**Problem 1 (10pts)**

Would the solution to 1.17 in Borodovsky and Ekisheva change if you had the following additional information? There are twice as many boxes with 6-base cutters as with 4-base cutters arranged in no particular order in absent-minded researcher’s lab. If the answer is yes what are the posterior probabilities?

Yes, there being twice as many packages of 6-base cutters as 4-base cutters changes the independent probability of each base cutter, which in turn changes the posterior probability. Now,

P(4-cutters) = 0.33

P(6-cutters) = 0.67

So the new Bayes’ theorem formula can be written as

P(4-cutters | X = 5) = [P(X = 5 | 4-cutters)\*P(4-cutters)] /

[P(X = 5 | 4-cutters)\*P(4-cutters) + P(X = 5 | 6-cutters)\*P(6-cutters)]

= (0.00588 \* 0.33) / [(0.00588 \* 0.33) + (0.00143 \* 0.67)]

= 0.669

There is now only a 66.9% chance that the 4-base cutter was used.

**Problem 2 (10pts)**

Align two sequences shown below using Smith-Waterman algorithm.  
Use match score of 4, mismatch score of -4 and gap penalty score of 2.

Show:  
a) dynamic programming matrix with scores (as it shown in Figure 2.6 in Durbin et al.)  
b) trace back pointers  
c) total score

sequences:  
sequence 1:  
AGAGCTCACAA

sequence 2:  
AGTAGCTTCCAAA

Note, that you applied Needelman Wunsch algorithm to these sequences in homework 1.

\ indicates the value came from the diagonal cell

> indicates the value came from the left cell

^ indicates the value came from the above cell

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | A | G | A | G | C | T | C | A | C | A | A |
|  | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A | 0 | \4 | >2 | \4 | >0 | >0 | >0 | >0 | \4 | >0 | \4 | \4 |
| G | 0 | ^2 | \8 | >6 | \8 | >6 | >4 | >2 | >0 | \0 | ^2 | ^2 |
| T | 0 | ^0 | ^6 | \4 | ^6 | \4 | \10 | >8 | >6 | >4 | >2 | \0 |
| A | 0 | \4 | ^4 | \10 | >8 | >6 | ^8 | \6 | \12 | >10 | \8 | \6 |
| G | 0 | ^2 | \8 | ^8 | \14 | >12 | >10 | >8 | ^10 | \8 | \6 | \4 |
| C | 0 | ^0 | ^6 | ^6 | ^12 | \18 | >16 | \14 | >12 | \14 | >12 | >10 |
| T | 0 | ^0 | ^4 | ^4 | ^10 | ^16 | \22 | >20 | >18 | >16 | >14 | >12 |
| T | 0 | ^0 | ^2 | ^2 | ^8 | ^14 | \20 | \18 | \16 | \14 | \12 | \10 |
| C | 0 | ^0 | ^0 | ^0 | ^0 | \12 | ^18 | \24 | >22 | \20 | >18 | >16 |
| C | 0 | ^0 | ^0 | ^0 | ^0 | ^10 | ^16 | \22 | \20 | \26 | >24 | >22 |
| A | 0 | \4 | >2 | \4 | >2 | ^8 | ^14 | ^20 | \26 | >24 | \30 | \28 |
| A | 0 | \4 | >2 | \6 | >4 | ^6 | ^12 | ^18 | \24 | \22 | \28 | \34 |
| A | 0 | \4 | >2 | \6 | >4 | ^4 | ^10 | ^16 | \22 | \20 | \26 | \32 |

Total score: 34

Alignment:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | G | - | A | G | C | - | T | C | A | C | A | A |
| A | G | T | A | G | C | T | T | C | - | C | A | A |

**Problem 3 (40 pts)**

For an alignment shown below derive log-odds score for the column 1 only by

1. Computing observed probability values

qTA = fTA / fTA + fTT

= 4 / 4 + 6 = 0.4

qTT = fTT / fTA + fTT

= 6 / 4 + 6 = 0.6

1. Expected probability values

eTA = 2\*pT\*pA

= 2 \* 0.8 \* 0.2 = 0.32

eTT = pT\*pT

= 0.8 \* 0.8 = 0.64

1. Log-odds score

sTA = log2(qTA/eTA)

= log2(0.4/0.32) = 0.3219

Adjusted for BLOSUM = sTA \* 2 = 1 (rounded to nearest integer)

sTT = log2(qTT/eTT)

= log2(0.6/0.64) = -0.0931

Adjusted for BLOSUM = sTT \* 2 = -1 (rounded to nearest integer)

TAGCTT

AAGCTC

T-GGTT

TGGCAT

TACCTT

**Problem 4 (30pts)**

Using 1st order models shown in homework 3 solution determine the probability of coding region in frame two for DNA fragment AGTAGCTTCCAG. Use only parameters provided in the homework solution (posted in course content). Show all of your work.

Work can be found in the attached excel file. The probability of a coding region in frame two is 0.00874.

**Problem 5 (60 pts)**

For the zero order hidden Markov model defined in homework 3 determine probability of coding state at the last nucleotide of the sequence AGTAG. Use parameters provided in the homework solution (posted in course content). Show all of your work.

Work can be found in attached excel file. Probability that last nucleotide is in coding state is 3.44 x 10-5.

**Problem 6 (10pts)**

p-value and e-value are used to assess the significance of the alignment. Can you think of additional ways of evaluating the strength of the alignment **other than bit Score**?

The relative entropy, H(q2||p), measures the difference between two distributions, q2 ( q x q) and p. As a result, we can determine the amount of information to expect from each aligned base pair in the alignment. Other entropy measures, like Shannon entropy, also show the variability among the aligned sequences, where a high entropy indicates more variability. In addition, in progressive multiple alignment, Feng and Doolittle created a measure of distance among the sequences in alignment. This measure is based on alignment scores, but can show us the evolutionary distances of the sequences in a linear fashion. Another measure of the strength of an alignment is the posterior distribution. This is a combination of the prior and likelihood distributions, demonstrating the information that is contained in the alignment. This distribution can give us an idea of which alignment is best out of all possible alignments for the sequences.

**Problem 7 (30 pts)**

Provide **detailed** reasoning for the following statement. For comparison/alignment of closely related species BLOSUM80 is a better choice than BLOSUM62.

The substitution matrices are made by taking a block of sequences that have been aligned, deriving the log-odds score, as seen in problem 3, and adjusting the final result to get an integer. The general practice when making these matrices, is to name them with the percent identity of the alignment. Therefore, BLOSUM80 was derived from an alignment block that had 80% identity, while BLOSUM62 was derived from a block that had 62% identity. These identities are an indication of relative divergence, where a higher identity means a more recent divergence, and thus the sequences are more closely related. Using BLOSUM80 to score the alignment of more closely related species provides a more critical scoring system such that we can narrow in on the differences among the sequences. If BLOSUM62 were used on closely related species, the slight differences wouldn’t be as noticeable. For example, a non-polar aromatic amino acid that has mutated to another of the same type would not be scored as harshly using BLOSUM62, making it difficult to tell from the score that there was a mutation at all. However, with BLOSUM80, the score would easily reflect that mutation compared to the amino acid being conserved.

**Problem 8 (10pts)**

If you are given an aligned block of sequences to derive model parameters for HMM profile from observed counts. What is(are)possible limitation(s) of this approach? What would you suggest as solution(s)?

One limitation would be the closeness of the sequences that have been aligned. These sequences could be relatively distant, resulting in parameters that are not accurate for what you may be trying to do, especially if you are wanting to observe a particular species or protein. To solve this issue, you could create the phylogenetic tree and evaluate the distances in an effort to remove any sequences that are too distant.