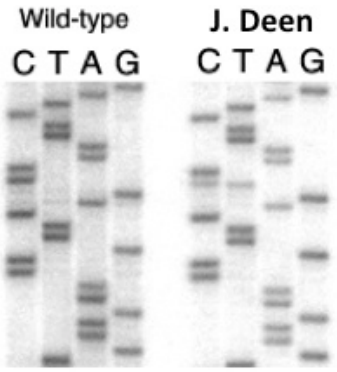
DUE: Friday Feb12th at the BEGINNING of class.
Hand In: Answer the questions on paper, number your answers, show all work (as necessary), identify key assumptions, and indicate all collaborations and assistance received or given.

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.

5 pt (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?

Reading the gel directly, J.Deen's mutated sequence is:

wt TGAAGAACCGTTCAGCCAATTCTAG

deen TGAAGAACCGTTCAGTCAATTCTAG

and its (reverse complement):

CTA GAA TTG ACT GAA CGG TTC TTC A

I'm pretty confident because the gel is reasonably clear, but in the position of the mutation the two apparent bands (one a C and the other a T) are a bit faint so I'll accept other answers if they are justified.

Most likely this individual is heterozygous at this position and the easiest way to confirm this is to sequence this fragment again.

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

The mutated sequence translates into:

LELTERFF

The C->T change (G->A on reverse complement) results in an Alanine mutating into a Threonine. Threonine is larger and hydroxylic whereas Alanine is tiny. Both are hydrophobic. Whether this mutation is causal for this disease cannot be determined from the sequencing gel alone. The disease is a dominant mutant so having it as a heterozygote is at least consistent.

The most common mistake was not properly accounting for the reverse complement.

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

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

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

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

In Phred+33 system encodings go from ASCII 33 ("!") to ASCII 73 ("I"). Hence a quality score encoded by "I" are very good (quality score 40).

3 pt (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?

The poorest quality position is the "A" which has a quality score of "A" (32).
The probability of a mistake is: $P = 10^{-32/10} = 0.0006$
Therefore the accuracy is: 99.94%

4 pt (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?

In the Phred+64 encoding, scores go from ASCII 64 (“@”) to ASCII 104 (“h”).
In this scheme, “I” is an quality score of 9, which is pretty bad.
The worst position remains the same.

3. 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.

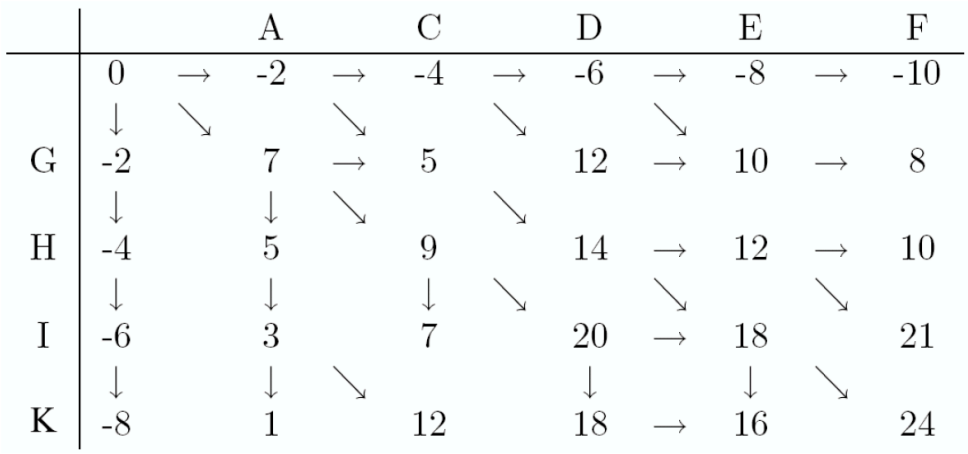
1 + 0 + 0 + -2 + 4 + 5 + 7 + -4*4 + 11 + 3 + 4 + 4 + 5 + 2 + 3 + 1 + 6 + 7 = 45

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.

1 + 0 + -1 + -4 + 6 + 9 + 12 + -9 + -1 * 4 + 16 + 4 + 6 + 7 + 8 + 2 + 4 + 1 + 9 + 11 = 78

Note that the affine gap penalty is (g + s*k) = -9 + 4*(-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

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

```
ACDEF
-GHIK
```

```
ACDEF
G-HIK
```

AHDEF
GCI-K

2 pt (b) Was this DP matrix generated by the Smith-Waterman or Needleman-Wunsch algorithm? How do you know?
Needleman-Wunsch, because Smith-Waterman never has negative scores.

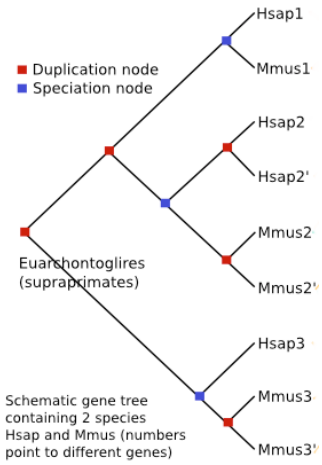
2 pt (c) For this DP matrix, is the gap penalty linear or affine? Explain and give the value(s).
Linear, which you can easily see from the uniform steps of 2 in the top row or first column. Value is -2.

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

	A	C	D	E	F
G	7	7	16	16	< 16
H	<7	2	9	< 0	< 0
I	<7	<2	11	4	9
K	<7	9	<11	< 4	6

I gave full credit for getting the blue (exactly discernable) values because the question was not clear on whether the inexact values (in red) were desired.

5. Consider the following phylogenetic tree:



4 pt (a) Determine whether the following gene pairs are orthologs or paralogs:
(i) Hsap3 and Mmus 1 paralogs, common anscestor is duplication
(ii) Hsap2 and Mmus 2 orthologs, common anscestor is speciation

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?
The obvious problem with this statement is that homology is a binary concept. Either two things descend from a common ancestor or they do not. There can

be no percentages. (Most likely they mean similarity). Knowing that you cannot have a percentage of homology was a required answer.

A less obvious possible problem is that a BLOSSUM40 matrix would be optimally tuned for sequences at 40% identity. Finding alignments at much higher identity (the 70% suggested in this statement) may indicate that they should have used a different scoring scheme. While it is perfectly possible to get a 70% identical alignment from a BLOSSUM40 matrix, it would prompt me to double check that this alignment was optimally scored (particularly if any subsequent findings of the paper depended on the alignment score or any assessment of its significance.)

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
```

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?

When you have no prior knowledge of the expected evolutionary distance between the two sequences, then you have to rely on the percent identity or similarity to determine the approximate relevant alignment distance. The gap penalty information cannot be considered informative. While there is some correlation between the gaps expected at a given evolutionary distance, the gap parameters are arbitrarily set (always). Whereas the scoring matrix reflects some amount of expectation given the evolutionary distance.

In this case, using those gap parameters, I can use Needle (an EMBOSS tool) to try a couple of different global alignment scoring matricies and see what happens. Because it is global alignment, the percent identity and percent similarity are informative about what scoring matrix is likely useful. [Note that local alignment this isn't always true because you don't always get the same region aligned for different scoring schemes.]

For example, an alignment by BLOSSUM90 shows these sequences are 0.3% identical (0.7% similar) suggesting that this scoring matrix is definitely wrong – if they really are 90% identical (which BLOSSUM90 suggests) then I should be

getting a very high percent identity/similarity.

An alignment using BLOSSUM62 shows these sequences are ~21% identical (~36% similarity), suggesting they are quite different. This scoring matrix is closer, but the scored identity/similarity is much smaller than the 62% assumed by this scoring scheme.

A realignment with a BLOSSUM30 matrix shows they are ~21% identical and ~42% similar. I note two things: the identity is the same as was determined using a BLOSSUM62. Suggesting these two sequences are about 21% identical. Second, the similarity improved relative to the BLOSSUM62 matrix. Ideally this should be scored using a BLOSSUM20, but at Needle the closest we can get is the BLOSSUM30.

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.

The output of Water (an EMBOSS tool) using BLOSSUM80 and the affine gap of -25 to open and -5 to extend gives the following alignment:

```
#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM80
# Gap_penalty: 25.0
# Extend_penalty: 5.0
#
# Length: 118
# Identity:      44/118 (37.3%)
# Similarity:    66/118 (55.9%)
# Gaps:           3/118 ( 2.5%)
# Score: 252.0
#
#
#=======================================

EMBOSS_001       163 EYPVDGSLVGLQSALRVDAFIPILPLIAEMKTGSYKRDHELALAGYALAF    212
                     ||.|||:.:|:...|.||....  .:|.:.:|||:.:..|:|::.|||||.
EMBOSS_001       200 EYRVDGTPLGMSQNLSVDVISD--SVIIDFKTGAPRDFHKLSITGYALAL    247

EMBOSS_001       213 ESQYEIPVDFGYLCYVNVIEGKIHNNCRLIVISDTLRQEFVEVRDRALRA    262
                     |:.||.|.|:|.|.|:|..|.. ....:.:.||:|||:.|:|.||..:..
EMBOSS_001       248 EAAYETPRDYGLLIYINNPEDP-RITYKPVYISNTLRRLFIEERDNIIDM    296

EMBOSS_001       263 IDDDVDPGLAKKCSADCP    280
                     :.:|.:|.....|...||
EMBOSS_001       297 LLEDAEPPKDLNCQPTCP    314
```

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