DUE: Friday Feb23rd at the BEGINNING of class.
Hand In: Answer the questions on paper, number your answers. Your work must be legible -- if your handwriting isn't great, type it up and print it.

For full credit you must identify key assumptions and provide reasoning (or show work) behind answers. Whenever possible, partial credit will be given if adequate work is shown. Remember I encourage working together, but you MUST indicate all collaborations and/or assistance received or given.

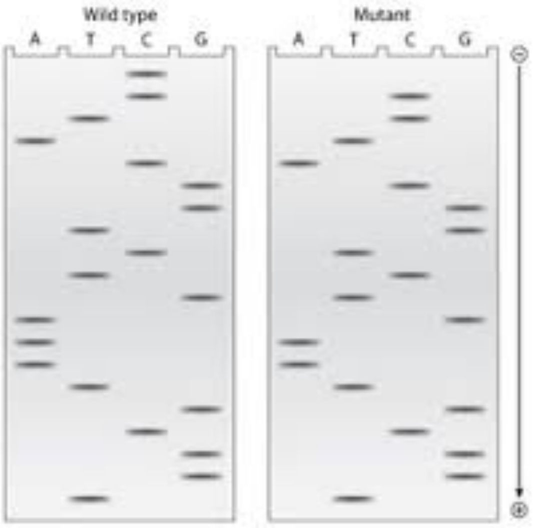
Note that showing work means that if you utilize software for assistance (programs you write or stock software such as Excel), you should indicate as such and provide sufficient details so that I can judge the work. That may mean sending me (by email) source code or associated Excel files.

Questions 1-5 are required of all sections. Questions 6 and 7 are advanced questions. Question 6 is required of Section MCDB 5520 and question 7 of section CSCI 7000. All questions are weighted equally in the overall grade. Those in the undergraduate sections may do the advanced questions for extra credit. Graduate students (5520 and 7000) may elect to do the other advanced problem for extra credit.

All problems have a maximum value of 10 points. Sub-problem values are marked when appropriate.

Questions

1. J. Deen has a family history of colon cancer consistent with hereditary non-polyposis colorectal cancer (HNPCC), an autosomal dominant form of colon cancer. Mutations in a family of genes, specifically MSH2 or MLH1, are involved in DNA repair have been linked to HNPCC. Your lab received Mr. Deen’s blood sample and has manually sequenced the MSH2 gene. The gel below shows the section of the sequence where you found a mutation. For comparison, a wildtype (known to be normal) individual is also sequenced. The two gels are as follows:



(6pt) (a) What is the mutation observed in J. Deen? How confident are you based on the gel above? What could you do to confirm this observation?

(4pt) (b) Does the mutation alter the protein sequence? How?

2. You are working to sequence and annotate a new species of bacteria recently discovered in a soil sample. After the latest round of assembly, you are specifically looking to annotate contig #18 (see Contig18.fasta on Canvas). As a first pass annotation, you plan to consider all open reading frames (ORFs).

(2pt) (1) Describe the pattern that an ORF finder looks for in bacterial sequences. For full credit, describe how many frames must be considered in your search.

(2pt) (b) Using NCBI's orf finder (https://www.ncbi.nlm.nih.gov/orffinder/) Copy and paste your DNA sequence in FASTA format into the search box. Set the minimal ORF length to 30 a.a., the genetic code to "standard", the start codon to 'ATG only', and ignore nested ORFs.

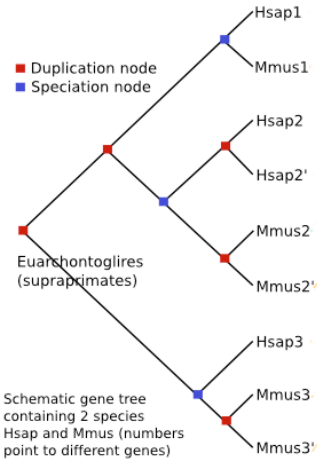
Under these settings, how many ORFs are predicted to reside in Contig18?

(2pt) (c) What is the reasoning behind ignoring nested ORFs?

[Note that you can select individual ORFs from either the list on the right or by clicking directly on the red/pink boxes in the sequence browser at the top.]
(2pt) (d) What is the longest ORF on the negative strand? (report frame, start, stop and length in amino acids)

(2pt) (e) What is the longest ORF on the positive strand? (report frame, start, stop and length in amino acids)

3. Consider the following phylogenetic tree:



(4 pt) (a) Determine whether the following gene pairs are orthologs or paralogs:

(i) Hsap3 and Mmus 1

(ii) Hsap2 and Mmus 2

(6 pt) (b) You are writing a manuscript for publication. In the latest draft, one of your co-authors has written, "Using the BLOSSUM40 matrix, we determined that our proteins are 70% homologous." What is wrong with this statement?

4. Consider the following alignment matrix:

			A		C		D		E		F
	0	→	-2	→	-4	→	-6	→	-8	→	-10
	↓	↘		↘		↘		↘			
G	-2		7	→	5		12	→	10	→	8
	↓		↓	↘		↘					
H	-4		5		9		14	→	12	→	10
	↓		↓		↓	↘		↘		↘	
I	-6		3		7		20	→	18		21
	↓		↓	↘			↓		↓	↘	
K	-8		1		12		18	→	16		24

(3 pt) (a) Write down all maximally scoring alignments for the dynamic programming matrix shown above.

(2 pt) (b) Was this DP matrix generated by the Smith-Waterman or Needleman-Wunsch algorithm? How do you know?

(2 pt) (c) For this DP matrix, is the gap penalty linear or affine? Explain and give the value(s).

(3 pt) (d) What is the scoring matrix, based on the above DP matrix. Note that you can infer some comparisons precisely whereas for others you can only infer

the bounds (i.e. score is < 0).

	A	C	D	E	F
G					
H					
I					
K					

5. Score the following protein sequence alignment:

```
RLINLMP----WVLATEYKNY
QFFPLMPPAPYWILATDFENY
```

Using:
(5 pt) (a) BLOSUM62 (ftp://ftp.ncbi.nih.gov/blast/matrices/BLOSUM62) and a linear gap penalty of -4.

(5 pt) (b) BLOSUM80 (available at: ftp://ftp.ncbi.nih.gov/blast/matrices/BLOSUM80) with affine gap penalties: gap open of -9 and gap extension of -1.

6. (Advanced, MCDB) Consider the following two protein sequences (give in Fasta format):

```
>Sulf-toko-ST0027
MFFTLSEIQLLSKRMKGFPRAISEELRGWHWNEPPLYPSSNTLLSVSDLTNGLCDSGRYVYLKHK
GIVPKVEAKIGNTIHTTYATAIETIKRLIYEHEDLDSVKLRTLMTDEFYNLKVEVIEVAKILWDH
IVSIYSAELEKARSKPFLRKDSLASLVIPFHVEYPVDGSLVGLQSALRVDAFIPILPLIAEMKTG
SYKRDHELALAGYALAFESQYEIPVDFGYLCYVNVIEGKIHNNCRLIVISDTLRQEFVEVRDRAL
RAIDDDVDPGLAKKCSADCPFLPHCKGG
>Ther_aggr-Csa1
MIRRVRGGFSTGSRAFPGFSGADDEGVLIGLETSQWLVEALILRRVMFRSIRRLYELARADPVDP
ELRGWSWDRLPLKPRAYLNLGVSEIASKYCETRRDIWLRRKTGARAEPTEPILTGRLIHDAISLA
LKETAKLLINNTEPYTAYQILSEKWRKLNPPKGYEKTVEKTYKATLITILGEAMYEKLVNETPQP
VAYSEYRVDGTPLGMSQNLSVDVISDSVIIDFKTGAPRDFHKLSITGYALALEAAYETPRDYGLL
IYINNPEDPRITYKPVYISNTLRRLFIEERDNIIDMLLEDAEPPKDLNCQPTCPLHGACNK
```

(5 pt) (a) You seek to obtain the global alignment using an affine gap penalty of -50 (gap open) and -1 (gap extension). What BLOSUM scoring matrix seems most appropriate for this alignment? Why?

(5 pt) (b) Calculate the best local alignment between the two sequences using BLOSUM80 (available at: ftp://ftp.ncbi.nih.gov/blast/matrices/BLOSUM80) with affine gap penalties: gap open of -25 and gap extension of -5.

Note that the EMBOSS suite of tools will likely be useful in this endeavor (http://www.ebi.ac.uk/Tools/emboss/).

7. (Advanced, CS) Describe how to achieve the best score by Smith-Waterman in linear space.