Proc. Nat. Acad. Sci. USA
Vol. 69, No. 1, pp. 4–6, January 1972

Matching Sequences under Deletion/Insertion Constraints

(algorithm/genetic homology/nucleotide sequence/amino-acid sequence)

DAVID SANKOFF

Centre de recherches mathématiques, Université de Montréal, c.p. 6128, Montréal 101

Communicated by Mark Kac, October 18, 1971

ABSTRACT Given two finite sequences, we wish to find the longest common subsequences satisfying certain deletion/insertion constraints. Consider two successive terms in the desired subsequence. The distance between their positions must be the same in the two original sequences for all but a limited number of such pairs of successive terms. Needleman and Wunsch gave an algorithm for finding longest common subsequences without constraints. This is improved from the viewpoint of computational economy. An economical algorithm is then elaborated for finding subsequences satisfying deletion/insertion constraints. This result is useful in the study of genetic homology based on nucleotide or amino-acid sequences.

A problem that arises in the study of evolution at the molecular level (1–4) is to find correspondences between two finite sequences. In its most basic form, the problem is simply to find the longest common subsequence of two sequences. This is termed a best match.

Definition. Let $\{a_i\}_1^m = (a_1, \cdots, a_m)$ and $\{b_i\}_1^n = (b_1, \cdots, b_n)$ be two sequences of elements from a finite set S. A *match* between $\{a_i\}_1^m$ and $\{b_i\}_1^n$ is any set M of pairs $(i,j) \in \{1, \cdots, m\} \times \{1, \cdots, n\}$ such that for all distinct $(i,j) \in M$, $(k,l) \in M$, either

$$\left.\begin{array}{l} i < k \\ j < l \end{array}\right\} \quad \text{OR} \quad \left\{\begin{array}{l} i > k \\ j > l \end{array}\right. \tag{1}$$

The *value* of M is the number of pairs $(i,j) \in M$ such that $a_i = b_j$. A *best match* is a match with maximum value. A *path* P to (i,j) is a match in which (i,j) is the pair with highest coordinates.

In genetics, the set S may be the set of amino acids that constitute protein (1), or $\{$A,C,G,U$\}$, the set of nucleotides found in RNA (5). Construction of matches that satisfy various criteria is the first step in one approach to a determination of the genetic relationship of two types of organism. For m and n equal to 10 or 15, the best match is usually obvious. For the values of m and n of interest in genetics (around 100), however, or to find matches that are not necessarily best, but satisfy some other correspondence criterion, any trial and error method becomes impossibly tedious. Needleman and Wunsch (1) were the first to discover an efficient way to find best matches and matches maximizing certain other criteria. This note describes a new algorithm designed to find matches of the highest value that satisfy constraints important in genetics.

Let $\delta(a_i,b_j) = 1$ if $a_i = b_j$ and $\delta(a_i,b_j) = 0$ otherwise. Following Needleman and Wunsch, for sequences $\{a_i\}_1^m$ and $\{b_i\}_1^n$ we define a matrix V where $V(i,j)$ is the highest value possible for a path to (i,j). Making the convention

$$V(i,0) = V(0,j) = 0 \tag{2}$$

for all $i \in \{0,\cdots,m\}$, $j \in \{0,\cdots,n\}$, we have

$$V(i,j) = \underset{\substack{h \leq i-1 \\ k \leq j-1}}{\text{maximum}}\, V(h,k) + \delta(a_i,b_j) \tag{3}$$

since a highest value path to (i,j) must contain (i,j), as well as the pairs of some highest value path to some pair (h,k), where both $h \leq i - 1$ and $k \leq j - 1$. Then the best match between $\{a\}_1^m$ and $\{b\}_1^n$ will have value

$$v = \underset{\substack{h \leq m \\ k \leq n}}{\text{maximum}}\, V(h,k). \tag{4}$$

This tells us the number of pairs in any best match. To actually construct a best match, proceed as follows. First find an (i,j) such that $a_i = b_j$, and $V(i,j) = v$, which must be possible by (2), (3), and (4) as long as $v > 0$. Then by (3), within $\{1,\cdots, i - 1\} \times \{1,\cdots,j - 1\}$ there must be an (h,k) such that $a_h = b_k$ and $V(h,k) = v - 1$. If we continue in this way, the set of pairs (i,j), (h,k),$\cdots$ so constructed satisfies (1) and so is a match. The algorithm will stop only after it produces a (g,l) such that $a_g = b_l$ and $V(g,l) = 1$, which will be the vth pair to be found. Therefore, we have constructed a best match.

In actual computation, it is more economical to calculate and store, instead of V, the matrix

$$\begin{aligned} W(i,j) &= \underset{\substack{h \leq i \\ k \leq j}}{\text{maximum}}\, V(h,k) \\ &= \max\{W(i - 1,j), W(i,j - 1), W(i - 1,j - 1) \\ &\qquad\qquad + \delta(a_i,b_j)\}. \end{aligned}$$

In this way we assure that in construction of the matrix, the number of search steps and arithmetic steps is proportional to mn. After construction of the matrix (either V or W), if the search for the v pairs for M satisfying (1) is started at position (m,n) and proceeds backwards along the mth row, then the $(m - 1)$st, and so on, no position need be examined more than once. As soon as a pair (i,j) is found, the rest of the search is confined to $\{1,\cdots,i - 1\} \times \{1,\cdots, j - 1\}$. Therefore the whole algorithm, including both matrix construction and search, can proceed in time proportional to mn.

Example

We calculate V and W for the two sequences AGCCAU and CCAGUCU, as depicted in (**5**).

V

		b_j	C	C	A	G	U	C	U
a_i	i ╲ j	0	1	2	3	4	5	6	7
	0	0	0	0	0	0	0	0	0
A	1	0	0	0	**1**	0	0	0	0
G	2	0	0	0	0	**2**	1	1	1
C	3	0	*1*	1	0	1	2	3	2
C	4	0	1	*2*	1	1	2	**3**	3
A	5	0	0	1	*3*	2	2	2	3
U	6	0	0	1	2	3	*4*	3	**4**

W (**5**)

		b_j	C	C	A	G	U	C	U
a_i	i ╲ j	0	1	2	3	4	5	6	7
	0	0	0	0	0	0	0	0	0
A	1	0	0	0	**1**	1	1	1	1
G	2	0	0	0	1	**2**	2	2	2
C	3	0	*1*	1	1	2	2	3	3
C	4	0	1	*2*	2	2	2	**3**	3
A	5	0	1	2	*3*	3	3	3	3
U	6	0	0	1	2	3	*4*	3	**4**

One best match for these sequences consists of pairs that are coordinates of boldface entries in (**5**); another is indicated by italic entries.

Definition. Let M be a match between two sequences. The deletion/insertion (DI) index of M is the number of successive pairs of pairs (i,j), $(k,l) \in M$ such that $k - i \neq l - j$.

As noted in (1), (5), and elsewhere, a match with a low DI index may seem to a geneticist to be a better indication of similarity than a match that has higher value, but also suffers from a higher DI index. Therefore we would like for $q \geq 0$, to find the best match M between $\{a_i\}_1^m$ and $\{b_i\}_1^n$ under the constraint that $DI(M) \leq q$.

Such constraints cannot be incorporated in the Needleman and Wunsch method by means of some combination of their "cell weights" and "gap penalties." It is possible to elaborate a suitable algorithm, however, by use of the fact that a path P to (i,j) with $DI(P) \leq q$ is the union of $\{(i,j)\}$ with either (*1*) a path P_1 to $(i - k,j - k)$, where $0 < k < \min\{i,j\}$, with $DI(P_1) \leq q$, or (*2*) a path P_2 to $(i - g,j - h)$, where $0 < g < i$ and $0 < h < j$, with $DI(P_2) \leq q - 1$.

We will construct matrices V_q for $q = 0,1,\cdots$, where $V_q(i,j)$ is the highest value possible for a path P to (i,j) satisfying $DI(P) \leq q$. As in (**2**), we define

$$V_q(i,0) = V_q(0,j) = 0 \qquad (\mathbf{6a})$$

for all $i \in \{0,\cdots,m\}$, $j \in \{0,\cdots,n\}$ and $q \in \{0,1,\cdots\}$. Then

$$V_0(i,j) = V_0(i - 1,j - 1) + \delta(a_i,b_j) \qquad (\mathbf{6b})$$

and

$$\begin{aligned} V_q(i,j) &= \underset{\substack{0<k<i,0<k<j\\0<g<i,0<h<j}}{\text{maximum}} \{V_q(i - k,j - k),V_{q-1}(i - g,j - h)\} \\ &\qquad + \delta(a_i,b_j) \\ &= \underset{\substack{0<g<i\\0<h<j}}{\text{maximum}} \{V_{q-1}(i - 1,j - h),V_q(i - 1,j - 1), \\ &\qquad V_{q-1}(i - g.j - 1)\} + \delta(a_i,b_j) \qquad (\mathbf{6c}) \end{aligned}$$

for all $i \in \{1,\cdots,m\}$, $j \in \{1,\cdots,n\}$ and $q \in \{1,2,\cdots\}$. A best match M satisfying $DI(M) \leq q$ will have value

$$v_q = \underset{\substack{0<h\leq m\\0<k\leq n}}{\text{maximum}} V_q(h,k). \qquad (\mathbf{7})$$

Now that we have v_q, we can find a suitable match. First find a pair (i,j) satisfying $a_i = b_j$ and

$$V_q(i,j) = v_q.$$

Such a pair exists by (**6**) and (**7**), as long as $v_q > 0$. By (**6c**) it is clear that if

$$V_q(i - 1,j - 1) \neq v_q - 1$$

then

$$\underset{\substack{0<g\leq i-1\\0<h\leq j-1}}{\text{maximum}} V_{q-1}(i - g,j - h) = v_q - 1$$

and we are in the same situation as after (**7**), i.e., we have completed the first step of the algorithm.

Otherwise, for some positive $k < \min\{i,j\}$ we have

$$\begin{aligned} V_q(i - 1,j - 1) &= V_q - \delta(a_i,b_j) = v_q - 1 \\ V_q(i - 2,j - 2) &= v_q - \delta(a_i,b_j) - \delta(a_{i-1},b_{j-1}) \\ &\cdot \\ &\cdot \\ &\cdot \\ V_q(i - k,j - k) &= v_q - \sum_{x=0}^{k-1} \delta(a_{i-x},b_{j-x}) \end{aligned}$$

but

$$V_q(i - k - 1,j - k - 1) < v_q - \sum_{x=0}^{k} \delta(a_{i-x},b_{j-x}).$$

Then there are $\sum_{x=0}^{k-1} \delta(a_{i-x},b_{j-x})$ pairs for our match on the matrix diagonal between (i,j) and $(i - k,j - k)$, inclusive. To find the next pair, we must invoke (**6c**) again to assert

$$\underset{\substack{0<g<i-k\\0<h<j-k}}{\text{maximum}} V_{q-1}(i - g,j - h) = v_q - \sum_{x=0}^{k-1} \delta(a_{i-x},b_{j-x}),$$

which completes the first cycle of the algorithm [i.e., return to (**7**)]. By the definition of V_q, the algorithm will stop only after v_q pairs have been found.

In a manner analogous to the unconstrained best-match case, it is possible to make this procedure more efficient by introduction of the matrices

$$W_q(i,j) = \max\{W_q(i - 1,j),\ V_q(i,j),\ W_q(i,j - 1)\}$$

for $q = 0,1,\cdots$, so that

$$V_q(i,j) = \max\{V_q(i-1,j-1), W_{q-1}(i-1,j-1)\} + \delta(a_i,b_j)$$

for $q = 1,2,\cdots$. Then for each successive q, the matrix V_q can be constructed in "computer time" proportional to mn from W_{q-1}, and W_q can be constructed from V_q in similar time. A search strategy similar to the one we used in the best-match case will assure that no pair (i,j) need be searched more than once, so that the total number of computations to find best matches under all possible *DI* constraints is $mnq_{\max}$, where $q_{\max}$ is usually much less than m or n. To find $q_{\max}$, we first calculate v for unconstrained best matches, and the minimum q for which $v_q = v$ is $q_{\max}$.

Example

		b_j	C	C	A	G	U	C	U
	j	0	1	2	3	4	5	6	7
a_i	i								
	0	0	0	0	0	0	0	0	0
A	1	0	0	0	1	0	0	0	0
G	2	0	0	0	0	2	0	0	0
C	3	0	1	1	0	0	2	1	0
C	4	0	1	2	1	0	0	3	1
A	5	0	0	1	3	1	0	0	3
U	6	0	0	0	1	3	2	0	1

V_0

		b_j	C	C	A	G	U	C	U
	j	0	1	2	3	4	5	6	7
a_i	i								
	0	0	0	0	0	0	0	0	0
A	1	0	0	0	1	1	1	1	1
G	2	0	0	0	1	2	2	2	2
C	3	0	1	1	1	2	2	2	2
C	4	0	1	2	2	2	2	3	3
A	5	0	1	2	3	3	3	3	3
U	6	0	1	2	3	3	3	3	3

W_0

For $q \geq 1$, the matrices V_q and W_q are just V and W, respectively, as shown in (**5**). The best matches with *DI* index equal to one are just the unconstrained best matches as in (**5**). The best matches with zero *DI* indices consist of the first three pairs of each unconstrained best match.

The advantage of this method over that of Needleman and Wunsch, as well as others that have been published, is that it does not depend on any arbitrarily imposed numerical criteria such as cell weights, but on a genetically meaningful criterion, the *DI* index. Further, this algorithm can be used as the basis for statistically testing hypotheses not only about the similarity of two sequences, but also about the number of deletions and insertions separating them (6).

I thank R. J. Cedergren for introducing me to this problem.

1. Needleman, S. B. & Wunsch, C. D. (1970) "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *J. Mol. Biol.*, **48,** 443–453.

2. Margoliash, E., Fitch, W. M. & Dickerson, R. E. (1969) 'Molecular Expression of Evolutionary Phenomena in the Primary and Secondary Structures of Cytochrome *c*,' in "Structure, Function, and Evolution in Proteins," *Brookhaven Symposia in Biology, no. 21* (Brookhaven National Laboratory, Upton, N.Y.), pp. 295–305.

3. Gibbs, A. J. & McIntyre, G. A. (1970) "The Diagram, a Method for Comparing Sequences," *Eur. J. Biochem.* **16,** 1–11.

4. Haber, J. E. & Koshland, D. E. (1970) "An Evaluation of the Relatedness of Proteins Based on Comparison of Amino Acid Sequences," *J. Mol. Biol.* **50,** 617–639.

5. Barker, W. C., Wallace, D. C. & Dayhoff, M. O. (1969) "5S Ribosomal RNA," in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M.O. (National Biomedical Research Foundation, Silver Spring, Md.), Vol. 4, pp. 95–98.

6. Sankoff, D. & Cedergren, R. J. (1971) "A Test for Nucleotide Sequence Homology," *Technical Report 122* (Centre de recherches mathématiques, Université de Montréal, Montreal, Canada).