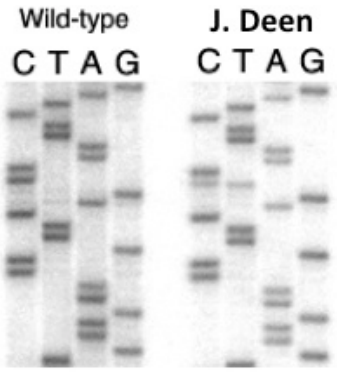
DUE: Friday Feb12th at the BEGINNING of class.
Hand In: Answer the questions on paper, number your answers, show all work (as necessary), identify key assumptions, and indicate all collaborations and assistance received or given.

Note that showing work means that if you utilize software for assistance (programs you write or stock software such as Excel or R), 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.

Your work must be legible -- if your handwriting isn't great, type it up and print it.

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.



The wild-type sequence is TGAAGAACCGTTCAGCCAATTCTAG and the protein is encoded on the reverse complement (e.g. the protein sequence is: LELAERFF).

** Note that the original posted version of this homework was incorrect in the wildtype sequence specified by the gel.

(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?

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

2. Consider the following read returned from the sequencing facility:

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
```

```
GCATGTGGTGAGGTGGTAGTGATGGTGATATAGAGTGGTAGTATAAGTGT
+
IIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIIIIIIIAIIGIICI
```

Recommendation: Refer to the Wikipedia page on FASTQ format for the encoding schemes discussed below.

(a) Assume the quality scores are encoded using the Sanger offset (Phred+33). Is this sequence of generally good quality?

(b) Under this encoding, what base is the lowest quality? (You may circle it in the above) What is the probability of this position being?

(c) You realize that you were mistaken in the encoding and it is given in the Illumina 1.3+ (Phred+64) format. Under this encoding scheme, is this sequence of generally good quality? Is the worst position still the one you circled in question b?

3. Score the following protein sequence alignment:

RLINLMP----WVLATEYKNY
QFFPLMPPAPYWILATDFENY

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

(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.

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

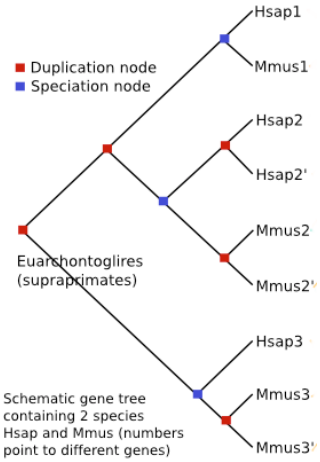
(a) Write down all maximally scoring alignments for the dynamic programming matrix shown above.

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

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

(d) What is the scoring matrix, based on the above DP matrix.

5. Consider the following phylogenetic tree:



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

(i) Hsap3 and Mmus 1

(ii) Hsap2 and Mmus 2

(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?

6. (Advanced) 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
```

(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?

(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/).