Notes on Dynamic-Programming Sequence Alignment

Introduction. Following its introduction by Needleman and Wunsch (1970), dynamic programming has become the method of choice for "rigorous" alignment of DNA and protein sequences. For a number of useful alignment-scoring schemes, this method is guaranteed to produce an alignment of two given sequences having the highest possible score.

For alignment scores that are popular with molecular biologists, dynamic-programming alignment of two sequences requires quadratic time, i.e., time proportional to the product of the two sequence lengths. In particular, this holds for *affine gap costs*, that is, scoring schemes under which a gap of length k is penalized $g + ek$, where g is a fixed "gap-opening penalty" and e is a "gap-extension penalty" (Gotoh, 1982). (More general alignment scores, which are more expensive to optimize, were considered by Waterman *et al*., 1976, but have not found wide-spread use.) Quadratic time is necessitated by the inspection of every pair (i, j), where i is a position in the first sequence and j is a position in the second sequence. For many applications, e.g., database searches, such an exhaustive examination of position pairs may not be worth the effort, and a number of faster methods have been proposed, such as Blast.

The Dynamic-Programming Alignment Algorithm. It is quite helpful to recast the problem of aligning two sequences as an equivalent problem of finding a maximum-score path in a certain graph, as has been observed by a number of authors, including Myers and Miller (1989). This alternative formulation allows the problem to be visualized in a way that permits the use of geometric intuition. We find this visual imagery critical for keeping track of the low-level details that arise in development and implementation of dynamic-programming alignment algorithms.

An *alignment* of two sequences, say S and T, is a rectangular array of symbols having two rows, such that removing all dash characters from the first row (if any are there) gives S, and removing dashes from the second row gives T. Also, we do not allow columns containing two dash symbols. For instance,

```
AAGCAA-A
A-GCTACA
```

is an alignment of `AAGCAAA` and `AGCTACA`.

For the current discussion, we assume the following simple alignment-scoring scheme. For each possible aligned pair $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$, where each of x and y is either a normal sequence entry or the symbol "−", there is an assigned score $\sigma(\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right])$. The score of a pairwise alignment is defined to be the sum of the σ-values of its aligned pairs (i.e., columns). For instance if we score each match (i.e., column of identical symbols) 1, and each other column −1, then the above alignment scores $1 - 1 + 1 + 1 - 1 + 1 - 1 + 1 = 2$.

Recall that a directed graph $G = (V, E)$ consists of a set V of *nodes* (also called *vertices*) and a set E of *edges*. The edge from node u to node v, if it exists, is denoted $u \to v$. A sequence of consecutive edges $u_1 \to u_2, u_2 \to u_3, \cdots, u_{k-1} \to u_k$ is a *path* from u_1 to u_k. If each edge $u \to v$ is assigned a score $\sigma(u \to v)$, then the *score* of such a path is $\sum_{i=1}^{k-1} \sigma(u_i \to u_{i+1})$.

We now describe the relationship between maximum-score paths and optimal alignments. Consider two sequences, $A = a_1 a_2 \cdots a_M$ and $B = b_1 b_2 \cdots b_N$. That is, A contains M symbols and B contains N symbols, where the symbols are from an arbitrary "alphabet" that does not contain the dash symbol, "−". The *alignment graph* for A and B, denoted $G_{A,B}$, is an edge-labeled directed graph. The nodes of $G_{A,B}$ are the pairs (i, j) where $i \in [0, M]$ and $j \in [0, N]$. (We

use the notation $[p, q]$ for the set $\{p, p+1, \cdots, q-1, q\}$.) When graphed, these nodes are arrayed in $M+1$ rows (row i corresponds to a_i for $i \in [1, M]$, with an additional row 0) and $N+1$ columns (column j corresponds to b_j for $j \in [1, N]$). The edge set for $G_{A,B}$ consists of the following edges, labeled as indicated.

1. $(i-1, j) \rightarrow (i, j)$ for $i \in [1, M]$ and $j \in [0, N]$, labeled $\begin{bmatrix} a_i \\ - \end{bmatrix}$
2. $(i, j-1) \rightarrow (i, j)$ for $i \in [0, M]$ and $j \in [1, N]$, labeled $\begin{bmatrix} - \\ b_j \end{bmatrix}$
3. $(i-1, j-1) \rightarrow (i, j)$ for $i \in [1, M]$ and $j \in [1, N]$, labeled $\begin{bmatrix} a_i \\ b_j \end{bmatrix}$

Fig. 1 provides an example of the construction.

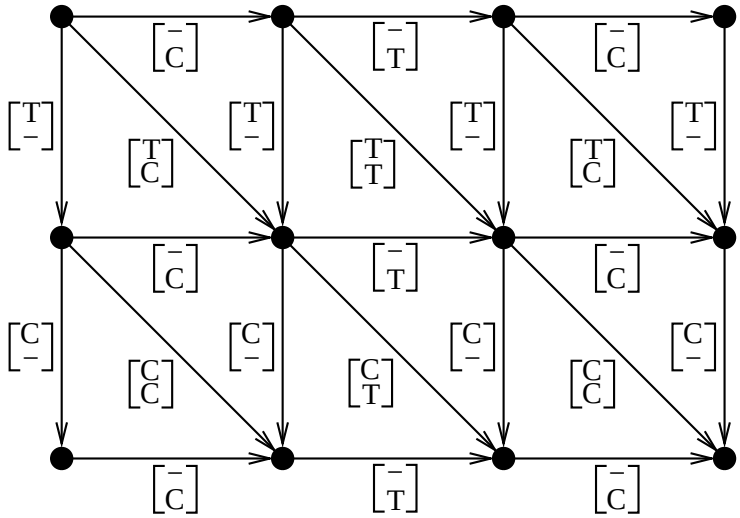

FIG. 1. Alignment graph $G_{A,B}$ for the sequences $A = TC$ and $B = CTC$.

It is instructive to look for a path from $(0, 0)$ (the upper left corner of the graph of Fig. 1) to $(2, 3)$ (the lower right) such that the labels along the path "spell" the alignment:

```
-TC
CTC
```

The first aligned pair is $\begin{bmatrix} - \\ C \end{bmatrix}$, so the first edge must be horizontal. The second pair is $\begin{bmatrix} T \\ T \end{bmatrix}$, so the second edge must be diagonal. The third pair is $\begin{bmatrix} C \\ C \end{bmatrix}$, so the third edge must be diagonal. Generally, when a path descends from row $i-1$ to row i, it picks up an aligned pair with top entry a_i. A path from $(0, 0)$ to (M, N) has zero or more horizontal edges, then a vertical or diagonal edges to row 1, then zero or more horizontal edges, then an edge to row 2, then $\cdots$, so the top entries of the labels along the path are a_1, a_2, $\cdots$, possibly with some interspersed dashes. Similarly, the bottom entries spell B if dashes are ignored, so the aligned pairs spell an alignment of A and B. Indeed, alignments are in general equivalent to paths, as we now state more precisely.

Fact: Let $G_{A,B}$ be the alignment graph for sequences A and B. With each path from $(0, 0)$ to (M, N) associate the alignment formed by concatenating the edge labels along the path, i.e., the alignment "spelled" by the path. Then every such path determines an alignment of A and B, and every alignment of A and B is determined by a unique path. In other words, there is a one-to-one correspondence between paths in $G_{A,B}$ from $(0, 0)$ to (M, N) and alignments of A and B. Furthermore, if the score $\sigma(\pi)$ is assigned to each edge of $G_{A,B}$, where π is the aligned pair labeling that edge, then a path's score is exactly the score of the corresponding alignment.

At each node, the score is computed from the scores of immediate predecessors and of entering edges, which are pictured in Fig. 2. The procedure of Fig. 3 computes the maximum alignment score by considering rows of $G_{A,B}$ in order, sweeping left to right within each row. $S[i, j]$ denotes the maximum score of a path from $(0, 0)$ to (i, j). Lines 7-10 mirror Fig. 2. In row 0 there is but a single edge entering a node (lines 2-3), and similarly for column 0 (line 5). This is a quadratic-space procedure since it uses the $(M+1)$-by-$(N+1)$ array S to hold all node-scores.

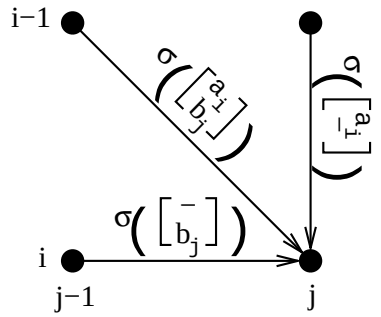

FIG. 2. Edges entering node (i, j) and their scores.

1. $S[0, 0] \leftarrow 0$
2. **for** $j \leftarrow 1$ **to** N **do**
3. $\quad S[0, j] \leftarrow S[0, j-1] + \sigma(\begin{bmatrix} - \\ b_j \end{bmatrix})$
4. **for** $i \leftarrow 1$ **to** M **do**
5. $\quad S[i, 0] \leftarrow S[i-1, 0] + \sigma(\begin{bmatrix} a_i \\ - \end{bmatrix}\}$
6. $\quad$ **for** $j \leftarrow 1$ **to** N **do**
7. $\quad\quad Vertical \leftarrow S[i-1, j] + \sigma(\begin{bmatrix} a_i \\ - \end{bmatrix})$
8. $\quad\quad Diagonal \leftarrow S[i-1, j-1] + \sigma(\begin{bmatrix} a_i \\ b_j \end{bmatrix})$
9. $\quad\quad Horizontal \leftarrow S[i, j-1] + \sigma(\begin{bmatrix} - \\ b_j \end{bmatrix})$
10. $\quad\quad S[i, j] \leftarrow \max\{Vertical, Diagonal, Horizontal\}$
11. **write** "Maximum alignment score is" $S[M, N]$

FIG. 3. Score-only alignment algorithm.

An Example. Consider sequences AGG of length $M = 3$ and ACGT of length $N = 4$. The algorithm systematically fills in entries of a 4-by-5 matrix, where the entry at the intersection of row i and column j is the highest score of any alignment between the first i entries of the first sequence and the first j entries of the second sequence. For this example, suppose that a match scores 1 and any other column scores −1.

Consider computation of the 2,2-entry, given:

		A	C	G	T
	0	−1	−2	−3	−4
A	−1	1	0	−1	−2
G	−2	0	?		

What is the best of the three ways to fill in the entry? Moving from the entry 0 to the left of the new position, or down from the 0 just above, we would add −1 (horizontal or vertical moves always pay the cost for a gap, which is −1 in this example). Coming from the 1 just above and left, we would add the score for aligning A to C, getting 0, so this is the best move and fills in a 0. Now, what about the next entry?

		A	C	G	T
	0	−1	−2	−3	−4
A	−1	1	0	−1	−2
G	−2	0	0	?	

Again the unique best move is down the diagonal, which gives the score 1. Filling in the entire matrix this way, gives:

		A	C	G	T
	0	−1	−2	−3	−4
A	−1	1	0	−1	−2
G	−2	0	0	1	0
G	−3	−1	−1	1	0

Thus the best score for aligning these two sequences is 0. Before reading further, find two alignments that attain this score.

Determining the Alignment. Given the filled-in score array, an optimal alignment can be constructed (in reversed order of its columns) by a ‘‘traceback’’ operation, which walks backwards along an optimal path from the lower right corner to the upper left corner. The basic operation is to identify which of the three possibilities was chosen to reach the current position, which gives (1) the nature of the corresponding alignment column (horizontal move means insertion, diagonal move means substitution, and vertical move means deletion) and (2) tells the previous position on an optimal path.

For instance, the lower-right corner in the above score array can be attained with either a horizontal or a diagonal move. Thus, the last column of an optimal alignment can be either the insertion dash-over-T or the substitution G-over-T.

Local Alignment. In many applications, a global (i.e., end-to-end) alignment of the two given sequences is inappropriate; instead, a local alignment (i.e., involving only a part of each sequence) is desired. In other words, one seeks a high-scoring path that need not terminate at the corners of the dynamic-programming grid (Smith and Waterman, 1981). The highest local alignment score can be computed as follows:

$$S[i,j] \leftarrow \max \begin{cases} 0 & \text{if } 0 \le i \le M \text{ and } 0 \le j \le N \\ S[i-1,j] + \sigma(\begin{bmatrix} a_i \\ - \end{bmatrix}) & \text{if } 1 \le i \le M \text{ and } 0 \le j \le N \\ S[i-1,j-1] + \sigma(\begin{bmatrix} a_i \\ b_j \end{bmatrix}) & \text{if } 1 \le i \le M \text{ and } 1 \le j \le N \\ S[i,j-1] + \sigma(\begin{bmatrix} - \\ b_j \end{bmatrix}) & \text{if } 0 \le i \le M \text{ and } 1 \le j \le N \end{cases}$$

Further complications arise when one seeks k best alignments, where $k > 1$. For computing an arbitrary number of non-intersecting and high-scoring local alignments, Waterman and Eggert (1987) developed a very time-efficient method.

REFERENCES

Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705-708.

Needleman, S. B. and C. D. Wunsch (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48**, 443-453.

Smith, T. F. and M. S. Waterman (1981) Identification of common molecular sequences. *J. Mol. Biol.* **197**, 723-728.

Waterman, M. S., T. F. Smith and W. A. Beyer (1976) Some biological sequence metrics. *Adv. Math.* **20**, 367-387.

Waterman, M. S. and M. Eggert (1987) A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.* **197**, 723-728.