A Class Note on Various 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

September 30, 2008

In this note, we shall briefly discuss how to modify the dynamic programming methods to copy with three scoring schemes that are frequently used in biological sequence analysis.

1 Affine Gap Penalties

For aligning DNA and protein sequences, affine gap penalties are considered more appropriate than the simple scoring scheme discussed in the previous sections. "Affine" means that a gap of length k is penalized $\alpha + k \times \beta$, where α and β are both nonnegative constants. In other words, it costs α to open up a gap plus β for each symbol in the gap. Figure 1 computes the score of a global alignment of the two sequences ATACATGTCT and GTACGTCGG under affine gap penalties, where a match is given a bonus score 8, a mismatch is penalized by a score -5, and the penalty for a gap of length k is $-4-k\times 3$.

In order to determine if a gap is newly opened, two more matrices are used to distinguish gap extensions from gap openings. Let $D(i,j)$ denote the score of an optimal alignment between $a_1a_2\ldots a_i$ and $b_1b_2\ldots b_j$ ending with a deletion. Let $I(i,j)$ denote the score of an optimal alignment between $a_1a_2\ldots a_i$ and $b_1b_2\ldots b_j$ ending with an insertion. Let $S(i,j)$ denote the score of an optimal alignment between $a_1a_2\ldots a_i$ and $b_1b_2\ldots b_j$.

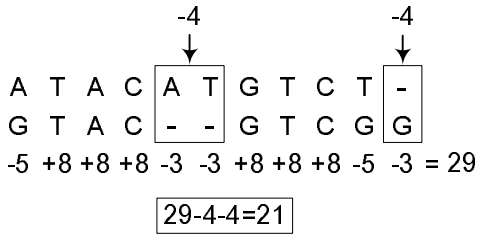

Figure 1: The score of a global alignment of the two sequences ATACATGTCT and GTACGTCGG under affine gap penalties.

By definition, $D(i,j)$ can be derived as follows.

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

This recurrence can be explained in an alternative way. Recall that $D(i,j)$ denotes the score of an optimal alignment between $a_1a_2\ldots a_i$ and $b_1b_2\ldots b_j$ ending with a deletion. If such an alignment is an extension of the alignment ending at $(i-1,j)$ with a deletion, then it costs only β for such a gap extension. Thus, in this case, $D(i,j)=D(i-1,j)-\beta$. Otherwise, it is a new deletion gap and an additional gap-opening penalty α is imposed. We have $D(i,j)=S(i-1,j)-\alpha-\beta$.

In a similar way, we derive $I(i,j)$ as follows.

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

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

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

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

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

2 Constant Gap Penalties

Now let us 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 $a_1 a_2 \ldots a_i$ and $b_1 b_2 \ldots b_j$ ending with a deletion. Let $I(i,j)$ denote the score of an optimal alignment between $a_1 a_2 \ldots a_i$ and $b_1 b_2 \ldots b_j$ ending with an insertion. Let $S(i,j)$ denote the score of an optimal alignment between $a_1 a_2 \ldots a_i$ and $b_1 b_2 \ldots b_j$. With proper initializations, $D(i,j)$, $I(i,j)$ and $S(i,j)$ can be computed by the following recurrences. In fact, these recurrences can be easily derived from those for the affine gap penalties by setting β to zero. A gap penalty is imposed when the gap is just opened, and the extension is free of charge.

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

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

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

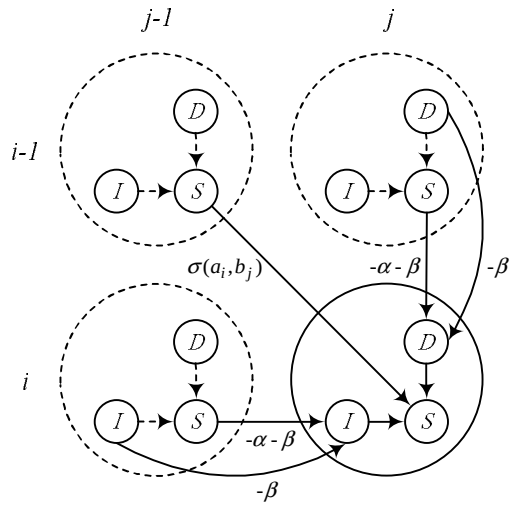

Figure 2: There are seven ways entering the three grid points of an entry (i, j).

3 Restricted Affine Gap Penalties

Another interesting scoring scheme is called the restricted affine gap penalties, in which 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.

In order to deal with the free long gaps, two more matrices $D'(i, j)$ and $I'(i, j)$ are used to record the long gap penalties in advance. With proper initializations, $D(i, j)$, $D'(i, j)$, $I(i, j)$, $I'(i, j)$, and $S(i, j)$ can be computed by the following recurrences:

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

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

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

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

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