

Homework Assignment 3, Due Monday October 6, 2008

CSCI 6490/4490 Algorithms for Computational Biology

September 24, 2008

1. **Written assignment:** Problem 6.20. Do the first two sub-problems.

Additional readings on global alignment, local alignment, and dynamic programming:

www.uga.edu/RNA-Informatics/Readings/background/Needleman-Wunsch1970.pdf

www.uga.edu/RNA-Informatics/Readings/background/SmithWaterman1981.pdf

www.uga.edu/RNA-Informatics/Readings/background/HowDPWorks.pdf

2. **Implementation:** Implementing a heuristic algorithm (*center-star*) for multiple sequence alignment. The idea of this algorithm is summarized in the following:

- Input: k sequences S_1, \dots, S_m , each of length at most n ;
- Find a *center sequence*: apply the global pairwise alignment algorithm on every pair of S_i, S_j of sequences. Let $Score(i, j)$ be the optimal alignment score between S_i and S_j . Then sequence S_c is designated as the center sequence if the summation of scores

$$\sum_{j \neq c} Score(c, j) \geq \sum_{j \neq i} Score(i, j) \text{ for } i = 1, \dots, m$$

- Multiple alignment: then for $i = 1, \dots, m, i \neq c$, sequence S_i is aligned to the center sequence S_c progressively. Note that the center sequence S_c is also progressively changed due to its alignments with sequences S_1, \dots, S_{i-1} . The strategy for gapping is “once gap, forever gap”. That is, if S_c , as a result its previous alignments with S_1, \dots, S_{i-1} , has a gap in position p , then the alignment between S_c and S_i should have a gap in the corresponding position for both sequences. On the other hand, if in this new alignment between S_c and S_i , there is a new gap inserted into sequence S_c , then the gap should also be inserted into the corresponding position in the previous alignments between S_c and $S_k, k = 1, \dots, i - 1$.
- Output: a center sequence S_c and a multiple sequence alignment for S_1, \dots, S_m .

You are required to find a center sequence only. This task asks you to be familiar with the global alignment algorithm and its implementation. You can choose to align nucleic acid DNA or protein sequences. PAM 250 and BLOSUM 62 are recommended for protein sequence alignment.

You will get bonus credits for completing both the center sequence and the multiple alignment tasks.