

Spring 2008
10-810 / MSCBIO2070: Computational Genomics
Problem Set 1: Biological Sequence Analysis
Due date: Feb. 5, 2008, before class

The contact TA for this assignment is Jacob Joseph (jmjoseph@andrew.cmu.edu). Jacob's office hours are at Mellon Institute 646C, on Wednesdays, 4-5pm, or by appointment.

Where implementation is required, supply source code in Python, Java, MatLab, C, C++, or R. Please include any auxiliary files and short description of how to execute your code. Submission of this source code may be by email to the contact TA.

Collaboration is permitted, but solutions must be completed individually. Refer to the course website for complete policies.

1. Consider a genome that can be divided into GC-poor (state A) and GC-rich (state B) fragments for which the GC content is 33% and 67%, respectively. Suppose that state A transitions to state B with probability 0.005 and state B transitions to state A with probability 0.01. Assuming that the two states have equal initial probabilities, implement the forward-backward algorithm, the Viterbi algorithm, and posterior decoding. Use these to compute the likelihood of every 50-bp subsequence of the genomic sequence posted to the course website. Estimate the hidden state probabilities using the Viterbi algorithm and posterior decoding. Find all segments for which the posterior decoding probability of state B is greater than 50%.
2. (a) Read the MEME and AlignACE papers and discuss the similarities and differences between the methods (e.g., algorithms, underlying models, etc.).
 - Bailey and Elkan. *The value of prior knowledge in discovering motifs with MEME*. Proc Int Conf Intell Syst Mol Biol (1995) 3:21-29.
<http://www.sdsc.edu/~tbailey/papers/ismb95.pdf>
 - Roth, Hughes, Estep, and Church. *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantization*. Nat Biotechnol (1998) 16:939-945.
<http://www.nature.com/nbt/journal/v16/n10/pdf/nbt1098-939.pdf>(b) Implement the MEME algorithm for motif detection and run it on the set of sequences posted to the course schedule.
3. (a) Locally align the words "CABERNET" and "CARMENERE" using dynamic programming. Use the parameters $M = 2$ (match), $m = -2$ (mismatch), and $g = -1$ (gap). Present all optimal alignments. Please include the full dynamic programming matrix,

and highlight the traceback of each optimal alignment. Provide all optimal alignments. What is the longest alignment?

- (b) In a setting the parameters for a local alignment, describe why must $2g < m < M$.
4. (a) Global alignment of two sequences, A and B may be accomplished by completing a distance matrix D , as:

$$D[i, j] = \min \begin{cases} D[i-1, j] + 1 \\ D[i-1, j-1] \\ D[i, j-1] + 1 \end{cases}$$

Generalize this to align three sequences. Describe the time complexity.

- (b) Several gap scoring functions were introduced in class. Consider an alignment model in which each insertion or deletion incurs a cost of one unit, but a maximum cost of k units may be incurred for $\geq k$ consecutive insertions or $\geq k$ deletions. Provide an algorithm to calculate the minimum edit distance under this model between two strings in $O(mn)$ time, independent of k .
5. (a) Consider the below distances between six taxa. Which of the Neighbor Joining or UPGMA algorithms is best suited to construction of a rooted tree from this distance matrix? Why? Iterate this algorithm, showing the intermediate distance matrices. Draw the resulting tree.

	B	C	D	E	F
A	16	16	16	16	2
B		14	14	14	16
C			10	10	16
D				10	16
E					16

- (b) Consider the below distance matrix for three taxa. Provide and compare the resultant trees from Neighbor Joining and UPGMA. Explain which is most correct.

	B	C
A	4	7
B		5

- (c) Neighbor Joining produces an unrooted tree. How may such a tree be rooted?