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

Assignment 2

Due: February 12, 2015 (at the beginning of the class)

1. (a) Implement the *Smith-Waterman* algorithm for local pairwise alignment with the **linear** gap penalty function. For scoring use the similarity matrix BLOSUM62 available at BLOSUM62.txt and the linear gap penalty function with constant $d = 4$. You also need to implement the traceback procedure that will find an optimal 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) Use your program to find the best 3 matches (local alignments) between the sequence number 1000 (`ZSC25_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 alignment score, and
   - *the alignment.*

   Notes.

   - 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 (`BMR1A_HUMAN`) and the first 50 sequences in the above list of proteins from the uniprot website:

     `TBA`

2. Find (local alignment) matches of the query sequence $MAAALIRRLLRG$ and the following sequence s:

   `SRLHMMVRRMGRVPGIKFSKEKTTWVDVVNRRLVVEKCGSTPSDTSSEDGVRRIVHLYTTSDDF`

   using the BLAST algorithm as follows:

   (a) Find all words in the query sequence.

   (b) Find all neighborhood words scoring at least $T = 13$ (using BLOSUM62 matrix).

   (c) Find all hits of neighborhood words in s.

   (d) Extend each hit in both directions (first extend in the right direction and then the left direction). Stop extending if the score drops by 5 (or more) from the last maximal score.

   (e) Report extended hits with score at least 16.

   Hint. You should be able to find 10 words, 10 neighborhood words (in total), 3 hits, 3 extendeds hits of lengths 7, 6 and 4, respectively, and only 2 of them should meet the score requirement of 16.

3. Consider a dishonest casino with 98 fair dice, one loaded dice (with the probabilities as in the texbook: $p_6 = 1/2$, $p_1 = \cdots = p_5 = 1/10$) and one two-six dice that has two six faces and no five face (and otherwise it's fair). Assume that casino picked a dice from 100 dice at random and then rolled $D = 6, 4, 6, 1$ with this dice.

   (a) Calculate the likelihood probability $\mathbf{Pr}(D|M)$ for each model M (Fair, Loaded or Two-Six).

   (b) Determine which dice was most likely picked using the MLE method (this method does not use the prior knowledge we have about the models).

   (c) Calculate the posterior probability $\mathbf{Pr}(M|D)$ for each model M.

   (d) Determine which dice was most likely picked using the MAP Bayesian method.

   (e) Use the probability distribution of three models to determine the probability of observing $D' = 6, 5$ after seeing outcomes D (assuming the dice was not switched). Compare it to the probabilities of observing D' based on MLE and MAP estimators, i.e., models determined by these methods in part (b) and (d).

4. Consider a Markov chain without an end state. Show that the sum of the probabilities of all possible sequences of length L:

$$\sum_{x_1, x_2, \ldots, x_L} \Pr(x_1 x_2 \ldots x_L) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_L} \Pr(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}$$

   is equal to 1.

   Hint. Use induction on L. Recall what is the sum $\sum_j a_{ij}$ equal to?

5. Use the data for the '+' and '−' models for CpG islands available in the textbook on pages 51–52 to determine which of the following sequences are likely parts of CpG islands:

   (a) `tccgttccgt`

   (b) `accatcctcg`

   (c) `cggcaaaatc`

Bonus. Consider a dishonest casino with **nine** fair dice (with probability of each outcome 1/6) and **one** loaded dice with unknown probabilities (p_i for outcome i, for every $i = 1, \ldots, 6$). Assume that before each roll, casino randomly chooses one the ten dice.

   (a) Given observed sequence of rolls $D = d_1, \ldots, d_N$, which parameters $\Theta = p_1, \ldots, p_6$ maximize likelihood of observed data D? That is what is Θ_{MLE}?

   (b) Obviously, Θ_{MLE} does not correspond directly to the relative frequencies of outcomes in the data. Can you come with a simple way how to correctly guess Θ_{MLE} for problems like this?

   Hints. The correct answer will make use of N and $n_1, \ldots, n_6$, where n_i is the number of occurrences of outcome i in sequence D. Follow the Lagrange multiplier method from the lecture notes (remember to take the ln of the probability to simplify the derivatives).