Biology/CS/SE 123A
Bioinformatics I

Fall 2014

Homework Four

Please hand in the solutions to the following problems on Tuesday, November 11, 2014. Hand in a hard copy (required) containing your solutions.

Problem One Problems from the textbook – Chapter Four

1) Upon scoring pairs of aligned amino acids, why are pairs such as isoleucine and leucine, scored more highly than other pairs such as isoleucine and lysine?
2) What is the minimum percentage identity that can reasonably be accepted as significant when comparing two protein sequences? Explain making sure you include what is meant by "twilight zone" in your answer.
3) When we are deciding on which substitution matrix to use, is there such a thing as one correct scoring scheme for all circumstances? Explain.
4) The rows in the substitution matrices of Figure 4.4 on page 83 are color coded. Explain what each color represents. Write one or two sentences on each property (beyond what is written in the caption).
5) Explain each of the following terms:
    a) PAM b) BLOSUM c) PET91
    d) STR e) SLIM f) PHAT
6) What are the two findings that were obtained from structural analysis of the sequence of amino acids w.r.t. insertions and deletions of amino acids?
7) When should we set a high gap penalty and should we set a low gap penalty?
8) Why should amino acids such as tryptophan have higher gap penalties (when aligned with gaps) than other amino acids such as glycine?
9) a) What are the Smith-Waterman and Needleman-Wunsch algorithms?
    b) What are they used for?
    c) Which one is a special case of the other?
10) What do sequence conservations in multiple alignments identify?
11) a) What do the authors mean when they claim that Needleman–Wunsch and Smith-Waterman methods are rigorous?
    b) What programming technique are these methods based on?
12) a) Is BLAST a rigorous method? Explain.
    b) Is SSEARCH a rigorous method? Explain.
13) Is the default gapped setting of BLAST adequate for most applications?
14) What are the differences between blastn, blastp, and blastx?

15) Consider Figure 4.12 (A) on page 99. Why does the caption of the figure (on
page 98) claim that hits above the arrow are significant, while the ones below are not. Fully explain.

16) Name two actions one could undertake to reduce the large number of hits one gets upon blasting a sequence against a database of sequences.

17) a) What are low complexity regions?
b) Are they desirable? Why?

18) Why does it make sense to resubmit a query sequence sometime after it was found to have no match in the database?

19) a) What is the simplest method of constructing a pattern or motif?
b) How does the method work? In other words, explain the method.

20) a) What are logos?
b) How are they constructed?
c) What do the size of the letters indicate?

Problem Two

Go to NCBI and retrieve the record (data entry) with accession number Z48051.

a) What is the name of the gene and what is its locus (Example 7p2)?
b) How many exons and introns does the gene have?
c) Under the second misc_feature, we have: “polymorphic (TAAA)n". Explain in your own words the meaning of this entry in the annotation of Z48051.
d) Exon 7 is from base pair 14658 to base pair 14678. What are the amino acids produced by exon 7?
e) Exon 5 is from base pair 11860 to base pair 11880. What are the amino acids produced by exon 5?

Problem Three

The following sequence represents the last 285 nucleotides of exon 8 of the human gene for myelin oligodendrocyte glycoprotein (MOG).

```
>Part of exon 8 of H.sapiens gene for myelin oligodendrocyte glycoprotein
TTCAAGACCAGCCTGGCCAACATGGTGAAACCCCATCTCTACTAAAAATACAAACAATTAACTGAGCATA
GTGGTGGGCACCTATAATACCAGCTACTCCGGAGGCTGAGGCAGGAGAATCGCTTGAACCCAGGAGGCAG
AGGTTGCAGTGAGCTGAGATCGCGCCACTGCACTCTAGCCGGAGTGACAGAGTAAGACTCTGTCTCAAAA
ATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAAAAATAATAATACAAGTTTTCATAAGCACA
CTTCT
```

Examine the sequence using the dot plot at:
http://www.vivo.colostate.edu/molkit/dnadot/

Use: a) Window Size = 9, Mismatch Limit = 0.
b) Window Size = 9, Mismatch Limit = 2.

What can you infer from the results? Include snapshots of the dot plot output to support your analysis.

Problem Four

The beta globin gene ends with the following nucleotide sequence:

... ctg gcc cac aag tat cac taa

a) Translate the above sequence and give the resulting amino acids.
b) Write the nucleotide sequence of a single base change producing a silent mutation in the region. Recall that a silent mutation is one that leaves the amino acid sequence unchanged.
c) Write the nucleotide sequence and the translation to an amino acid sequence, of a single base change producing a missense mutation in the region.
d) Write the nucleotide sequence and the translation to an amino acid sequence, of a single base change producing a nonsense mutation in the region.
e) Write the nucleotide sequence and the translation to an amino acid sequence, of a single base change producing a mutation in this region that would lead to improper chain termination resulting in extension of the protein.

Problem Five

CLUSTAL Omega is available for PCs and also on a Web site at EMBL-EBI in the United Kingdom: http://www.ebi.ac.uk/Tools/msa/clustalo/

The file "myh16_sequences.txt" contains eight nucleotide sequences from the masticatory myosin heavy chain (MYH16). Copy and paste the 8 sequences in FASTA format into the CLUSTAL OMEGA data window. Make sure you have "DNA" from dropdown window in Step 1. Under STEP 2, click on "More options ..." and choose "input" under "ORDER". Click on "Submit" to align the eight sequences.

Note that the eight sequences are from the masticatory myosin heavy chain (MHY16):

- Accession Number: AY350716.1 Woolley monkey (Lagothrix lagotricha) masticatory myosin heavy chain (MYH16) gene, exon 18 and partial cds
- Accession Number: AY350717.1 Rhesus macaque (Macaca nemestrina) masticatory myosin heavy chain (MYH16) gene, exon 18 and partial cds
- Accession Number: AY350718.1 Pigtailed Macaque) (Macaca mulatta) masticatory myosin heavy chain (MYH16) gene, exon 18 and partial cds

- Accession Number: AY350719.1 Orangutan (Pongo pygmaeus) masticatory myosin heavy chain (MYH16) gene, exon 18 and partial cds
- Accession Number: AY350720.1 Gorilla (Gorilla gorilla) masticatory myosin heavy chain (MYH16) gene, exon 18 and partial cds
- Accession Number: AY350721.1 Bonobo (Pan paniscus) masticatory myosin heavy chain (MYH16) gene, exon 18 and partial cds
- Accession Number: AY350722.1 Chimpanzee (Pan troglodytes) masticatory myosin heavy chain (MYH16) gene, exon 18 and partial cds
- Accession Number: BK001410 Human (Homo Sapiens), Homo sapiens myosin heavy chain (MYH16) pseudogene, exon 19 (nucleotides 36298-36383 from BK001410)

The alignment itself will appear under "CLUSTAL O (1.2.0) multiple sequence alignment". Choose "Show Colors" (assuming the link for color appears).

a) Copy the alignment obtained (found under: "CLUSTAL O(1.2.0) multiple sequence alignment"), paste it and submit it as the solution to this part of the problem.

b) What kind of substitutions do we have? Look at the columns and count the number of transitions and of transversions.

c) What kind of indels (deletions/insertions) can you see? What are (or might be) the consequences of the indels? Fully explain your answer.

Note that at the top of the page, there is link to "Result Summary". If you click on it, you will find a link to "JalView", a JAVA alignment viewer and editor. Go there and increase the font to view a different way of showing the alignment of the eight sequences.

Problem Six

Who is/are the Nobel Prize recipient(s) in Chemistry of 2013? Write a paragraph or two on his/their work that lead to the award. Alternatively, you can answer the same question on the Nobel Prize recipient(s) in Physiology or Medicine (instead of Chemistry).

Problem Seven

Given the following two sequences S and T:

S = GCTAGTCAGATCTGACGCTA
T = GATGGTCACATCTGCCGC

a) Construct a simple dot plot (window size = 1). Recall, you need to place each sequence along an axis, and place a dot in the plot for each identical pair of nucleotides. Does your plot reveal any regions of similarity?

b) Construct a dot plot using a sliding window of size 4 with stringency value = 3. Does this plot reveal any regions of similarity between the two sequences?

c) Which dot plot is better, the plot with window size = 1 (part a) or the plot with window size = 4 and stringency value = 3 (part b)? Why?