CSI 5126 Algorithms in Bioinformatics — Fall 2010
Computer Science — University of Ottawa

Homework Assignment #3 (100 points, weight 15%)
Due: Tuesday, November 30

Genome Alignment and Phylogeny Reconstruction

1. (20 marks) (genome alignment) Consider two sequences $S[1..n]$ and $T[1..m]$. We aim to find the set of all strings P such that (1) P appears exactly once in S, (2) P appears exactly once in T and its reverse complement, (3) P is maximal. Give a detailed (efficient) algorithm to find the set P. Analyse its running time.

2. (40 marks) (multiple sequence alignment)
   Consider the following set of sequences.

$$\begin{aligned} S_1 &= \texttt{ACTCTCGATC} \\ S_2 &= \texttt{ACTTCGATC} \\ S_3 &= \texttt{ACTCTCTATC} \\ S_4 &= \texttt{ACTCTCTAATC} \end{aligned}$$

   Use a matching score of +1 and mismatch/indel score of -1.

   (a) (10 marks) Compute the optimal **pairwise** global alignment for each pair of strings from the above list. You don't have to run the algorithm step by step, and may write the answer directly when obvious.

   (b) (10 marks) Compute the multiple sequence alignment for $\{S_1, S_2, S_3, S_4\}$ using the **center star method**. Please show your steps.

   (c) (20 marks) Compute the multiple sequence alignment for $\{S_1, S_2, S_3, S_4\}$ using **ClustalW**. Please show your steps.

3. (20 marks) For the following M, construct the corresponding ultrametric tree. Show each step.

M	S_1	S_2	S_3	S_4	S_5
S_1	0	20	20	20	8
S_2		0	16	16	20
S_3			0	10	20
S_4				0	20
S_5					0

4. (20 marks) Compute a phylogenetic tree for the following set of DNA sequences such that the total cost of your phylogenetic tree is at most twice that of the optimal solution (hint: use

the approximation algorithm given in section 7.2.1.2):

$$
\begin{aligned}
S_1 &= \texttt{ACCGT} \\
S_2 &= \texttt{ACGTT} \\
S_3 &= \texttt{CCGTA} \\
S_4 &= \texttt{GTCCT} \\
S_5 &= \texttt{AGCTT}
\end{aligned}
$$