Structural Bioinformatics 2004-05 Semester B Assignment 2.

Due to Mar 15

1. Consider that only the optimal sequence alignment score has to be computed, i.e. we are not interested in the alignment itself. Show how to achieve this using only $O(min(n, m))$ space.

2. Compute the number of different optimal alignments (for simplicity consider only global alignment). Notice that there is not necessary only one optimal alignment. A different traceback paths correspond to different alignments. Give an algorithm that runs in $O(nm)$ time.

3. Definition: if t is a **subsequence** of s then s is a **supersequence** of t. Problem: Design an algorithm to find the shortest common supersequence between two given sequences.

4. Input: (1) two sequences, (2) amino acid substitution matrix $M(a,b)$, (3) secondary structure assignment for each protein sequence, i.e. $SSE(s[i]) \in \{Helix, Strand, Other\}$, (4) Score and gap penalty for alignment between secondary structure elements:

   $MSSE(SSE(a), SSE(b)) =$

   $$\begin{cases} 1 & SSE(a) == SSE(b)\ AND\ (a \neq' -'\ AND\ b \neq' -') \\ 0 & SSE(a) \neq SSE(b)\ AND\ (a \neq' -'\ AND\ b \neq' -') \\ -2 & (a\ OR\ b ==' -')\ AND\ (SSE(a)\ OR\ SSE(b) == Helix\ OR\ Strand)\ AND \\ & ('-'\ is\ inside\ of\ Helix/Strand) \\ -1 & otherwise \end{cases}$$

   Devise dynamic programming algorithm that creates alignment (s_1^*, s_2^*) that maximizes $\sum M(s_1^*[i], s_2^*[i]) + \sum MSSE(s_1^*[i], s_2^*[i])$.

5. A RNA folding problem or prediction of RNA secondary (planar) structure. Let S be a string of n characters over the RNA alphabet A,C,U,G. We define a *pairing* as a set of disjoint pairs of characters in S, only (A,U) and (C,G) pairs are allowed. If we draw S as a circular string (on the plane), we define a *nested* pairing where each pairing is connected by a segment (hydrogen bond) inside the circle, and where the segments do not cross each other. Find the size of the largest nested pairing in $O(n^3)$ time. Hint: use dynamic programming.