The Assignment

Please hand in the solutions to the following problems on Wednesday, August 11, 2010. Hand in a hard copy and a USB key containing your solutions.

Problem 1

A) The nucleotide sequence of one DNA strand of a double helix is given. Write the complementary sequence found on the other strand. Notice that the new sequence you will write is on the lower strand. Do not forget to label the ends of your sequence.

```
5′ --- CACTGTCATGGCTTTTGATCAAAAAAA --- 3′
```

B) Search the Web for an on-line tool that will find the complement of a DNA sequence. Write down the URL.

C) Suppose that the DNA molecule from part a) is transcribed and the lower strand (from 3' to 5') is used as the template strand. What is the RNA sequence obtained from the transcription? Label the 5' and 3' ends of the molecule.

D) What is the difference between the RNA molecule you obtained and the given sequence of part a)?

Problem 2

Complete the following table. Assume that

- the columns represent transcriptional and translational alignments;
- the top DNA strand is in the 5' to 3' direction and is the coding strand, while the bottom DNA strand is the template strand and goes from 3' to 5'.

G												**DNA double helix**
						T	**C**	**A**				
	C	**U**										**mRNA transcribed**
									C	**G**	**A**	**Appropriate tRNA anticodon**
			Met (M)									**Amino acids incorporated into protein**

Problem 3

A) The following is a sequence of bases within a segment of a RNA molecule.

```
5'--- CACUGUCACGGCUUUAGAUCAAAAAAA --- 3'
```

Write the amino acid sequence that would exist in the corresponding segment of the encoded polypeptide molecule. Assume that translation has been initiated and that this sequence is in the proper reading frame (first reading frame).

B) Find an on-line RNA translation tool. Write down its URL.
C) Write all six possible reading frames from 5' to 3' of the following sequence:

```
5′ --- GCACTAGTCATGGCTTTTGAC --- 3′
```

Problem 4

NCBI has a sample GenBank record at:

http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

Please go to that site, read the example and answer the following questions:

a) What does CDS stand for? Explain.

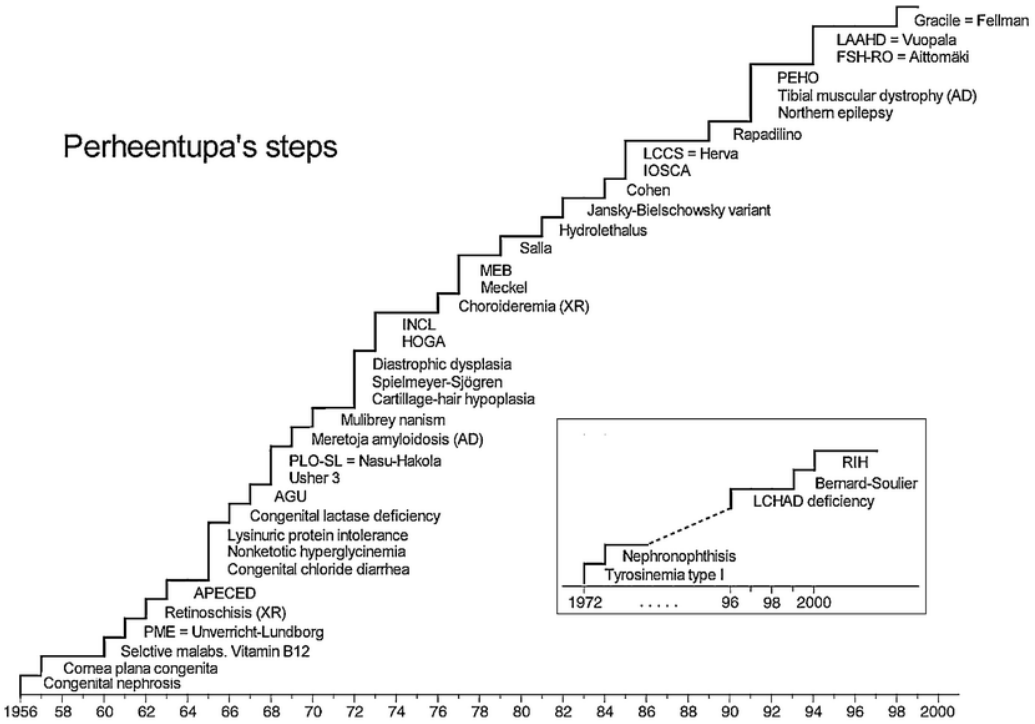
b) Is it better to search by using the actual accession number or the locus name? Why?

c) What does GI stand for? How does it differ from accession?

d) There are three occurrences of "CDS" under "Features". There are three occurrences of "CDS" under "Features".

1. Consider the first occurrence of "CDS". One of its subfields is "/translation". Explain why "/translation" starts with the specific sequence of amino acids: SSIYN.
2. Consider the third occurrence of "CDS". Explain why "/translation" ends with the specific sequence of amino acids: GSLF.

Problem 5



Choose a gene from the above figure and write a paragraph or two about it.

Problem 6

You are a researcher who has just obtained the following sequence:

```
> unknown sequence
tttcttgaccccttcatgggactcccaacaggggtaccccatttactcagcctgccctgc
tcaacctcttgcaggagggagcacacgagtgaacgagtgcaggaaccagctggctgcttt
agtgctgtgaggagtaaactccatgcaggccctgcagcagcaaccagtttttccagattt
gctcaaagcaatcccagtgagcatccacgtcaatgtcattctcttctctgccatccttat
tgtgttaaccatggtggggacagccttcttcatgtacaatgcttttggaaaacctttttga
aactctgcatggtcccctagggctgtaccttttgagcttcatttcaggctcctgtggctg
tcttgtcatgatattgtttgcctctgaagtgaaaatccatcacctctcagaaaaaattga
aattataaagaagggacttatgtctacaaaacgcaaagtgaaaaatataccacctcattc
tggctgactaaaggccacagctgagcctggaactgacccttccttcatcctcaacctgct
gtcctccagaaagcaccaaggaaaaagcagagaatgacagcaaacagatcactaggcctc
tgaccacaggtgctgagtactcagcagccctcatataataggtttgaaa
```

Use NCBI tools and databases to answer the following questions. Make sure to use blastn. For help, refer to “Blast Program Selection Guide” available at

http://www.ncbi.nlm.nih.gov/blast/producttable.shtml

A) Determine the organism and the name of the gene from which the sequence most probably comes from. Which tool did you use to find the answer?

B) What is the accession number of the GenBank record of the sequence that is the best match to the “unknown sequence”? Save the sequence of the best hit in a file in FASTA format.

C) Display the GenBank record using the accession number from B). From “Customize View” deselect “Show Sequence” under “Display options” and click on “Update View”. Print the record. Use this GenBank record to answer the questions below. On your printout of the record, highlight and label (e.g. Question #) the information used to answer each question. Attach the annotated record to your homework solution.

1) Which publication is the first (oldest) reference for this gene?

2) Which locus does the gene reside at?

3) How many human paralogues of this gene been identified? On which chromosome(s)?

4) Are there any known orthologues of this gene? If yes, in which organisms?

5) What is the name of the protein that is obtained from the gene? What is the GenBank accession number of the protein record?

6) At which position in the given sequence does the translation start?

7) How many amino acid residues are encoded by this gene? How did you determine that?

8) What is the function of the protein encoded by the gene?

9) What is the name of the disorder caused by the mutations in this gene?

10) A protein domain is a region of a protein with a well-defined structure and function. Which domain(s) is present in the protein encoded by this gene?

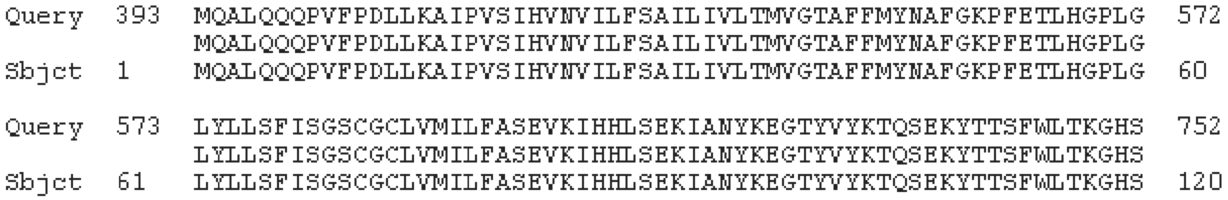
D) Perform "translated blast search" of NCBI protein database (blastx) with the "unknown" sequence as query. What is the accession number of the best hit? Did the search confirm the origin and function of the protein found in Steps A) through C)? Save the amino acid sequence of the best hit in a file in FASTA format.

E) Use BLAST to perform a pairwise alignment between the gene sequence from part B) with the protein sequence from part D). From the main page of BLAST, choose "Align two (or more) sequences using BLAST (bl2seq)" under "Specialized Blast". Enter the nucleotide sequence in the first window and the protein in the second. Choose "blastx" from the top. How similar are these two sequences?

F) Explain why the alignment starts at position 393 in the query sequence.

G) Examine the output carefully.

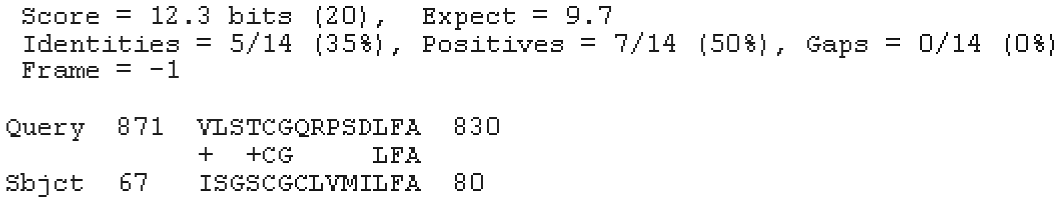
i) Explain why the two sequences below (taken from the bl2seq output), do not have the same range.



```
Query  393  MQALQQQPVFPDLLKAIPVSIHVNVILFSAILIVLTMVGTAFFMYNAFGKPFETLHGPLG  572
            MQALQQQPVFPDLLKAIPVSIHVNVILFSAILIVLTMVGTAFFMYNAFGKPFETLHGPLG
Sbjct  1    MQALQQQPVFPDLLKAIPVSIHVNVILFSAILIVLTMVGTAFFMYNAFGKPFETLHGPLG  60

Query  573  LYLLSFISGSCGCLVMILFASEVKIHHLSEKIANYKEGTYVYKTQSEKYTTSFWLTKGHS  752
            LYLLSFISGSCGCLVMILFASEVKIHHLSEKIANYKEGTYVYKTQSEKYTTSFWLTKGHS
Sbjct  61   LYLLSFISGSCGCLVMILFASEVKIHHLSEKIANYKEGTYVYKTQSEKYTTSFWLTKGHS  120
```

ii) Explain in your own words what is meant by "Identities" and "Positives" in the second alignment:



```
 Score = 12.3 bits (20),  Expect = 9.7
 Identities = 5/14 (35%), Positives = 7/14 (50%), Gaps = 0/14 (0%)
 Frame = -1

Query  871  VLSTCGQRPSDLFA  830
            +  +CG     LFA
Sbjct  67   ISGSCGCLVMILFA  80
```

H) Find an on-line resource for protein domain prediction.

1) Write the URL of the resource below:

2) Use the amino acid sequence from part D) as a query and find all protein domains in the protein. Which protein domains were identified in the sequence? Compare with Part C) Question 10).

3) What are the coordinates of each domain? Print the results of the prediction and submit with your homework. Highlight the information used to answer Questions 2 and 3.