ADVANCES IN APPLIED MATHEMATICS 2, 482–489 (1981)

Comparison of Biosequences

TEMPLE F. SMITH

Northern Michigan University, Marquette, Michigan 48955

AND

MICHAEL S. WATERMAN

Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545

Homology and distance measures have been routinely used to compare two biological sequences, such as proteins or nucleic acids. The homology measure of Needleman and Wunsch is shown, under general conditions, to be equivalent to the distance measure of Sellers. A new algorithm is given to find similar pairs of segments, one segment from each sequence. The new algorithm, based on homology measures, is compared to an earlier one due to Sellers.

1. DISTANCE AND SIMILARITY

Both distance and similarity measures have been designed for the comparison of pairs of biological molecules. The basis of such comparisons is the information from the biochemist as to the linear sequence of elements comprising such macromolecules, as the DNA of the gene. These distance and/or similarity measures have been used by the biologist to obtain information about processes of molecular evolution. The simplest and most fundamental of these are the point mutation (the conversion of one sequence element into another) and the insertion or deletion of sequence elements.

In 1970 Needleman and Wunsch [1] introduced their homology (similarity) algorithm. From a mathematical viewpoint, their work lacks rigor and clarity. But their algorithm has become widely used by the biological community for sequence comparisons.

The two molecules under consideration will be denoted by $\mathbf{a} = a_1a_2\ldots a_n$ and $\mathbf{b} = b_1b_2\ldots b_m$. The basic problem is to find the alignment of $\mathbf{a}$ and $\mathbf{b}$ with the highest similarity. To be specific, we define an alignment of $\mathbf{a}$ and $\mathbf{b}$ by $A(\mathbf{a},\mathbf{b}) = [(a_{i_1}, b_{j_1}), (a_{i_2}, b_{j_2}), \ldots : 1 \le i_1 < i_2 < \cdots \le n, 1 \le j_1 < j_2 < \cdots \le$

0196-8858/81/040482-08$05.00/0

m]. Each $a_k(b_k)$ not appearing in the subsequence $a_{i_1}a_{i_2}\ldots$ $(b_{j_1}b_{j_2}\ldots)$ will be considered an insertion or deletion, depending on the point of view. For display of an alignment, the null element Δ will be inserted in the sequences to indicate insertions/deletions. Thus, the alignment $A = \{(a_1, b_3), (a_2, b_4), (a_4, b_5)\}$ for $\mathbf{a} = a_1a_2a_3a_4$ and $\mathbf{b} = b_1b_2b_3b_4b_5$ is displayed

$$\begin{array}{cccccc} \Delta & \Delta & a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 & \Delta & b_5 \end{array}$$

Frequently, insertions/deletions of length greater than one are used. Then the display becomes

$$\begin{pmatrix} \multicolumn{2}{c}{\Delta} \\ b_1 & b_2 \end{pmatrix} \begin{array}{ccccc} a_1 & a_2 & a_3 & a_4 \\ b_3 & b_4 & \Delta & b_5 \end{array}$$

Similarity measures are based on two weight functions. The first, $s(a_i, b_j)$, measures the degree of "similarity" between two elements a_i, b_j. For ease of notation, let (a_i, b_j) be known as a match of type k if it is assigned weight α_k:

$$s(a_i, b_j) = \alpha_k.$$

The other function necessary is $w_k \geq 0$, the weight assigned to an insertion/deletion of length k. Now let λ_k be the number of matches of type k and Δ_k be the number of insertions/deletions of length k. The similarity measure between $\mathbf{a}$ and $\mathbf{b}$ is then

$$S(\mathbf{a}, \mathbf{b}) = \max_A \left\{ \sum \alpha_i\lambda_i - \sum w_k\Delta_k \right\}.$$

The first theorem provides an algorithm to calculate S. The statement in Needleman and Wunsch [1] is less general, unclearly stated, and does not have a proof.

THEOREM 1. *Let* $S_{0j} = -w_j$, $S_{i0} = -w_i, 0 \leq j \leq m, 0 \leq i \leq n$. *If* $S_{ij} = S(a_1a_2\ldots a_i, b_1b_2\ldots b_j)$, *then*

$$S_{ij} = \max\Big\{ S_{i-1,j-1} + s(a_i, b_j), \max_{k\geq 1} \{S_{i,j-k} - w_k\},$$
$$\max_{l\geq 1} \{S_{i-l,j} - w_l\} \Big\}.$$

Proof. Let A be an optimal alignment for $a_1\ldots a_i$, $b_1\ldots b_j$. There are three cases: (i) a_i is matched with b_j. Then the remaining sequences

$a_1 \dots a_{i-1}, b_1 \dots b_{j-1}$ must be optimally aligned and S_{ij} equals

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

(ii) a_i is a member of an insertion/deletion of length k and S_{ij} equals

$$S_{i-k,j} - w_k.$$

(iii) b_j is a member of an insertion/deletion of length l and S_{ij} equals

$$S_{i,j-l} - w_l.$$

T. F. Smith and S. M. Ulam realized that for the purposes of taxonomic tree construction, it would be appropriate to have a metric D defined for biological sequences. The mathematical community was made aware of this problem by Ulam. P. H. Sellers learned of this problem and solved it in 1972 [2].

As in our discussion of similarity we let (a_i, b_j) be known as a match of type k if it is assigned weight

$$d(a_i, b_j) = \beta_k.$$

Here, d is a "distance" between a_i and b_j, and d is required to be a metric on the set of sequence elements. Sellers only allowed insertions/deletions of length one, but the generalization to insertions/deletions of length k was later made by Waterman et al. [6]. Deletions of length k are assigned weight $x_k \geq 0$. The distance measure between **a** and **b** is then

$$D(\mathbf{a}, \mathbf{b}) = \min_A \left\{ \sum_i \beta_i \lambda_i + \sum_k x_k \Delta_k \right\}.$$

The next theorem was given in [6] and generalizes the work of Sellers [2]. The proof follows the general lines of Theorem 1, but the inclusion of longer insertions/deletions is more difficult.

THEOREM 2. *If $x_1 \leq x_2 \leq \cdots$ and d is a metric on the set of sequence elements, then D is a metric on the set of sequences. Let $D_{i0} = x_i$ and $D_{0j} = x_j$ for $0 \leq i \leq n, 0 \leq j \leq m$. If $D_{ij} = D(a_1 a_2 \dots a_i, b_1 b_2 \dots b_j)$, then*

$$D_{ij} = \min \Big\{ D_{i-1,j-1} + d(a_i, b_j), \min_k \{D_{i,j-k} + x_k\}, \min_l \{D_{i-l,j} + x_l\} \Big\}.$$

Until recently [5], it was not known whether the Needleman-Wunsch algorithm and the Sellers algorithm were equivalent or not. This was largely due to the differences in the way the algorithms were formulated and to the question not being clearly stated. Here the two algorithms are defined to be equivalent if given the weights for one algorithm there is a choice of weights for the second algorithm such that the set of alignments achieving the maximum value for Needleman–Wunsch is equal to the set of alignments achieving the minimum value for Sellers. The following theorem is contained in [5].

THEOREM 3. *The Needleman–Wunsch similarity algorithm is equivalent to the Sellers algorithm. The equivalence is established by setting*

$$\beta_i = \max_j \{\alpha_j\} - \alpha_i$$

and

$$x_k = k/2 \max_j \{\alpha_j\} + w_k.$$

Proof. As above, let $\lambda_i =$ the number of matches of type i and $\Delta_k =$ the number of deletions of length k. The proof is based on the observation that

$$n + m = 2\sum_i \lambda_i + \sum_k k\Delta_k.$$

To be specific, suppose a Needleman–Wunsch algorithm is given. Let

$$\alpha_M = \max_i \alpha_i$$

and

$$\beta_i = \alpha_M - \alpha_i.$$

Then,

$$S = \max_A \left\{ \sum_i \alpha_i \lambda_i - \sum_k w_k \Delta_k \right\}$$

$$= \max_A \left\{ \alpha_M \sum_i \lambda_i - \sum_i \beta_i \lambda_i - \sum_k w_k \Delta_k \right\}.$$

But

$$\sum_i \lambda_i = \frac{n+m}{2} - \sum_k \frac{k}{2} \Delta_k,$$

so

$$S = \max_{A}\left\{\alpha_M \frac{n+m}{2} - \sum_i \beta_i \lambda_i - \sum_k \left(\frac{\alpha_M k}{2} + w_k\right)\Delta_k\right\}$$

$$= \alpha_M \frac{n+m}{2} - \min_{A}\left\{\sum_i \beta_i \lambda_i + \sum_k \left(\frac{\alpha_M k}{2} + w_k\right)\Delta_k\right\}.$$

Therefore, the algorithms are equivalent if

$$\beta_i = \alpha_M - \alpha_i$$

and

$$x_k = \frac{\alpha_M k}{2} + w_k.$$

Of course, not all choices of similarity α_i will induce a metric d on the sequence alphabet. But the cases of interest are all included. For example, the simplest cases have

$$\begin{aligned} s(a,b) &= 0 \quad \text{if } a \neq b, \\ &= 1 \quad \text{if } a = b \end{aligned}$$

and

$$\begin{aligned} d(a,b) &= 1 \quad \text{if } a \neq b, \\ &= 0 \quad \text{if } a = b. \end{aligned}$$

While $d = 1 - s$, we still must use $x_k = k/2 + w_k$.

2. Maximum Similarity Segments

Frequently two sequences which are, overall, no closer together than expected by random will have segments which are quite similar. For a sequence $a = a_1 a_2 \ldots a_n$, a segment is defined to be a subsequence $a_i a_{i+1} \ldots a_j$ where $1 \leq i \leq j \leq n$. After these similar segments are located, it is the task of a biologist to access their significance.

This problem of locating similar patterns within two sequences has been worked on by Sellers [3]. His approach is via distance measures and is quite involved. Our own approach makes use of similarity measures and is simpler. See [4] for an announcement of this result. The connection between these two approaches is discussed in the section following this one.

As before, $\mathbf{a} = a_1a_2\dots a_n$ and $\mathbf{b} = b_1b_2\dots b_m$ are two biological sequences. We are given a similarity measure $s(a, b)$ between two sequence elements and a deletion weight w_k for deletions of length k. Define H_{ij} to be the maximum similarity of two segments that end in a_i and b_j, or zero, whichever is larger.

$$H_{ij} = \max\{0, S(a_xa_{x+1}\dots a_i, b_yb_{y+1}\dots b_j): 1 \le x \le i \text{ and } 1 \le y \le j\}.$$

Zero arises from the view that negative values of H_{ij} represent less similar alignments than no association between the segments.

THEOREM 4. *Set $H_{i0} = H_{0j} = 0$ for $1 \le i \le n$ and $1 \le j \le m$. Then*

$$H_{ij} = \max\left\{H_{i-1,j-1} + s(a_i, b_j), \max_{1\le k\le i}\{H_{i-k,j} - w_k\}, \max_{1\le l\le j}\{H_{i,j-l} - w_k\}, 0\right\}.$$

Proof. The proof is similar to that of Theorem 1. If the best segments have an alignment with a_i and b_j matched, the value of H_{ij} must be

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

If a_i is a member of an insertion/deletion of length, k, H_{ij} must be equal to

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

The case of b_j a member of an insertion/deletion is similar. Finally, H_{ij} equals zero if none of the above situations result in positive similarity.

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 then sequentially determined with a traceback procedure ending with an element of H equal to zero. This procedure both identifies the segments and 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 which is not associated with the first traceback and which has an alignment ending in a match.

Table 1 gives the matrix H for sequences AAUGCCAUUGACGG and CAGCCUCGCUUAG. Here $s(a, b) = 1$ if $a = b$, $s(a, b) = -\frac{1}{3}$ if $a \ne b$, and $w_k = 1 + k/3$. The maximum element is $H_{10,8} = 3.33$ and the corresponding members of the alignment are indicated by underlined traceback. The maximum similar segments are

—GCCAUUG—
—GCCΔUCG—

which has five matches, one mismatch, and one deletion.

TABLE I

Maximum Similarity Calculation

	Δ	C	A	G	C	C	U	C	G	C	U	U	A	G
Δ	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
A	0.00	0.00	1.00	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66
U	0.00	0.00	0.00	0.66	0.33	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.66
G	0.00	0.00	0.00	1.00	0.33	0.00	0.00	0.66	1.00	0.00	0.00	0.66	0.66	1.00
C	0.00	1.00	0.00	0.00	2.00	1.33	0.33	1.00	0.33	2.00	0.66	0.33	0.33	0.33
C	0.00	1.00	0.66	0.00	1.00	3.00	1.66	1.33	1.00	1.33	1.66	0.33	0.00	0.00
A	0.00	0.00	2.00	0.66	0.33	1.66	2.66	1.33	1.00	0.66	1.00	1.33	1.33	0.00
U	0.00	0.00	0.66	1.66	0.33	1.33	2.66	2.33	1.00	0.66	1.66	2.00	1.00	1.00
U	0.00	0.00	0.33	0.33	1.33	1.00	2.33	2.33	2.00	0.66	1.66	2.66	1.66	1.00
G	0.00	0.00	0.00	1.33	0.00	1.00	1.00	2.00	3.33	2.00	1.66	1.33	2.33	2.66
A	0.00	0.00	1.00	0.00	1.00	0.33	0.66	0.66	2.00	3.00	1.66	1.33	2.33	2.00
C	0.00	1.00	0.00	0.66	1.00	2.00	0.66	1.66	1.66	3.00	2.66	1.33	1.00	2.00
G	0.00	0.00	0.66	1.00	0.33	0.66	1.66	0.33	2.66	1.66	2.66	2.33	1.00	2.00
G	0.00	0.00	0.00	1.66	0.66	0.33	0.33	1.33	1.33	2.33	1.33	2.33	2.00	2.00

3. Conclusion

The problem of locating segments of maximum similarity is that of finding segments I_0 and J_0 satisfying

$$\max_{I,J} S(I, J) = S(I_0, J_0)$$

when I is a segment of **a** and J is a segment of **b**. The proof of Theorem 3 shows

$$S(\mathbf{a}, \mathbf{b}) + D(\mathbf{a}, \mathbf{b}) = \alpha_M(n+m)/2.$$

From this it might seem that the problem of segments of maximum similarity is equivalent to finding $\min_{I,J} D(I, J)$ but, if **a** and **b** have any elements in common, this minimum must be zero.

From this discussion, it is clear that the problem is more difficult to approach through the distance measures. This is the approach taken by Sellers [3]. He first defines **a** *most resembles* a segment I of **b** *locally* if

$$d(\mathbf{a}, I) \leq d(\mathbf{a}, L)$$

and

$$d(\mathbf{a}, I) \leq d(\mathbf{a}, J)$$

for all segment L and J satisfying $L \subset I \subset J \subset \mathbf{b}$. One of his algorithms gives all segments $I \subset \mathbf{b}$ most resembling $\mathbf{a}$ locally.

The final algorithm in [3] finds all segments $I \subset \mathbf{a}$ and $J \subset \mathbf{b}$ such that I most resembles J locally. Applying the equality $S + D = \alpha_M n + m/2$, it can be seen that maximum similarity segments are found by this algorithm. However, the algorithm for maximum similarity segments runs in half the steps and is a great deal easier to implement. In addition, the Sellers algorithm does not contain an intrinsic optimality criterion for choosing segments I and J. Our algorithm orders the segments by the value of the similarity measure S.

REFERENCES

1. S. B. NEEDLEMAN AND C. D. WUNSCH, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.* **48** (1970), 443–453.
2. P. H. SELLERS, On the theory and computation of evolutionary distances, *SIAM J. Appl. Math.* **26** (1974), 787–793.
3. P. H. SELLERS, The theory and computation of evolutionary distances: Pattern recognition, *Algorithms*, in press.
4. T. F. SMITH AND M. S. WATERMAN, The identification of common molecular subsequences, *J. Mol. Biol.* **147** (1981), 195–197.
5. T. F. SMITH, M. S. WATERMAN, AND W. M. FITCH, Comparative biosequence metrics. *J. Mol. Evol.*, in press.
6. M. S. WATERMAN, T. F. SMITH, AND W. A. BEYER, Some biological sequence metrics, *Adv. in Math.* **20** (1976), 367–387.