UBC, CPSC 445: Algorithms for Bioinformatics (Winter 2 2014/2015)

Assignment 4

Due: March 19, 2015 (at the beginning of the class)

1. Design an HMM that generates sequences of length between 1 and n with a uniform distribution, i.e., the probability that a sequence of length i, where $1 \leq i \leq n$, gets generated is $1/n$. Your HMM should have a Begin state, an End state and n additional states ($n+2$ states all together). Specify the transition probabilities (no need to specify emission probabilities since we are not concerned about what sequences the HMM generates, just about the length of those sequences).

2. Design a pair HMM for pairwise *semi-global* alignment (i.e., alignment with "overlap matches") which does not allow a gap in one sequence directly after a gap in other sequence. Remember to specify transition probabilities (in terms of δ, ϵ, τ and η) and emission probabilities (in terms of q_a's and p_{ab}'s). To specify the transition probabilities it's recommended that you draw a state transition diagram (as we did for global and local pairwise alignment).

3. (a) Implement the *Viterbi* algorithm for *Pair HMM* for the local pairwise alignment (*LPA*). You can either modify the Pair HMM so that it does not use silent states or pay attention in your implementation in which order the states are computed and that i and j do not move on silent states. Use the log space for the Viterbi matrix. The parameters used to specify the transition probabilities are

$$\delta = 0.08 \qquad \epsilon = 0.35 \qquad \tau = 0.002 \qquad \eta = 0.12$$

   The q_a and p_{ab} probabilities used to specify the emission probabilities are based on BLOSUM62 matrix: qp.txt. You also need to implement the traceback procedure that will find a Viterbi path (and use it reconstruct an optimal local alignment). Use a programming language of your choice. Submit your code through Connect.

   Make sure that your code works correctly before proceeding to part (b).

   (b) (Same as Question 1(b) of Assignment 2.) Use your program to find the best 3 matches (local alignments) between the sequence number 1000 (ZN768_HUMAN) and the first 999 sequences in the list from Question 4 of Assignment 1 (uniprot list).

   Your answer should consists of 3 top ranking matches (you don't need to consider suboptimal alignments, i.e., for each of the 999 sequences consider only one optimal alignment to the query sequence number 1000). For each match provide the following information:

   - the name of the protein from uniprot list (for example, 1A1L1_HUMAN),
   - *starting positions* of the match in this protein and in the query sequence number 1000,
   - the *length* of the alignment,
   - the natural log of the Viterbi path ($\ln \mathbf{Pr}(x, y, \pi^*)$), and
   - the first **60** characters of *the alignment.*

Notes.

- **Important notes on initialization.** The initialization in the textbook is *incorrect.* According to the text book $v_k(i,0) = v_k(0,j)$ are initialized to 0. However, this is not correct, since there is a path in the pair HMM (for the global pairwise alignment) that would generate a sequence of length i and the empty sequence: start at the Begin state, transition to the X state and loop there. For the pair HMM for the global pairwise alignment this is only such a path, hence, $v_X(i,0)$ should be equal to $\delta\epsilon^{i-1}\prod_{t=1}^{i} q_{x_t}$ which is not 0. This is the reason why the initialization in my lecture notes looks more complicated (but more correct). The idea is to initialize $v_k(0,0)$ for every k, and let the recurrence compute the correct values for all $v_k(i,j)$ including $v_k(i,0)$ and $v_k(0,j)$. However, if the recurrence tries to query $v_k(i,-1)$ or $v_k(-1,j)$, such a path is not possible, and should be ignored. For example, consider $v_Y(i,0)$. According to recurrences, this is equal to $q_{y_0}\max\{v_M(i,-1)\delta, v_Y(i,-1)\epsilon\}$. This would seems as a mistake since y_0 does not exist (recall $y = y_1 \dots y_m$), but since both $v_M(i,-1) = v_Y(i,-1) = 0$, the max is 0 and the product is then 0 as well, not matter what y_0 is. Hence, the recurrence should compute that $v_Y(i,0) = 0$.

Second, with the pair HMM for the *local pairwise alignment*, it's not correct to initialize $v_B(0,0) = 1$ and $v_k(0,0) = 0$ for all $k \neq B$. This is because this model (unlike the pair HMM for the global pairwise alignment) has silent states. For example, it's possible to get to the silent state between RX_1 and RY_1 (let's call it S_1) without generating any output (to any sequence) with probability η. Hence, $v_{S_1}(0,0) = \eta \neq 0$. *Recommended solution how to deal with it*: initialize only the begin state B: $v_B(0,0) = 1$ and $v_B(i,j) = 0$ for every $(i,j) \neq (0,0)$. Let the recurrence compute all remaining $v_k(i,j)$ including $v_k(0,0)$ for every $k \neq B$.

- Remember that with the silent states, the order in which the states are computed (for each (i,j)) matters. Make sure that your order is as follows: it starts with all non-silent states $(RX_1, RY_1, M, X, Y, RX_2, RY_2)$ followed by silent states ordered so that there is no transition from any silent state to a preceding silent state in this order.
- The complete list of residues can be found at wiki. Note that letter U represents selenocysteine, so in your analysis you can replace it with cysteine C.
- **Sample input and output:** 3 best matches between the sequence number 100 (`BMP6_HUMAN`) and the first 50 sequences in the above list of proteins from the uniprot website (only the first 60 columns of each alignment are shown):

```
Index=13 Name=ACHB_HUMAN ln Pr=-3117.485
START POS in ACHB_HUMAN: 63  START POS in BMP6_HUMAN: 273  LENGTH: 133
LISLNEKDEEMSTKVYLDLEWTDYRLSWDPAEHDGIDSLRITAES-VWLPDVVLLNNNDG
LISIYQVLQEHQHR-DSDLFLLDTRVVW--ASEEGWLEFDITATSNLW---VVTPQHNMG

Index=32 Name=AL1A1_HUMAN ln Pr=-3120.289
START POS in AL1A1_HUMAN: 320  START POS in BMP6_HUMAN: 107  LENGTH: 54
VRRSVERAKKYIL--GNPLTPG-VTQGPQIDKEQYDKI-LDLIESGKKEGAKLE
LRQQEEQQQQQQLPRGEP-PPGRLKSAPLFMLDLYNALSADNDEDGASEGERQQ

Index=1 Name=1A1L1_HUMAN ln Pr=-3122.361
START POS in 1A1L1_HUMAN: 328  START POS in BMP6_HUMAN: 64  LENGTH: 81
SGLRFGTLYT-ENQDVATAVASLC----RYHGLSGLVQYQMAQLLRDRDWINQVYLPENH
SGFLYRRLKTQEKREMQKEILSVLGLPHRPRPLHGLQQPQPPALRQQEEQQQQQQLPRGE
```

4. Consider three DNA sequences ACT, ATC and ACC. Find two alignments of these sequences, m and m', such that m is better than m' when considering the SP (Sum of Pairs) scoring model and m' is better than m when considering the Minimum Entropy scoring model (in other words, $S_{\text{SP}}(m) > S_{\text{SP}}(m')$, since in the SP scoring model, we are trying to maximize scores, and $S_{\text{ME}}(m) > S_{\text{ME}}(m')$, since in the Minimum Entropy scoring model, we are trying to minimize scores).

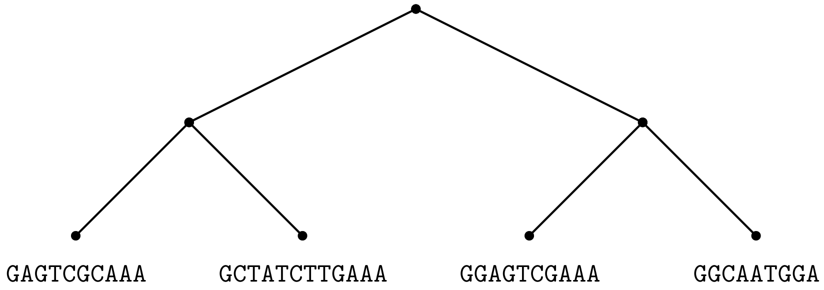
For the SP scoring model assume matches are scored by +3, mismatches by -2 and gaps by $-d$, where $d = 1$. For the Minimum Entropy scoring model assume the gap symbol $(-)$ is one of the regular symbols, when calculating probabilities of symbols in each column of the alignment.

Remark. Neither m nor m' need to be optimal alignments. Recall that $s(-,-) = 0$.

5. Use the Feng-Doolittle method to find an MSA (multiple sequence alignment) of the following 4 sequences:

```
GAGTCGCAAA
GCTATCTTGAAA
GGAGTCGAAA
GGCAATGGA
```

Use the following guide tree:



and to find an alignment of two sequences use the following tool: NW alignment tool with X as a neutral symbol. (Energy model: match score +3, mismatch score -2 and linear gap penalty with $d = 1.5$.)

The reported final MSA should not contain any X's. What is the SP score of the MSA (using the above energy model)? Is this alignment optimal?