Reprinted from *J. Mol. Biol.* (1981) **147**, 195–197

Identification of Common Molecular Subsequences

T. F. SMITH AND M. S. WATERMAN

J. Mol. Biol. (1981), **147**, 195–197

Identification of Common Molecular Subsequences

The identification of maximally homologous subsequences among sets of long sequences is an important problem in molecular sequence analysis. The problem is straightforward only if one restricts consideration to contiguous subsequences (segments) containing no internal deletions or insertions. The more general problem has its solution in an extension of sequence metrics (Sellers 1974; Waterman *et al.*, 1976) developed to measure the minimum number of "events" required to convert one sequence into another.

These developments in the modern sequence analysis began with the heuristic homology algorithm of Needleman & Wunsch (1970) which first introduced an iterative matrix method of calculation. Numerous other heuristic algorithms have been suggested including those of Fitch (1966) and Dayhoff (1969). More mathematically rigorous algorithms were suggested by Sankoff (1972), Reichert *et al.* (1973) and Beyer *et al.* (1979), but these were generally not biologically satisfying or interpretable. Success came with Sellers (1974) development of a true metric measure of the distance between sequences. This metric was later generalized by Waterman *et al.* (1976) to include deletions/insertions of arbitrary length. This metric represents the minimum number of "mutational events" required to convert one sequence into another. It is of interest to note that Smith *et al.* (1980) have recently shown that under some conditions the generalized Sellers metric is equivalent to the original homology algorithm of Needleman & Wunsch (1970).

In this letter we extend the above ideas to find a pair of segments, one from each of two long sequences, such that there is no other pair of segments with greater similarity (homology). The similarity measure used here allows for arbitrary length deletions and insertions.

Algorithm

The two molecular sequences will be $\underset{\sim}{A} = a_1 a_2 \ldots a_n$ and $\underset{\sim}{B} = b_1 b_2 \ldots b_m$. A similarity $s(a,b)$ is given between sequence elements a and b. Deletions of length k are given weight W_k. To find pairs of segments with high degrees of similarity, we set up a matrix H. First set

$$H_{k0} = H_{0l} = 0 \text{ for } 0 \leq k \leq n \text{ and } 0 \leq l \leq m.$$

Preliminary values of H have the interpretation that H_{ij} is the maximum similarity of two segments *ending* in a_i and b_j, respectively. These values are obtained from the relationship

$$H_{ij} = \max\{H_{i-1,j-1} + s(a_i,b_j), \max_{k\geq 1}\{H_{i-k,j} - W_k\}, \max_{l\geq 1}\{H_{i,j-l} - W_l\}, 0\}, \quad (1)$$

$1 \leq i \leq n$ and $1 \leq j \leq m$.

0022-2836/80/090195-03 $02.00/0

The formula for H_{ij} follows by considering the possibilities for ending the segments at any a_i and b_j.

(1) If a_i and b_j are associated, the similarity is

$$H_{i-1,j-1}+s(a_i,b_j).$$

(2) If a_i is at the end of a deletion of length k, the similarity is

$$H_{i-k,j}-W_k.$$

(3) If b_j is at the end of a deletion of length l, the similarity is

$$H_{i-k,j}-W_l.$$

(4) Finally, a zero is included to prevent calculated negative similarity, indicating no similarity up to a_i and b_j.†

The pair of segments with maximum similarity is found by first locating the maximum element of H. The other matrix elements leading to this maximum value are than sequentially determined with a traceback procedure ending with an element of H equal to zero. This procedure identifies the segments as well as produces the corresponding alignment. The pair of segments with the next best similarity is found by applying the traceback procedure to the second largest element of H not associated with the first traceback.

A simple example is given in Figure 1. In this example the parameters $s(a_ib_j)$ and W_k required were chosen on an *a priori* statistical basis. A match, $a_i=b_j$, produced an $s(a_ib_j)$ value of unity while a mismatch produced a minus one-third. These values have an average for long, random sequences over an equally probable four letter set of zero. The deletion weight must be chosen to be at least equal to the difference between a match and a mismatch. The value used here was $W_k = 1·0+1/3*k$.

	Δ	C	A	G	C	C	U	C	G	C	U	U	A	G
Δ	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0
A	0·0	0·0	1·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	1·0	0·0
A	0·0	0·0	1·0	0·7	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	1·0	0·7
U	0·0	0·0	0·0	0·7	0·3	0·0	1·0	0·0	0·0	0·0	1·0	1·0	0·0	0·7
G	0·0	0·0	0·0	$\underline{1·0}$	0·3	0·0	0·0	0·7	1·0	0·0	0·0	0·7	0·7	1·0
C	0·0	1·0	0·0	0·0	$\underline{2·0}$	1·3	0·3	1·0	0·3	2·0	0·7	0·3	0·3	0·3
C	0·0	1·0	0·7	0·0	1·0	$\underline{3·0}$	1·7	1·3	1·0	1·3	1·7	0·3	0·0	0·0
A	0·0	0·0	2·0	0·7	0·3	$\underline{1·7}$	2·7	1·3	1·0	0·7	1·0	1·3	1·3	0·0
U	0·0	0·0	0·7	1·7	0·3	1·3	$\underline{2·7}$	2·3	1·0	0·7	1·7	2·0	1·0	1·0
U	0·0	0·0	0·3	0·3	1·3	1·0	2·3	$\underline{2·3}$	2·0	0·7	1·7	2·7	1·7	1·0
G	0·0	0·0	0·0	1·3	0·0	1·0	1·0	2·0	$\underline{3·3}$	2·0	1·7	1·3	2·3	2·7
A	0·0	0·0	1·0	0·0	1·0	0·3	0·7	0·7	2·0	3·0	1·7	1·3	2·3	2·0
C	0·0	1·0	0·0	0·7	1·0	2·0	0·7	1·7	1·7	3·0	2·7	1·3	1·0	2·0
G	0·0	0·0	0·7	1·0	0·3	0·7	1·7	0·3	2·7	1·7	2·7	2·3	1·0	2·0
G	0·0	0·0	0·0	1·7	0·7	0·3	0·3	1·3	1·3	2·3	1·3	2·3	2·0	2·0

FIG. 1. H_{ij} matrix generated from the application of eqn (1) to the sequences A-A-U-G-C-C-A-U-U-G-A-C-G-G and C-A-G-C-C-U-C-G-C-U-U-A-G. The underlined elements indicate the trackback path from the maximal element 3·30.

† Zero need not be included unless there are negative values of $s(a,b)$.

Note, in this simple example, that the alignment obtained:

-G-C-C-A-U-U-G-
-G-C-C — U-C-G-

contains both a mismatch and an internal deletion. It is the identification of the latter which has not been previously possible in any rigorous manner.

This algorithm not only puts the search for pairs of maximally similar segments on a mathematically rigorous basis but it can be efficiently and simply programmed on a computer.

Northern Michigan University T. F. SMITH

Los Alamos Scientific Laboratory
P.O. Box 1663, Los Alamos
N. Mex. 87545, U.S.A. M. S. WATERMAN

Received 14 July 1980

REFERENCES

Beyer, W. A., Smith, T. F., Stein, M. L. & Ulam, S. M. (1979). *Math. Biosci.* **19**, 9–25.
Dayhoff, M. O. (1969). *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Silver Springs, Maryland.
Fitch, W. M. (1966). *J. Mol. Biol.* **16**, 9–13.
Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453.
Reichert, T. A., Cohen, D. N. & Wong, A. K. C. (1973). *J. Theoret. Biol.* **42**, 245–261.
Sankoff, D. (1972). *Proc. Nat. Acad. Sci., U.S.A.* **61**, 4–6.
Sellers, P. H. (1974). *J. Appl. Math.* (*Siam*), **26**, 787–793.
Smith, T. F., Waterman, M. S. & Fitch, W. M. (1981). *J. Mol. Evol.* In the press.
Waterman, M. S., Smith, T. F. & Beyer, W. A. (1976). *Advan. Math.* **20**, 367–387.

Note added in proof: A weighting similar to that given above was independently developed by Walter Goad of Los Alamos Scientific Laboratory.