J. theor. Biol. (1984) **108,** 333–337

Efficient Sequence Alignment Algorithms

MICHAEL S. WATERMAN

Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113, *U.S.A.*

(*Received* 23 *December* 1982, *and in revised form* 25 *July* 1983)

Sequence alignments are becoming more important with the increase of nucleic acid data. Fitch and Smith have recently given an example where multiple insertion/deletions (rather than a series of adjacent single insertion/deletions) are necessary to achieve the correct alignment. Multiple insertion/deletions are known to increase computation time from $O(n^2)$ to $O(n^3)$ although Gotoh has presented an $O(n^2)$ algorithm in the case the multiple insertion/deletion weighting function is linear. It is argued in this paper that it could be desirable to use concave weighting functions. For that case, an algorithm is derived that is conjectured to be $O(n^2)$.

Introduction

With the advent of rapid methods for determining nucleic acid sequences, there is increased interest in computer methods for comparing these sequences. Sequencing is estimated to be proceeding at the rate of 10^6 bases per year and various data bases are being structured to organize the sequences and attendant information. Relevant portions of the data base are searched for sequences similar to ones recently determined; and rapid, efficient, and meaningful algorithms are necessary.

The Needleman & Wunsch (1970) algorithm was the first rapid method in the biological literature for determining sequence homology and was followed by the metrics of Sankoff (1972) and Sellers (1974*a*,*b*) for finding the distance between two sequences. Kruskal (1983) recently presented an extensive review of these and related methods and applications.

Sellers' work was generalized to include multiple insertion/deletions by Waterman, Smith & Beyer (1976). A review of the use of these techniques in sequence analysis was given by Smith, Fitch & Waterman (1981). Fitch & Smith (1983) recently studied these algorithms for a region of DNA coding for alpha and beta hemoglobin in chicken, where the alignment is assumed known by comparison of the many corresponding protein sequences, and they determined that a specific range of weights for multiple insertion/deletions were necessary to obtain correct alignments. Therefore, it is important to include these in application algorithms.

In this paper we review recent work of Gotoh (1982) who presented an $O(n^2)$ algorithm for the case of linear insertion/deletion functions. An extension to concave insertion/deletion functions is obtained here which has computation complexity close to $O(n^2)$.

Linear Insertion/Deletions

To make the discussion specific, the algorithm of Waterman *et al.* (1976) will now be presented. $\mathbf{a} = a_1 a_2 \ldots a_n$ and $\mathbf{b} = b_1 b_2 \ldots b_m$ are the sequences of interest and $D(\mathbf{a}, \mathbf{b})$ will denote the (minimum) distance between $\mathbf{a}$ and $\mathbf{b}$ where $d(x, y)$ is a dissimilarity measure between the sequence elements, and deletions of length k are assigned weight $w(k)$. We take $D_{i0} = w(i)$, $D_{0j} = w(j)$ and $D_{ij} = D(a_1 a_2 \ldots a_i, b_1 b_2 \ldots b_j)$ and proceed by

$$D_{ij} = \min \{D_{i-1,j-1} + d(a_i, b_j), \min \{D_{i,j-k} + w(k): 1 \le k \le j\},$$
$$\min \{D_{i-l,j} + w(l): 1 \le l \le i\}\}.$$

Of course, the final result is $D_{n,m} = D(\mathbf{a}, \mathbf{b})$. The computational efficiency of the algorithm is of order

$$\sum_{i,j} (i+j) = O(n^2 m + m^2 n)$$

which, when $n = m$, is $O(n^3)$.

Paul Haeberli of the University of Wisconsin brought it to our attention that this algorithm can be improved to $O(n^2)$ when $w(k)$ is a linear function of k, although he did not carefully state the condition on $w(k)$ or give a complete proof. Much the same observation was made by Goad & Kanehisa (1982) regarding the secondary structure algorithm of Waterman & Smith (1978). More recently Gotoh (1983) gave a complete and clear proof for the new algorithm.

Concave Insertion/Deletions

The intuitive argument for multiple insertion/deletion functions is that a deletion of fourteen bases (say) should not be thought of as fourteen independent single deletions but as one deletion event which has weight less than the sum the weights of fourteen single deletions. This reasoning implies a general inequality

$$w(k+l) \le w(k) + w(l), \qquad k, l \ge 1.$$

It is possible to give a slightly improved set of inequalities which proves useful in this problem. If we agree that bases are increasingly easier to

insert (delete) as the insertion (deletion) length grows, then the resulting inequalities are

$$w(m+k+l)-w(m+k)\leq w(k+l)-w(k), \qquad k, l, m\geq 1.$$

If equality is required, then linearity follows:

$$w(k)=a+b(k-1).$$

The assumption of linearity is now transparent; $w(k)$ linear means that after the first deletion (insertion) each successive addition of a base has equal cost. We argue that strict inequality could be preferable. Since the function w is increasing and has decreasing differences, it is concave downward, here simply referred to as concave. Such functions as

$$w(k)=a+b\log(k), \qquad a, b>0$$

are concave and have intuitive appeal.

Next we make the assumption of linearity, $w(k)=a+b(k-1)$ and review Gotoh's proof. He presents the above recursion for D_{ij} in the form

$$D_{ij}=\min\{D_{i-1,j-1}+d(a_i, b_j), E_{i,j}, F_{i,j}\},$$

where

$$E_{i,j}=\min\{D_{i,j-k}+w(k): 1\leq k\leq j\},$$

$$F_{i,j}=\min\{D_{i-l,j}+w(l): l\leq l\leq i\},$$

where $E_{00}=F_{00}=D_{00}=0$, $E_{i0}=D_{i0}=w(i)$, $F_{0j}=D_{0j}=w(j)$. Gotoh then observes

$$\begin{aligned}E_{i,j}&=\min\{D_{i,j-1}+a, \min\{D_{i,j-k}+w(k): 2\leq k\leq j\}\}\\&=\min\{D_{i,j-1}+a, \min\{D_{i,j-1-l}+w(l+1): 1\leq l\leq j-1\}\}\\&=\min\{D_{i,j-1}+a, \min\{D_{i,j-1-l}+w(l): 1\leq l\leq j-1\}+b\}\\&=\min\{D_{i,j-1}+a, E_{i,j-1}+b\}.\end{aligned}$$

Of course

$$F_{i,j}=\min\{D_{i-1,j}+a, F_{i-1,j}+b\}$$

and the algorithm is of order $O(nm)$ or $O(n^2)$ in case $n=m$.

Now we study the problem of efficient algorithms for w concave. First of all, set

$$E_{i,j}=\min\{D_{i,k}+w(j-k): k\in[0, j-1]\}.$$

This means that, for some $l\in[0, j-1]$,

$$D_{i,l}+w(j-l)\leq D_{i,k}+w(j-k), \qquad 0\leq k\leq j-1.$$

If $k \geq l$, $w(j-l+1)-w(j-l) \leq w(j-k+1)-w(j-k)$, and

$$D_{i,l}+w(j-l)+w(j-l+1)-w(j-l) \leq D_{i,k}+w(j-k)+w(j-k+1)-w(j-k)$$

or

$$D_{i,l}+w(j+1-l) \leq D_{i,k}+w(j+1-k).$$

Therefore

$$E_{i,j+1}=\min\{D_{i,k}+w(j+1-k): k \in [0, l]; D_{ij}+w(1)\}.$$

The utility of the last equation is that minimization can be reduced to those D_{ij} associated with

$$S(i)=\{l: E_{i,l+1}=D_{i,l}+w(1)\}.$$

In fact

$$E_{i,j}=\min\{D_{i,k}+w(j-k): k \in S(i)\}.$$

The computational complexity of this algorithm depends on the rate of growth of $S(i)$ as a function of sequence length. On a row of length m, how many times will length one deletions be optimal? We conjecture that this function does not grow faster than $\log(m)$.

For further improvement of the algorithm, notice that whenever

$$D_{i,j-1}+w(1) < D_{i,k}+w(j-k), j-1 > k \in S(i),$$

it is possible to calculate when $D_{i,k}$ can be superior to $D_{i,j-1}$. That is, solve

$$D_{i,j-1}+w(1+h)=D_{i,k}+w(j-k+h).$$

If $j+h > m$, then $D_{i,k}$ never need be considered again and k can be removed from $S(i)$. Otherwise, D_{ik} need not be considered until calculation of $E_{i,j+h}$. While we cannot establish the computational complexity of this last result on first principles, the practical implication is an $O(n^2)$ algorithm.

Conclusion

More development of sequence comparison algorithms will occur in the near future. The rapidly increasing data base will force those working on algorithms for molecular biology to construct algorithms that are more efficient and that answer increasingly more special and involved questions. We are, for example, applying our results on concave deletion functions to algorithms for RNA secondary structure. One of the interesting aspects will be the effects that molecular biology and computer science will have as they continue to communicate and cooperate.

The Department of Biochemistry and Biophysics of the University of California at San Francisco and System Development Foundation are each gratefully acknowledged for providing partial support of this work.

REFERENCES

FITCH, W. M. & SMITH, T. F. (1983). *Proc. natn. Acad. Sci. U.S.A.* **80,** 1382.
GOTOH, O. (1982), *J. mol. Biol.* **162,** 705.
GOAD, W. B. & KANEHISA, M. I. (1982). *Nucleic Acids Res.* **10,** 247.
KRUSKAL, J. B. (1983). *SIAM Rev.* **25,** 201.
NEEDLEMAN, S. & WUNSCH, C. (1970). *J. mol. Biol.* **42,** 245.
SANKOFF, D. (1972). *Proc. natn. Acad. Sci. U.S.A.* **69,** 4.
SMITH, T. F., FITCH, W. M. & WATERMAN, M. S. (1981). *J. mol. Evol.* **18,** 38.
SELLERS, P. (1974*a*). *J. Comb. Theory Ser.* A **16,** 253.
SELLERS, P. (1974*b*). *J. SIAM* **26,** 787.
WATERMAN, M. S., SMITH, T. F. & BEYER, W. A. (1976). *Adv. Math.* **20,** 367.
WATERMAN, M. S. & SMITH, T. F. (1978). *Math. Bio. Sci.* **42,** 257.