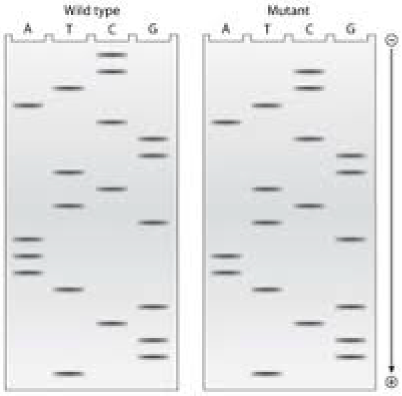
Homework 2

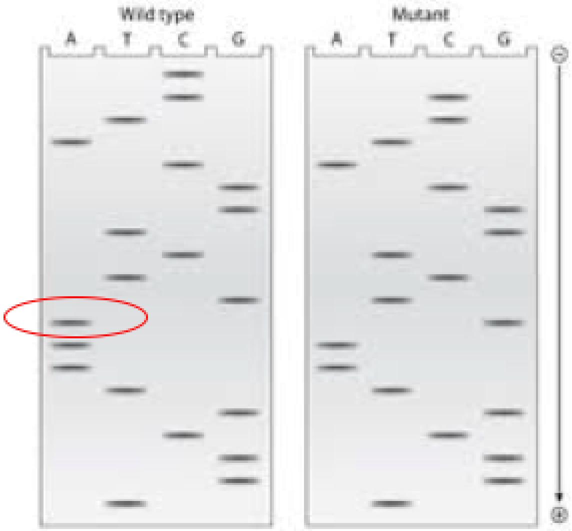
Dieu My Nguyen | MCDB 5520 | Feb 23, 2018

1. J. Deen has a family history of colon cancer consistent with hereditary nonpolyposis 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?

In this given gel, we observe a deletion mutation of nucleotide A. I can't be very confident if this is the only gel showing the results. Plus, "all sequencing strategies have some amount of errors/mistakes. These can come from humans (user error), chemistry, technology (reading wrong color), etc." (Lecture Jan 29 slide 16). Missing DNA bands may mean that the small bands were electrophoresed off the gel or that bands of similar size were not resolved. Confidence in the results can be increased by repeating the sequencing many times. We should also try other sequencing techniques: mass spectrometry can detect single nucleotides based on their mass rather than size and can detect single-nucleotide mutations in a fragment like the one we have here. And why not try other methods too: capillary and parallel sequencing.



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

This mutation does alter the protein sequene by causing a frame shift mutation. In the mutant, UAG gives rise a a STOP codon and this may cause the polypeptide chain to be shorter than the wild type.

Wildtype:
DNA: TGG CGT AAA GTC TGG CAT CC
mRNA: ACC GCA UUU CAG ACC GUA GG
Protein: Thr Ala Phe Gln Thr Val

Mutant:
DNA: TGG CGT AAG TCT GGC ATC C
mRNA: ACC GCA UUC AGA CCG UAG
Protein: Thr Ala Phe Arg Pro STOP

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) (a) 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.

An ORF finder looks for ORFs consisting of a series of codons that specify the amino acid sequence that a gene codes for. The OFR begins with a start codon (usually ATG) and ends with a termination codon (TAA, TAG, or TGA). A double-stranded DNA has 6 ORFs: 3 in 1 direction and 2 in the reverse on the complementary strand. ORF scans work well for bacterial genomes to locate most of the genes in the sequencs, because some assumptions for simple ORF scans are valid for bacterial sequencing: relatively little non-coding DNA in genome, real genes don't overlap, no genes-within-genes. (Source: https://www.ncbi.nlm.nih.gov/books/NBK21136/ (https://www.ncbi.nlm.nih.gov/books/NBK21136/))

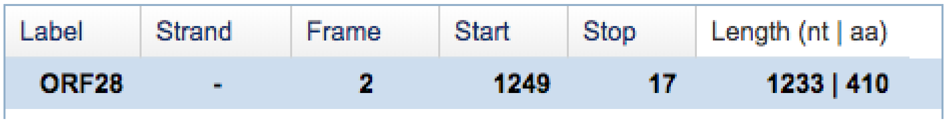
(2pt) (b) Using NCBI's orf finder (https://www.ncbi.nlm.nih.gov/orffinder/ (https://www.ncbi.nlm.nih.gov/orffinder/)). Copy and paste your DNA sequence in FASTA format into the search box. Set the minimal ORF length to 30 a.a., the genetic code to "standard", the start codon to 'ATG only', and ignore nested ORFs. Under these settings, how many ORFs are predicted to reside in Contig18?

37 ORFs.

(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.]

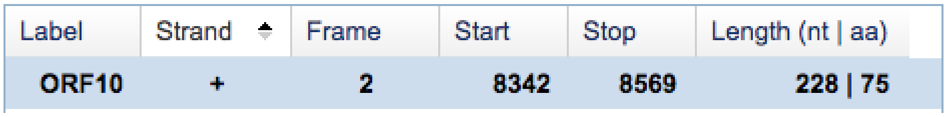
Bacterial genomes usually do not contain nested ORFs. (Source: in 2(a))

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



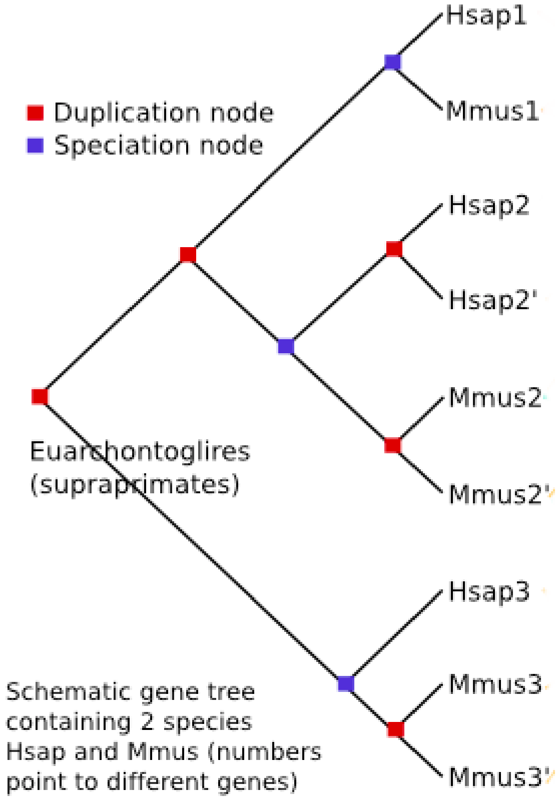
Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF28	-	2	1249	17	1233 | 410

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



Label	Strand ⬍	Frame	Start	Stop	Length (nt | aa)
ORF10	+	2	8342	8569	228 | 75

3. Consider the following phylogenetic tree:



(4 pt) (a) Determine whether the following gene pairs are orthologs or paralogs:
(i) Hsap3 and Mmus 1: Paralogs arose by duplication node Euarchontoglires.
(ii) Hsap2 and Mmus 2: Orthologs arose by speciation event.

(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?

BLOSUM40 would cluster sequences with identity of 40% or more. Moreover, BLOSUM40 is used for more divergent sequences than, for instance, BLOSUM80. Most importantly, similarity from sequence aligment does not indicate homology. Similar sequences may indicate evolutionary relationship, or not, or the converse. To make a conclusion about homology, we need to use other statistical methods: Karlin Alrshul stats, Bayes probability, classicaly stats with hypothesis testing. To assess whether an alignment indicates homology, we should know the probablity that the alignment can be expected from chance ("noise" in database).

4. Consider the following alignment matrix:

$$
\begin{array}{c|ccccccccccc}
 & & & A & & C & & D & & E & & F \\
\hline
 & 0 & \rightarrow & -2 & \rightarrow & -4 & \rightarrow & -6 & \rightarrow & -8 & \rightarrow & -10 \\
 & \downarrow & \searrow & & \searrow & & \searrow & & \searrow & & & \\
G & -2 & & 7 & \rightarrow & 5 & & 12 & \rightarrow & 10 & \rightarrow & 8 \\
 & \downarrow & & \downarrow & \searrow & & \searrow & & & & & \\
H & -4 & & 5 & & 9 & & 14 & \rightarrow & 12 & \rightarrow & 10 \\
 & \downarrow & & \downarrow & & \downarrow & \searrow & & \searrow & & \searrow & \\
I & -6 & & 3 & & 7 & & 20 & \rightarrow & 18 & & 21 \\
 & \downarrow & & \downarrow & \searrow & & & \downarrow & & \downarrow & \searrow & \\
K & -8 & & 1 & & 12 & & 18 & \rightarrow & 16 & & 24
\end{array}
$$

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

Because the matrix has negative values, we assume that it is generated by the Needleman-Wunsch algorithm. Thus, by tracing back from the lower right corner (where highest score is) to the upper left, we get the following maximally scoring alignments, in the order of top sequence, bottom sequence, score:

Alignment #1:
A C D E F
- G H I K
-2 5 14 18 24

Alignment #2:
A C D E F
G - H I K
7 5 14 18 24

Alignment #3:
A C D E F
G H I - K
7 9 20 18 24

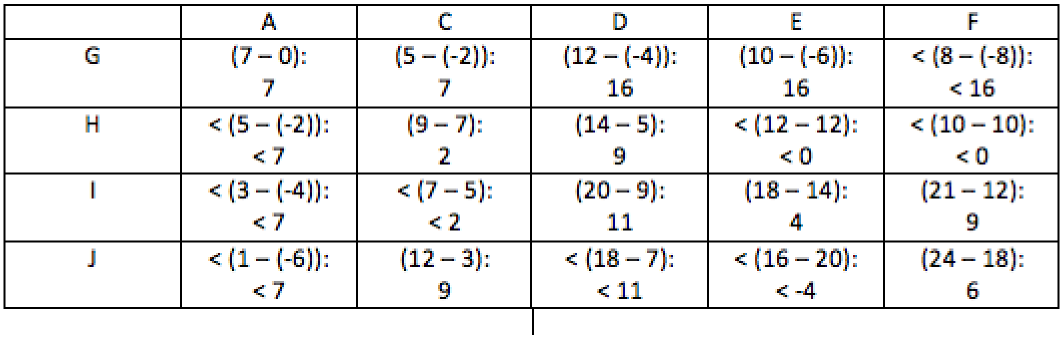
(2 pt) (b) Was this DP matrix generated by the Smith-Waterman or Needleman-Wunsch algorithm? How do you know?

Needleman-Wunsch, because in Smith-Waterman algorithm, no score is negative (and the highest value is not always at the bottom right of the matrix).

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

Linear gap penalty, by which evert gap receives a score of d = -2. We can see this in the maximally scoring alignments in 4(a). In alignment 1, the first gap shows a score of -2. In alignment 2, the gap is in the 2nd position, and we see that 7 + d = 7 + -2 = 5. Likewise, in alignment 3, the gap at the 4th position shows a reduction in the overall score by 2: 20 - 2 = 18.

(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	(7 – 0): 7	(5 – (-2)): 7	(12 – (-4)): 16	(10 – (-6)): 16	< (8 – (-8)): < 16
H	< (5 – (-2)): < 7	(9 – 7): 2	(14 – 5): 9	< (12 – 12): < 0	< (10 – 10): < 0
I	< (3 – (-4)): < 7	< (7 – 5): < 2	(20 – 9): 11	(18 – 14): 4	(21 – 12): 9
J	< (1 – (-6)): < 7	(12 – 3): 9	< (18 – 7): < 11	< (16 – 20): < -4	(24 – 18): 6

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.
1+0+0+(-2)+4+5+7+11+3+4+4+5+2+3+1+6+7 = 61
Linear model: cost of a run of k gaps is k*d: (length of gap)(-4) = 4(-4) = -16
Total score: 61 - 16 = 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+16+4+6+7+8+2+4+1+9+11 = 91
Affine gap model: cost of a run of k gaps is g+k*s: -9 + (4)(-1) = -13
Total score: 91 - 13 = 78

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?

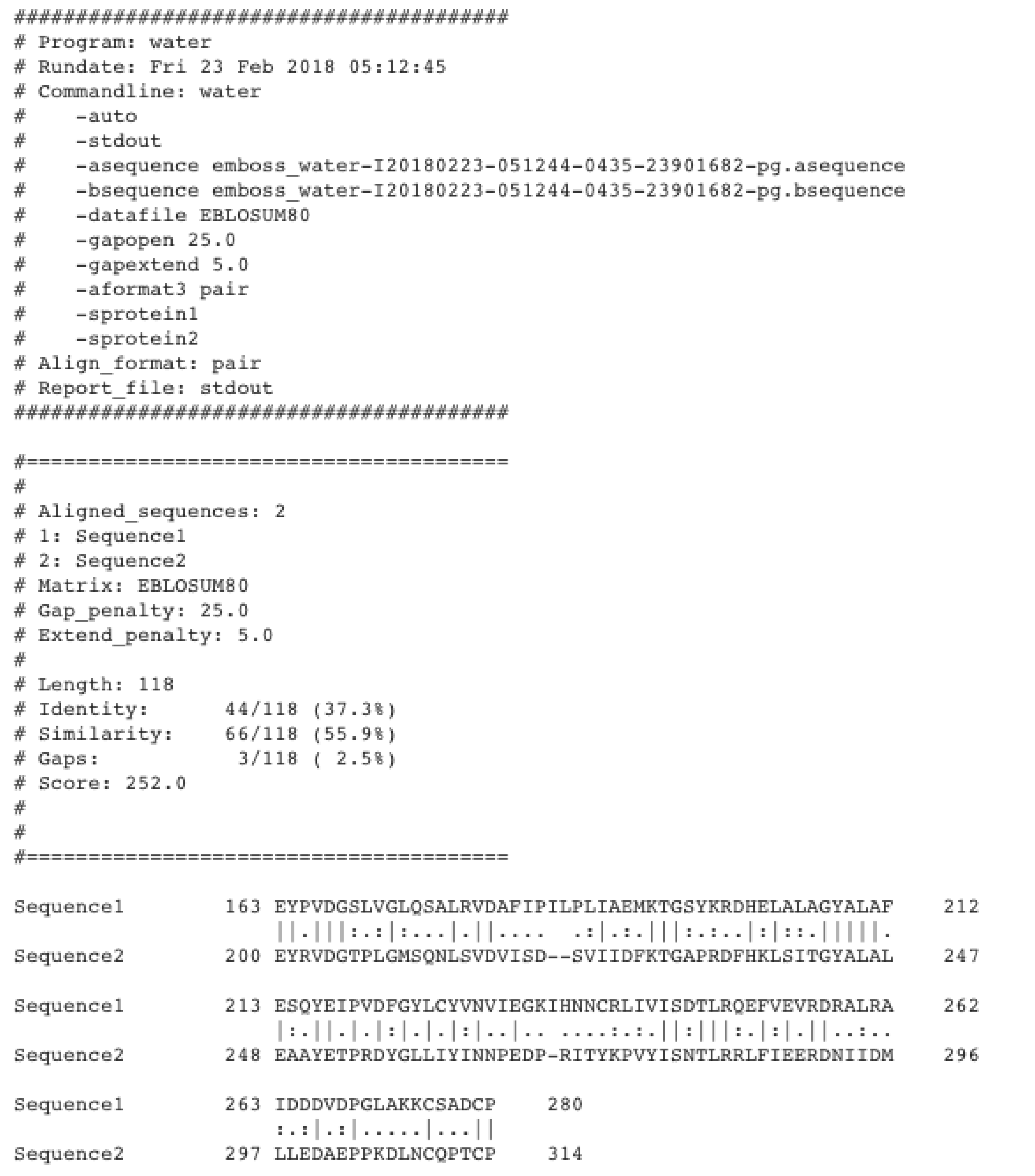
Because we don't have prior knowledge on the sequences, the standard BLOSUM62 seems like a place to start. If we knew that the sequences re closely related, we would choose matrices made for highly similar alignments (e.g. BLOSUM80), as we would for distantly related sequences. Interestingly, we want to set the gap penalty for opening to -50, quite high, indicating that a point mutation (indel) would have be impactful in reducing the overall alignment. If we already have a gap though, the cost of extending the gap is relatively low (-1). If we expect a point mutation to have such a high penalty score, we're probably assuming that the sequences are closely related and may try a BLOSUM matrix towards to high end (e.g. BLOSUM80).

Source:
http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/650/Use_scoring_matrice
(http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/650/Use_scoring_matric

(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/ (http://www.ebi.ac.uk/Tools/emboss/)).

For some reason, I couldn't input the sequence names as they are. So, Sequence1 = Sulf-toko-ST0027; Sequence2 = Ther_aggr-Csa1

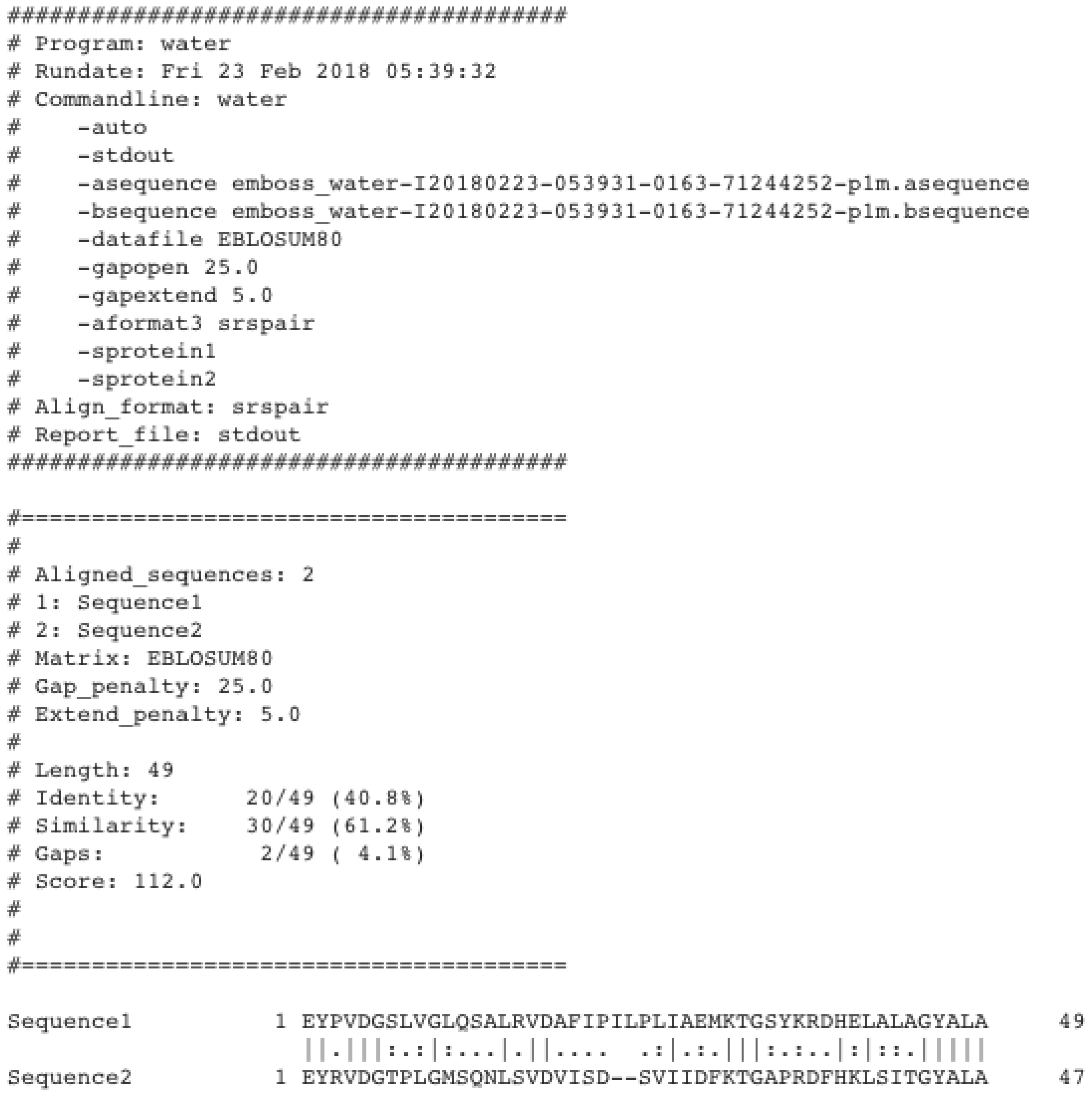
Getting best local alignments:



```
########################################
# Program: water
# Rundate: Fri 23 Feb 2018 05:12:45
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20180223-051244-0435-23901682-pg.asequence
#    -bsequence emboss_water-I20180223-051244-0435-23901682-pg.bsequence
#    -datafile EBLOSUM80
#    -gapopen 25.0
#    -gapextend 5.0
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: Sequence1
# 2: Sequence2
# 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
#
#
#=======================================

Sequence1        163 EYPVDGSLVGLQSALRVDAFIPILPLIAEMKTGSYKRDHELALAGYALAF    212
                     ||.|||:.:|:...|.||....  .:|.:.|||:.:..|:|::.|||||.
Sequence2        200 EYRVDGTPLGMSQNLSVDVISD--SVIIDFKTGAPRDFHKLSITGYALAL    247

Sequence1        213 ESQYEIPVDFGYLCYVNVIEGKIHNNCRLIVISDTLRQEFVEVRDRALRA    262
                     |:.||.|.|:|.|.|:|..|.. ....:.:.||:|||:.|:|.||..:..
Sequence2        248 EAAYETPRDYGLLIYINNPEDP-RITYKPVYISNTLRRLFIEERDNIIDM    296

Sequence1        263 IDDDVDPGLAKKCSADCP    280
                     :.:|.:|.....|...||
Sequence2        297 LLEDAEPPKDLNCQPTCP    314
```

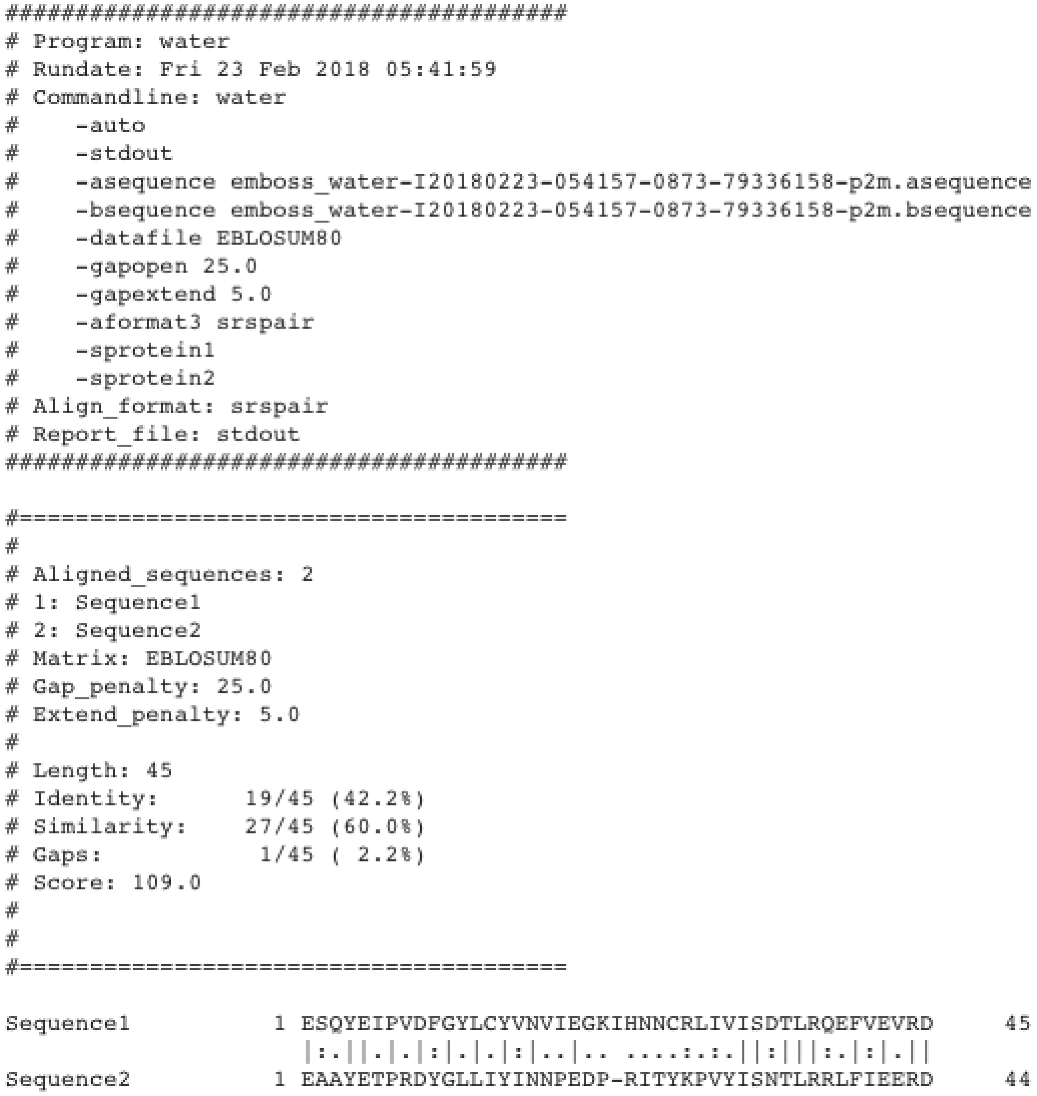
Getting scores for each of the 3 alignments, in same order as above:

(1) (Best one with highest score, 112)



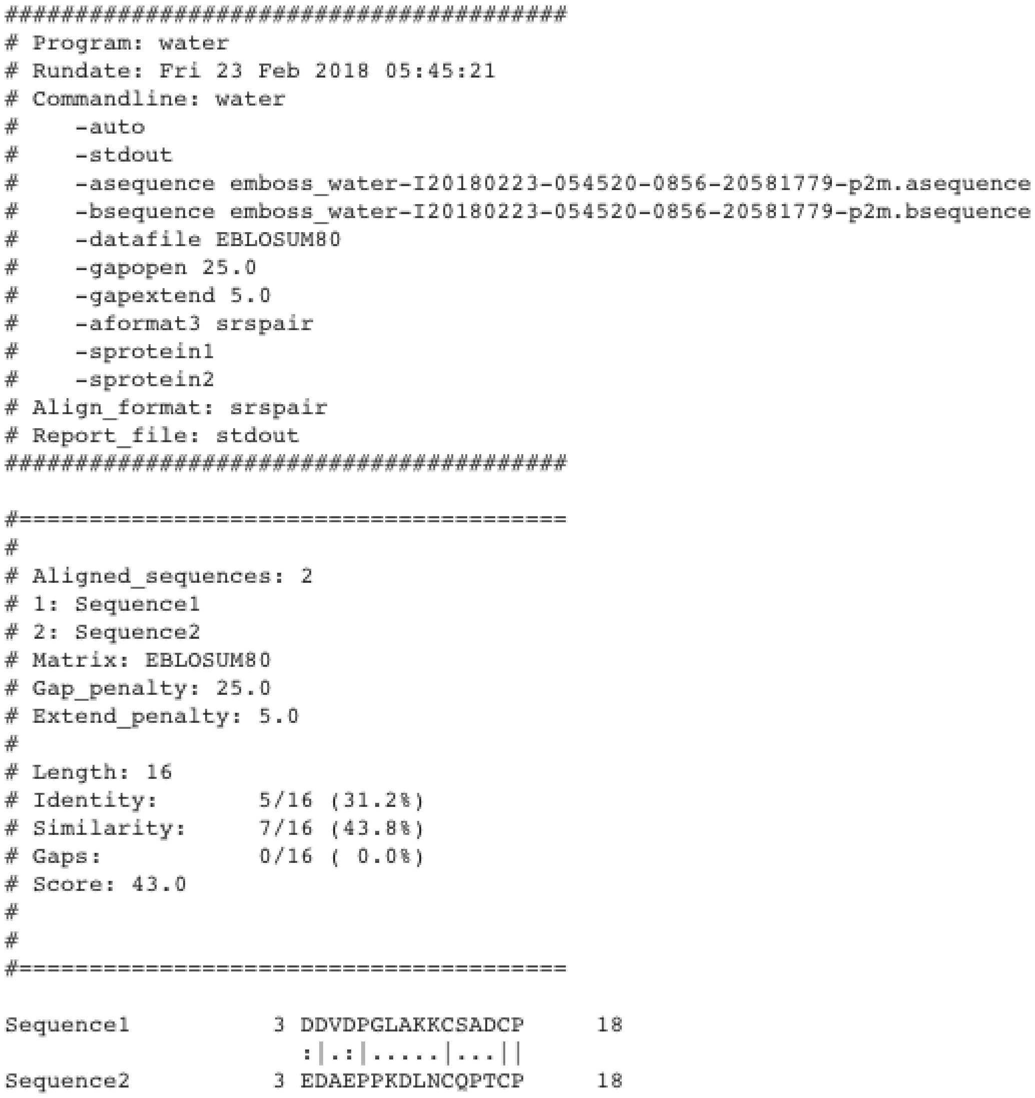
```
########################################
# Program: water
# Rundate: Fri 23 Feb 2018 05:39:32
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20180223-053931-0163-71244252-p1m.asequence
#    -bsequence emboss_water-I20180223-053931-0163-71244252-p1m.bsequence
#    -datafile EBLOSUM80
#    -gapopen 25.0
#    -gapextend 5.0
#    -aformat3 srspair
#    -sprotein1
#    -sprotein2
# Align_format: srspair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: Sequence1
# 2: Sequence2
# Matrix: EBLOSUM80
# Gap_penalty: 25.0
# Extend_penalty: 5.0
#
# Length: 49
# Identity:      20/49 (40.8%)
# Similarity:    30/49 (61.2%)
# Gaps:           2/49 ( 4.1%)
# Score: 112.0
#
#
#=======================================

Sequence1          1 EYPVDGSLVGLQSALRVDAFIPILPLIAEMKTGSYKRDHELALAGYALA     49
                     ||.|||:.:|:...|.||....  .:|.:.|||:.:..|:|::.|||||
Sequence2          1 EYRVDGTPLGMSQNLSVDVISD--SVIIDFKTGAPRDFHKLSITGYALA     47
```

(2)



```
########################################
# Program: water
# Rundate: Fri 23 Feb 2018 05:41:59
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20180223-054157-0873-79336158-p2m.asequence
#    -bsequence emboss_water-I20180223-054157-0873-79336158-p2m.bsequence
#    -datafile EBLOSUM80
#    -gapopen 25.0
#    -gapextend 5.0
#    -aformat3 srspair
#    -sprotein1
#    -sprotein2
# Align_format: srspair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: Sequence1
# 2: Sequence2
# Matrix: EBLOSUM80
# Gap_penalty: 25.0
# Extend_penalty: 5.0
#
# Length: 45
# Identity:      19/45 (42.2%)
# Similarity:    27/45 (60.0%)
# Gaps:           1/45 ( 2.2%)
# Score: 109.0
#
#
#=======================================

Sequence1          1 ESQYEIPVDFGYLCYVNVIEGKIHNNCRLIVISDTLRQEFVEVRD     45
                     |:.||.|.|:|.|.|:|..|.. ....:.:.||:|||:.|:|.||
Sequence2          1 EAAYETPRDYGLLIYINNPEDP-RITYKPVYISNTLRRLFIEERD     44
```

(3)



```
########################################
# Program: water
# Rundate: Fri 23 Feb 2018 05:45:21
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20180223-054520-0856-20581779-p2m.asequence
#    -bsequence emboss_water-I20180223-054520-0856-20581779-p2m.bsequence
#    -datafile EBLOSUM80
#    -gapopen 25.0
#    -gapextend 5.0
#    -aformat3 srspair
#    -sprotein1
#    -sprotein2
# Align_format: srspair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: Sequence1
# 2: Sequence2
# Matrix: EBLOSUM80
# Gap_penalty: 25.0
# Extend_penalty: 5.0
#
# Length: 16
# Identity:       5/16 (31.2%)
# Similarity:     7/16 (43.8%)
# Gaps:           0/16 ( 0.0%)
# Score: 43.0
#
#
#=======================================

Sequence1          3 DDVDPGLAKKCSADCP     18
                       :|.:|.....|...||
Sequence2          3 EDAEPPKDLNCQPTCP     18
```

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

For sequences A and B, the first row and column are filled with 0's. We start at the first column and fill in scores in the second column based on the scoring maxtrix and gap penalties chosen. We keep a variable that stores the matrix index of the highest score. For the third column, we don't need to know the scores of the first column, only the scores of the second column, to fill in the scores here. So we can forget about the first column. As we get scores for the third column, we update the index of the highest score if a score in the third column exceeds the current highest score. So, we need to store just 2 columns and the highest score's index. In Smith-Waterman, the best score is the max value in the table and it can be anywhere.