Algorithms for Biological Sequence Analysis (Midterm # 1)

Instructor: Kun–Mao Chao TA: Hung-Lung Wang

November 7, 2006

We are given two sequences $A = \langle a_1, a_2, \ldots, a_M \rangle$ and $B = \langle b_1, b_2, \ldots, b_N \rangle$. An *alignment* of A and B is obtained by introducing dashes into the two sequences such that the lengths of the two resulting sequences are identical and no column contains two dashes. Let Σ denote the input symbol alphabet. A score $\sigma(a, b)$ is defined for each $(a, b) \in \Sigma \times \Sigma$. The score of an alignment is the sum of σ scores of all columns with no dashes minus the penalties of the gaps.

Problem 1 (60%): In this problem, we employ a simple scoring scheme where each gap symbol is penalized by a nonnegative constant β. Let $S(i, j)$ denote the score of an optimal alignment between $\langle a_1, a_2, \ldots, a_i \rangle$ and $\langle b_1, b_2, \ldots, b_j \rangle$. With proper initializations, $S(i, j)$ can be computed by the following recurrences:

$$S(i,j) = \max \begin{cases} S(i-1,j) - \beta \\ S(i,j-1) - \beta \\ S(i-1,j-1) + \sigma(a_i, b_j) \end{cases}$$

- **(a) (15%):** Write down a complete pseudo-code for computing $S(M, N)$ in $O(MN)$ time and $O(MN)$ space. All initializations should be included in the pseudo-code.
- **(b) (15%):** Write down an $O(MN)$-time and $O(M + N)$-space version of the pseudo-code in (a).
- **(c) (10%):** Describe an $O(MN)$-time and $O(M + N)$-space approach for delivering an optimal alignment that achieves the score $S(M, N)$. Justify your answer. No pseudo-code is needed for this subproblem.
- **(d) (10%):** Describe an $O(MN)$-time and $O(M+N)$-space approach for delivering an optimal *local* alignment. Justify your answer. No pseudo-code is needed for this subproblem.
- **(e) (10%):** Assume that the maximum length of a gap allowed in an alignment is x. Give a method (as efficient as possible) for computing the score of an optimal alignment. (You may use more than one matrices.)

Problem 2 (15%): In affine gap penalties, a gap of length k is penalized by $\alpha + k \times \beta$, where α and β are both nonnegative constants.

- **(a) (10%):** Give the recurrence relations for computing the score of an optimal (global) alignment between A and B. Justify your recurrence relations.
- **(b) (5%):** Give the recurrence relations for computing the score of an optimal *local* alignment between A and B.

Problem 3 (15%): In restricted affine gap penalties, a gap of length k is penalized by $\alpha + f(k) \times \beta$, where α and β are both nonnegative constants, and $f(k) = \min\{k, c\}$ for a given positive integer c. Give the recurrence relations for computing the score of an optimal (global) alignment between A and B. Justify your answer.

Problem 4 (10%): Explain why the approach of dividing the dynamic-programming matrix by both the middle row and middle column is more space efficient for computing Δ-points (suboptimal points) than the approach of dividing the matrix by merely the middle row.