A Class Note on the Scoring Schemes

Kun-Mao Chao[1,2,3]

[1]Graduate Institute of Biomedical Electronics and Bioinformatics
[2]Department of Computer Science and Information Engineering
[3]Graduate Institute of Networking and Multimedia
National Taiwan University, Taipei, Taiwan 106
Email: kmchao@csie.ntu.edu.tw

November 5, 2007

1 Introduction

This note briefly discusses the scoring schemes for pairwise alignment. 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.

2 A Simple Scoring Scheme

Let us start with a very 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:

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

3 Affine Gap Penalties

In affine gap penalties, a gap of length k is penalized by $\alpha + k \times \beta$, where α and β are both nonnegative constants.

Let $D(i,j)$ denote the score of an optimal alignment between $\langle a_1, a_2, \ldots, a_i \rangle$ and $\langle b_1, b_2, \ldots, b_j \rangle$ ending with a deletion. Let $I(i,j)$ denote the score of an optimal alignment between $\langle a_1, a_2, \ldots, a_i \rangle$ and $\langle b_1, b_2, \ldots, b_j \rangle$ ending with an insertion. Let $S(i,j)$ denote the score of an optimal alignment between $\langle a_1, a_2, \ldots, a_i \rangle$ and $\langle b_1, b_2, \ldots, b_j \rangle$.

By definition, we have

$$\begin{aligned} D(i,j) &= \max_{0 \le i' \le i-1} \{S(i',j) - \alpha - (i-i') \times \beta\} \\ &= \max\{\max_{0 \le i' \le i-2} \{S(i',j) - \alpha - (i-i') \times \beta\}, S(i-1,j) - \alpha - \beta\} \\ &= \max\{\max_{0 \le i' \le i-2} \{S(i',j) - \alpha - ((i-1)-i') \times \beta - \beta\}, S(i-1,j) - \alpha - \beta\} \\ &= \max\{D(i-1,j) - \beta, S(i-1,j) - \alpha - \beta\}. \end{aligned}$$

Similarly,

$$\begin{aligned} I(i,j) &= \max_{0 \le j' \le j-1} \{S(i,j') - \alpha - (j-j') \times \beta\} \\ &= \max\{\max_{0 \le j' \le j-2} \{S(i,j') - \alpha - (j-j') \times \beta\}, S(i,j-1) - \alpha - \beta\} \\ &= \max\{I(i,j-1) - \beta, S(i,j-1) - \alpha - \beta\}. \end{aligned}$$

Therefore, with proper initializations, $D(i,j)$, $I(i,j)$ and $S(i,j)$ can be computed by the following recurrences:

$$D(i,j) = \max \begin{cases} D(i-1,j) - \beta \\ S(i-1,j) - \alpha - \beta \end{cases}$$

$$I(i,j) = \max \begin{cases} I(i,j-1) - \beta \\ S(i,j-1) - \alpha - \beta \end{cases}$$

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

4 Constant Gap Penalties

Now we consider the constant gap penalties where each gap, regardless of its length, is charged with a nonnegative constant penalty α.

Let $D(i,j)$ denote the score of an optimal alignment between $\langle a_1, a_2, \ldots, a_i \rangle$ and $\langle b_1, b_2, \ldots, b_j \rangle$ ending with a deletion. Let $I(i,j)$ denote the score of an optimal alignment between $\langle a_1, a_2, \ldots, a_i \rangle$ and $\langle b_1, b_2, \ldots, b_j \rangle$ ending with an insertion. 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, $D(i,j)$, $I(i,j)$ and $S(i,j)$ can be computed by the following recurrences:

$$D(i,j) = \max \begin{cases} D(i-1,j) \\ S(i-1,j) - \alpha \end{cases}$$

$$I(i,j) = \max \begin{cases} I(i,j-1) \\ S(i,j-1) - \alpha \end{cases}$$

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

5 Restricted Affine Gap Penalties

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 $f(k) = \min\{k, \ell\}$ for a given positive integer ℓ.

Three approaches were explained in the class.

Approach 1: maintain the length of a gap. Give an example to show how it fails.

Approach 2: an efficient way for computing $\max_{0 \le i' \le i-c}\{S(i',j) - \alpha - \ell \times \beta\}$.

Approach 3: use $D'(i,j)$ and $I'(i,j)$ to record the long gap penalties in advance.

$$D(i,j) = \max \begin{cases} D(i-1,j) - \beta \\ S(i-1,j) - \alpha - \beta \end{cases}$$

$$D'(i,j) = \max \begin{cases} D'(i-1,j) \\ S(i-1,j) - \alpha - \ell \times \beta \end{cases}$$

$$I(i,j) = \max \begin{cases} I(i,j-1) - \beta \\ S(i,j-1) - \alpha - \beta \end{cases}$$

$$I'(i,j) = \max \begin{cases} I'(i,j-1) \\ S(i,j-1) - \alpha - \ell \times \beta \end{cases}$$

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

6 Gap Length Limitation

Assume that the maximum length of a gap allowed in an alignment is ℓ.

Acknowledgements

Kun-Mao Chao was supported in part by NSC grants 94-2213-E-002-018 and 95-2221-E-002-126-MY3 from the National Science Council, Taiwan.