3’

+1 cat ccg cgc atg gca ttg gct aaa cgg atg att caa aac acc caa gtc aat gaa cgg tat ctg gtc ttg c

 H P R

+2 c atc cgc gca tgg cat tgg cta aac gga tga ttc aaa aca ccc aag tca atg aac ggt atc tgg tct tgc

 I R A

+3 ca tcc gcg cat ggc att ggc taa acg gat gat tca aaa cac cca agt caa tga acg gta tct ggt ctt gc

 S A L

 R T K

-3 gt agg cgc gta ccg taa ccg att tgc cta cta agt ttt gtg ggt tca gtt act tgc cat aga cca gaa cg

E R N

-2 gta ggc gcg tac cgt aac cga ttt gcc tac taa gtt ttg tgg gtt cag tta ctt gcc ata gac cag aac g

 P R A

-1 g tag gcg cgt acc gta acc gat ttg cct act aag ttt tgt ggg ttc agt tac ttg cca tag acc aga acg

3’ gtaggcgcgtaccgtaaccgatttgcctactaagttttgtgggttcagttacttgccatagaccagaacg5’

 <http://scienceblogs.com/oscillator/index.php?utm_source=rssTextLink&page=4>

|  |  |
| --- | --- |
|  | Symbol |
| Name  | 3-Letter | 1-Letter |
| Alanine | Ala | A |
| Arginine | Arg | R |
| Asparagine | Asn | N |
| Aspartic acid | Asp | D |
| Cysteine | Cys | C |
| Glutamic Acid | Glu | E |
| Glutamine | Gln | Q |
| Glycine | Gly | G |
| Histidine | His | H |
| Isoleucine | Ile | I |
| Leucine | Leu | L |
| Lysine | Lys | K |
| Methionine | Met | M |
| Phenylalanine | Phe | F |
| Proline | Pro | P |
| Serine | Ser | S |
| Threonine | Thr | T |
| Tryptophan  | Trp | W |
| Tyrosine | Tyr | Y |
| Valine | Val | V |
| Unknonwn or "other" |  | X |

Did you get a full set of amino acids with no \*s in all 6 open reading frames?

What if you get more than one ORF?

Look at your BLASTx and tBLASTx results – which of these ORFs were similar to proteins (real or predicted) in the database?

Do we have to consider introns?