A NONLINEAR MEASURE OF SUBALIGNMENT SIMILARITY AND ITS SIGNIFICANCE LEVELS

Stephen F. Altschul*† and Bruce W. Erickson*
*The Rockefeller University,
New York, NY 10021, U.S.A.
†Department of Applied Mathematics,
Massachusetts Institute of Technology,
Cambridge, MA 02139, U.S.A.

A new measure of subalignment similarity is introduced. Specifically, similarity \( s(l, c) \) is defined as the logarithm to the base \( p \) of the probability of finding \( c \) or fewer mismatches in a subalignment of length \( l \), where \( p \) is the probability of a match. Previous algorithms cannot use this measure to find locally optimal subalignments because, unlike Needleman–Wunsch and Sellers similarities, this measure is nonlinear. A new pattern recognition algorithm is described for finding all locally optimal subalignments of two nucleotide sequences. The DD algorithm can use \( s(l, c) \) or any other reasonable similarity function to assess the relative interest of subalignments. The DD algorithm searches only the diagonal graph, which lacks insertions and deletions. This search strategy greatly decreases the computation time and does not require an arbitrary choice of gap cost. The paths of the resulting DD graph usually draw attention to likely locations for insertions and deletions. A heuristic formula is derived for estimating significance levels for \( s(l, c) \) in the context of the lengths of the two aligned sequences. The DD algorithm has been used to find interesting subalignments between the nucleotide sequences for human and murine interleukin 2.

Introduction. An alignment of sequences \( S_1 \) and \( S_2 \) is a correspondence of the elements of \( S_1 \) with the elements of \( S_2 \). An arbitrary number of nulls (missing elements) may be inserted into each sequence provided that each element of one sequence is aligned with one element of the other sequence or with a null (shown as a dash) and two nulls are not aligned. For example, three alignments of the sequences CGGA and CGA are

\[
A_1 = \begin{align*}
CGGA \\
CG-A
\end{align*} \quad A_2 = \begin{align*}
CGGA \\
C-GA
\end{align*} \quad A_3 = \begin{align*}
CGG-A \\
-CGA-
\end{align*}.
\]

If a cost is assigned for each aligned pair of elements and for each alignment of an element with a null, the alignment cost is the sum of the costs for all aligned pairs. The distance \( d \) between \( S_1 \) and \( S_2 \) is the minimum alignment cost for all possible alignments, and an alignment whose cost is the distance is a metric alignment.

The SS algorithm (Sellers, 1974) finds all metric alignments of two specified
sequences. Normally, an aligned pair of identical elements (a match) is assigned a cost of 0, an aligned pair of non-identical elements (a mismatch) is assessed a cost of 1, and alignment of an element with a null (a point insertion or deletion) is charged an arbitrary null cost $U$. If $U$ is 2, the cost of alignment $A_3$ is 7 while alignments $A_1$ and $A_2$ each cost 2. Since the lowest cost for aligning these sequences is 2, both $A_1$ and $A_2$ are metric alignments but $A_3$ is not.

A subalignment of sequences $S_1$ and $S_2$ is an alignment of a segment of $S_1$ with a segment of $S_2$, where a segment is a contiguous series of elements. The SU algorithm (Sellers, 1980; Erickson and Sellers, 1983) finds all locally optimal (metric) alignments of one sequence with unspecified segments of another sequence. A subalignment is represented in a graph by a path (contiguous series of horizontal, diagonal and/or vertical edges). For example, Fig. 1b shows the path for alignment $A_1$. A diagonal edge aligns an element of $S_1$ with an element of $S_2$ (a match or mismatch), a horizontal edge aligns an element of $S_1$ with a null, and a vertical edge aligns an element of $S_2$ with a null. Each subalignment corresponds to a unique path.

![Figure 1. Four graphs for the sequences CGGA and CGA. (a) The complete graph; (b) metric alignment $A_1$; (c) the diagonal graph; (d) the DD graph for $s^* = 2$.](image)

**Subalignment Similarity**

**Length vs cost.** The definition of a locally optimal subalignment of two sequences presents a quandary. Subalignments with the smallest distance usually consist of just a few matches. But an alignment of two segments of length 20 with two mismatches is often more interesting than an alignment of two segments of length 5 with no mismatches. A balance is needed between finding longer subalignments and finding subalignments with fewer mismatches. Smith and Waterman (1981) suggested that maximizing the similarity measure of Needleman and Wunsch (1970) can be used to find interesting