

Bioinformatics: Life Sciences on Your Computer

Week 4 Practice Problems

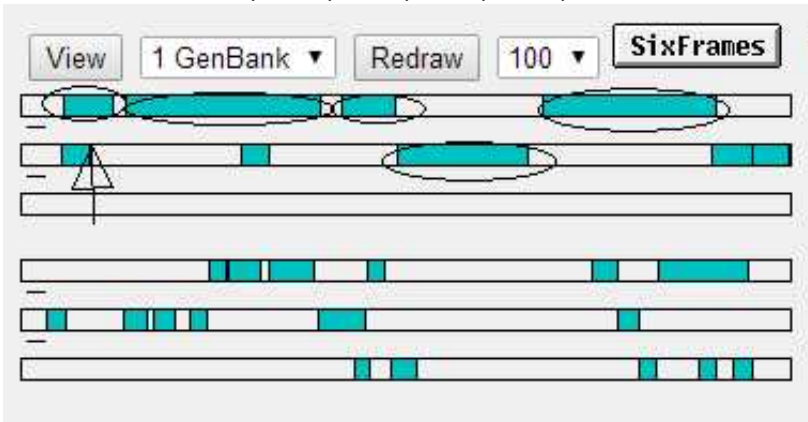
1. Find the nucleotide database with the accession number, KC333876.
 - a. What organism?
 - b. What gene names are listed in the definition field?
 - c. List the locations of the -35 signal and the -10 signal for the operon.
 - d. Why is the -10 signal more than 10 bases upstream from the start of the first CDS of the operon?
 - e. Why is there no -35 and -10 signal just upstream of the third CDS (second of operon)?
 - f. Run the FASTA DNA sequence through a gene prediction at [FGENESB](#). Choose Bacterial generic as the training set. List any discrepancies from the GenBank record.
 - g. How many different mRNA transcripts are needed to produce the six CDS regions listed in GenBank and by FGENESB?
 - h. Run BLAST using the predicted protein from [FGENESB](#) from the first CDS (should be on complement strand). Then run BLAST with the first protein sequence from the GenBank record. Based on your BLAST results, which query sequence do you believe to be more correct--the prediction or the NCBI-annotated protein? Briefly explain.
2. Search the nucleotide database for gi|41096.
 - a. What organism? List the gene names from the record title.
 - b. Use the [ORF Finder at NCBI](#) to find open reading frames. How many genes do there appear to be in the operon? Hint: look for long ORFs (rectangles with little space between).
 - c. List the locations of the three longest ORFs.
 - d. Try a BLAST search off this page. Click the longest ORF. It should turn pink. Then click BLAST. You will have to manually click View Report. What is the enzyme name?
 - e. Run the bacterial sequence (41096) at FGENESB. Choose *Escherichia coli* K12 as closest organism. What are the predicted locations?
 - f. How do those locations compare to the locations from ORF Finder and the GenBank record?
3. Access the sequence at <https://d396qusza40orc.cloudfront.net/bioinform/assignments/mvarigprt.fasta>. The sequence is from a bacterium that was recently sequenced but not annotated as of June 29, 2014. Run the web version of [Glimmer v3](#), now available at NCBI.
 - a. How many ORFs predicted?
 - b. How many ORFs are predicted to be on the minus strand?
 - c. Run the same sequence at GeneMark (<http://www.ncbi.nlm.nih.gov/genomes/MICROBES/genemark.cgi>). What is the only difference in the prediction?
4. Use the *Lactococcus* DNA at <https://d396qusza40orc.cloudfront.net/bioinform/assignments/lactococcus.txt>.
 - a. Run [FGENESB](#) (use *Lactococcus lactis* as closest organism) and report the location of two CDSs on an operon.
 - b. Then run [BPROM](#) to report the locations of the most appropriate -35 signal and the -10 signal.
 - c. At what nucleotide is the predicted transcription start site (promoter pos.)?

5. A complete genome of a simian immunodeficiency virus isolate has accession AF447763 at NCBI. Access that sequence.
- What is the location of the gag gene?
 - In addition to gag, pol and env, what other CDSs are present?
 - The host is listed as *Pan troglodytes schweinfurthii*. What is a more common name for this organism?
 - Retrieve the SIV Gag protein sequence (from the AF447763 record). Run a pairwise alignment at BLAST with the HIV-1 Gag protein with accession BAF32552. What is the % identity and % similarity?

Answers

- 1.a. *Aquicola tertiaricarbonis*
- b. iso gene, the ethABCD gene cluster and tnp gene (note: ethABCD is an operon with ethA, ethB, ethC and ethD)
- c. -35: 1552..1561; -10: 1574..1579
- d. It's 10 bases upstream from the transcription start site, not from the start codon.
- e. It's on the same mRNA as three other CDS regions.
- f. The first CDS has a different predicted START codon. This is a start codon, not a stop because the complementary strand genes are transcribed in the opposite direction. GenBank has the start codon at 1219 but FGENESB has it at 1297.
- g. There are three mRNAs. The middle four CDSs are all on the same mRNA.
- h. It is hard to say for sure, because incorrect sequences are present in GenBank, however the FGENESB prediction appears to be more correct.

- 2.a. *Escherichia coli*, celA, celB, celC, celD, celF



- b. five are circled, all on plus strand - very little space between. A 6th and a 7th on the right side could possibly be argued as well. Five or seven are both reasonable answers.
- c. 691..1944; 3394..4512; 2447..3289
 - d. PTS N'-diacetylchitobiose transporter subunit IIC
 - e. 286..606; 691..1944; 1905..2030; 2089..2439; 2450..3289; 3394..4512; 4532..4747
 - f. The three longest ORFs from ORF finder are matched perfectly, except the start codon of one--ORF Finder says 2447; FGENESB says 2450. The 6th and 7th genes from ORF Finder appear to be genes, but on a separate operon from the first five. GenBank has that start codon at 2447. GenBank also does not have the CDS at 1905..2030 and from 4532..4747.

3.

```
>Mannheimia varigena partial
orf00001      367      1452  +1      10.06
orf00002     1566      2192  +3      10.30
```

orf00004	2311	4254	+1	12.24
orf00005	4454	5548	+2	10.55
orf00006	5592	6890	+3	12.86
orf00007	7554	6946	-1	9.75
orf00008	8883	7618	-1	13.04
orf00009	9572	8892	-3	11.53
orf00010	9870	10952	+3	12.19
orf00011	11977	11000	-2	12.62

a. 10

b. 3

c. GeneMark predicts an extra partial gene at <2..181, meaning that if the prediction is correct, the first nucleotide in this partial sequence is in the middle of a gene (ORF) that starts before the beginning of this sequence.

4.a. 287..553; 556..2283

b. -35 at 190 and -10 at 210 is the most plausible location

c. TSS- at 225

5.a. 353..1927

b. vif, vpr, tat, rev, vpu, nef

c. Eastern chimpanzee (from a web search)

d. 60% identical, 70% similar