CENG 465 Spring 2008-2009

Written Assignment #2

Please use the template file given on COW to fill-in your answers, and upload your answers as a single .txt file to COW.

Problem 1 (30 Points):

Get the protein sequences for human IRGM (Accession #: **NP_001139277.1**) and the rat IRGM (Accession #: **NP_001012007.1**). You may use the Entrez Protein Search at NCBI web site (<u>http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein</u>) to fetch the sequences. At the detailed protein information page, you may use the FASTA link at the top of the page to get the sequences in the FASTA format. Compare the local alignment of these sequences under dynamic programming. You may use the SIM tool by Expasy (<u>http://www.expasy.org/tools/sim-prot.html</u>) to align the sequences.

- (a) Note the effect of varying gap open and extend penalties on the alignment. Report the top-5 local alignment scores obtained using BLOSUM62 on the following pair of (open, extend) penalties: (12,12), (12,4), (12,1), (8,1), (4,1). Just write down the scores and the lengths of the alignments, you do not need to show the actual alignment.
- (b) Which pair of (open, extend) penalties given in part (a) corresponds to the linear gap model?
- (c) Now align the human IRGM to chimpanzee IRGM (Accession #: XP_527077.1) with the default parameters (BLOSUM62, Gap Open: 12, Gap Extend: 4). Compare this alignment with that between human and rat IRGM (using same parameters) in terms of length and score. Are the conserved sites between human and mouse also conserved between human and chimpanzee?

	-	Т	G	С	Α	Т	Т	Α	С	G	G	Α
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
Α	-1	-1	-2	-3	0	-1	-2	-3	-4	-5	-6	-7
G	-2	-2	2	1	0	-1	-2	-3	-4	-1	-2	-3
Т	-3	1	1	1	0	3						
Т	-4											
С	-5											
G	-6											
Α	-7											

Problem 2 (40 Points):

Given the partially completed dynamic programming partial scores table above obtained for global alignment of the two sequences:

- (a) What is the gap penalty (assuming that a linear gap model is used)? What are the match and mismatch scores?
- (b) Fill in the rest of the table.
- (c) Show an optimal global alignment of the sequences. Also, show the traceback path on the partial scores table.
- (d) How many different optimal alignments exist? Why?

Problem 3 (30 Points):

- (a) What is the maximum length of the DNA sequence that is expected to occur at least once in its entirety in another DNA sequence of length 30000? In other words given two DNA sequences, A and B, where length(A)=30000, what should be the maximum length of B, so that it is expected to observe a perfect semi-global alignment (all matches, no mismatch or internal gaps) between A and B? Show the steps of your calculation.
- (b) What would be the E-value of such a perfect semi-global alignment if the sequence **B** in part (a) was of length 100?