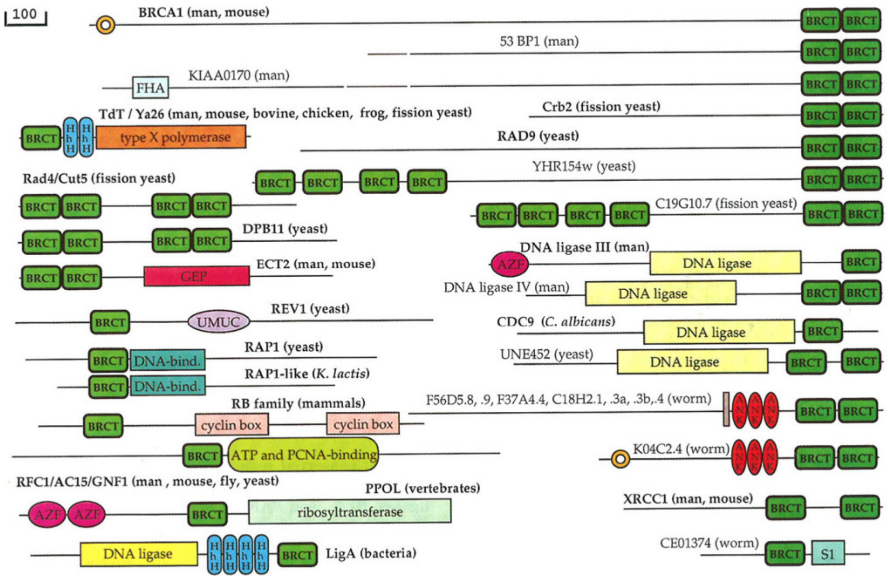
Local Sequence Alignment

Stephen Altschul

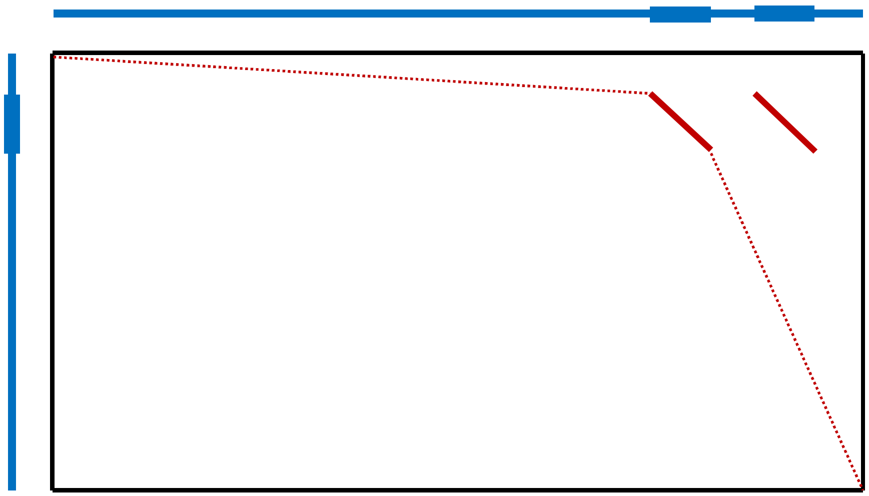
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

Local Alignment: Motivation

In the early days of protein sequence comparison, most known related proteins, were related over their whole lengths. However, soon proteins that shared only isolated regions of similarity were found. A schematic of a protein *superfamily* is shown here, with related *domains* represented by similar boxes.



The measure of global sequence similarity, and the Needleman-Wunsch alignment algorithm, was not well-adapted to finding such domains. A new definition of local similarity was required, along with a new algorithm for finding locally optimal alignments.



Local Alignment: Definition

During the 1970s and early 1980s, a variety of definitions for local alignment were proposed. The one that eventually gained the greatest popularity, along with an associated algorithm, is due to Smith & Waterman.

Smith & Waterman proposed simply that a local alignment of two sequences allow arbitrary-length segments of each sequence to be aligned, with no penalty for the unaligned portions of the sequences. Otherwise, the score for a local alignment is calculated the same way as that for a global alignment.

It would at first appear that the problem of finding an optimal local alignment should be significantly more complex than the problem of finding an optimal global alignment, because the start and stop positions of the alignment must be located as well. However, only a constant factor more calculation is necessary.

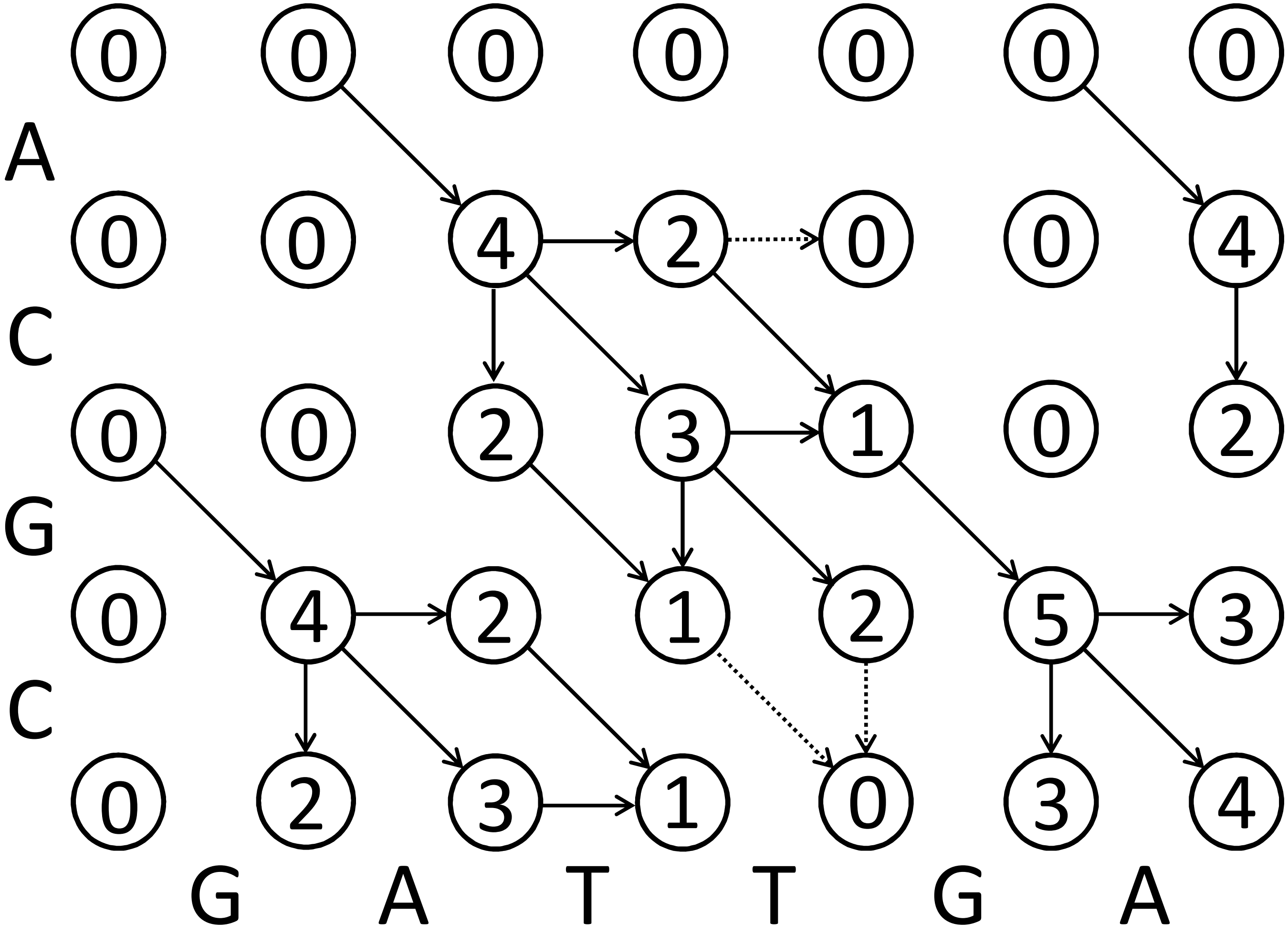
Smith, T.F. & Waterman, M.S. (1981) “Identification of common molecular subsequences.” *J. Mol. Biol.* **147**:195-197.

The Smith-Waterman Algorithm

Two modifications to Needleman-Wunsch:

1) Allow a node to start at 0.

2) Record the highest-scoring node, and trace back from there.



		G	A	T	T	G	A
	0	0	0	0	0	0	0
A	0	0	4	2	0	0	4
C	0	0	2	3	1	0	2
G	0	4	2	1	2	5	3
C	0	2	3	1	0	3	4

Why does this algorithm yield an optimal local alignment?

Scores: Match +4 Mismatch -1 Gap -2

Pseudocode for Finding Local Sequence Similarity

```
Local_Similarity(X,Y):
    S=0
    For i = 0,...,m:  SIM[i,0] = 0
    For j = 1,...,n:   SIM[0,j] = 0
    For i = 1,...,m:
        For j = 1,...,n:
            SIM[i,j] = max(
                0,
                SIM[i-1,j-1] + s(X[i],Y[j]),
                SIM[i-1,j]+g,
                SIM[i,j-1]+g
            )
            S=max(S,SIM[i,j])
        EndFor
    EndFor
Return S
```

Exercise: Generalize the code to include traceback information, and produce one optimal local alignment.

Multiplying all substitution and gap scores by a positive constant does not change the optimal alignment. Why?

Adding a constant k to all substitution scores, and $k/2$ to all gap scores, *can* change the optimal alignment. Why?

The Smith-Waterman Algorithm: Traceback

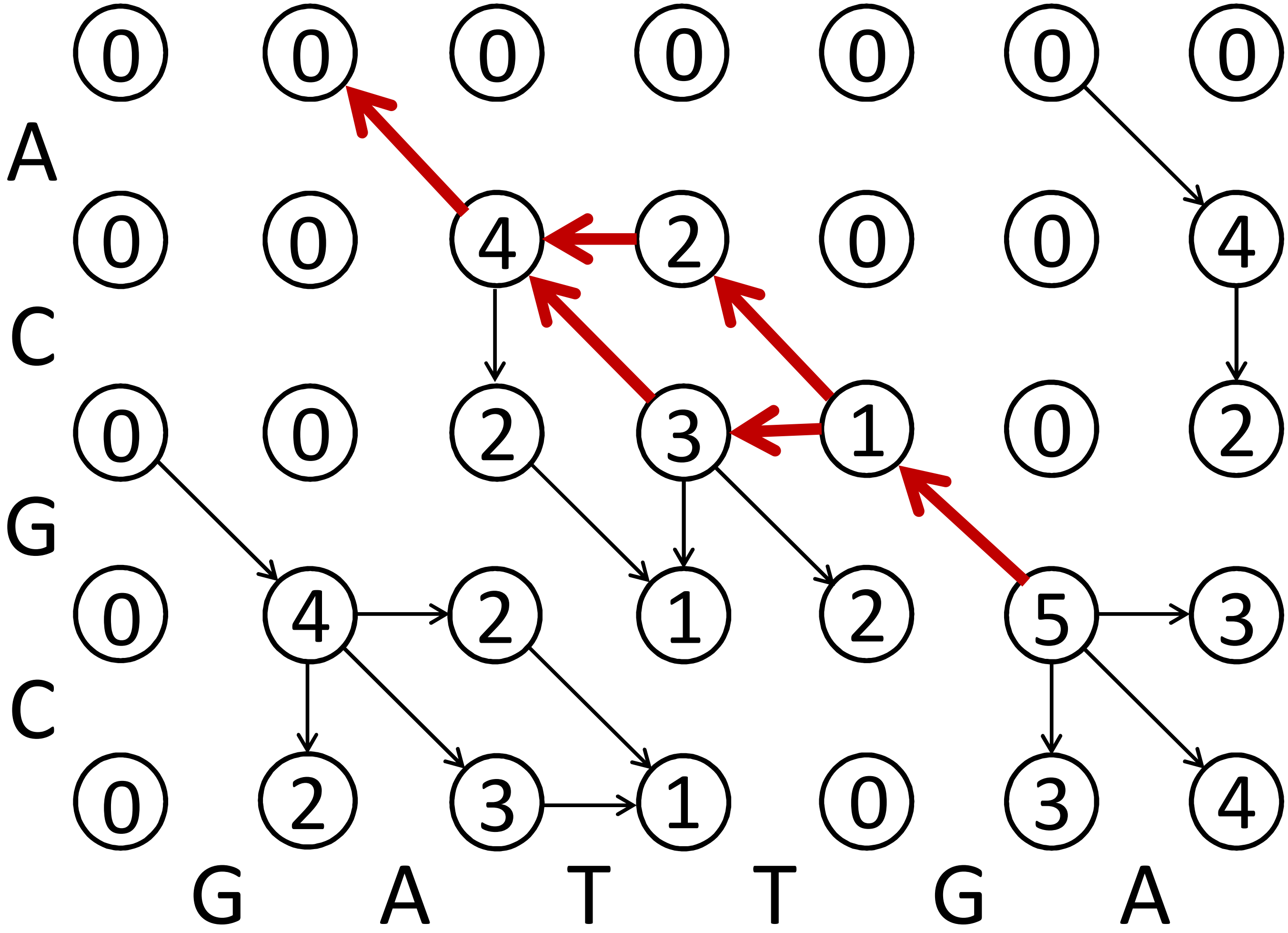
Optimal local alignments, or *subalignments*:

```
AC–G         A–CG
ATTG   and   ATTG
```

<u>Questions</u>:

Can one find other *locally optimal* subalignments?

How can they be defined?



		G	A	T	T	G	A
	0	0	0	0	0	0	0
A	0	0	4	2	0	0	4
C	0	0	2	3	1	0	2
G	0	4	2	1	2	5	3
C	0	2	3	1	0	3	4

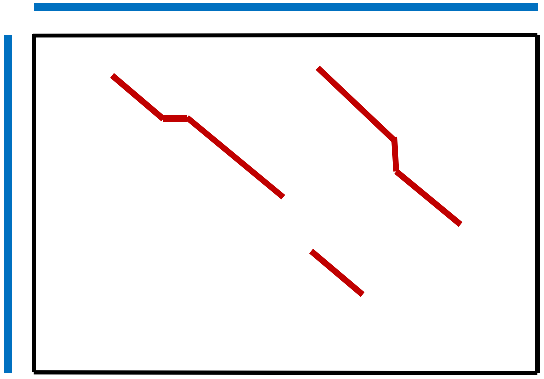
Scores: Match +4   Mismatch -1   Gap -2

Local optimality: Definitions and Algorithms

A definition of local optimality was proposed in 1984, along with an algorithm to find all locally optimal subalignments. [Sellers, P.H. (1984) *Bull. Math. Biol.* **46**:501-514.]

A subalignment is *locally optimal* if its score is greater than or equal to that of any subalignment it "touches".

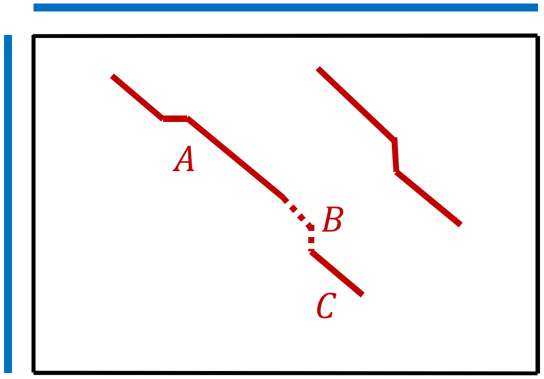
A provably $O(mn)$ algorithm for finding all locally optimal subalignments was subsequently described. [Altschul, S.F. & Erickson, B.W (1986) *Bull. Math. Biol*. **48**:633-660.]



Problem: By Sellers' definition, a strong subalignment can suppress, by means of intermediaries, subalignments it does not actually touch. This can be a particular problem if one is seeking internal approximate repeats.

One may advance an alternative definition to address this problem: A subalignment is *weakly locally optimal* if it touches no weakly locally optimal subalignment that has greater score (Altschul & Erickson, 1986). This definition is not circular, but recursive.

No $O(mn)$ algorithm for finding all weakly locally optimal subalignments of two sequences has been described, although several incorrect ones have been published.



$SIM(A) > SIM(B) > SIM(C)$

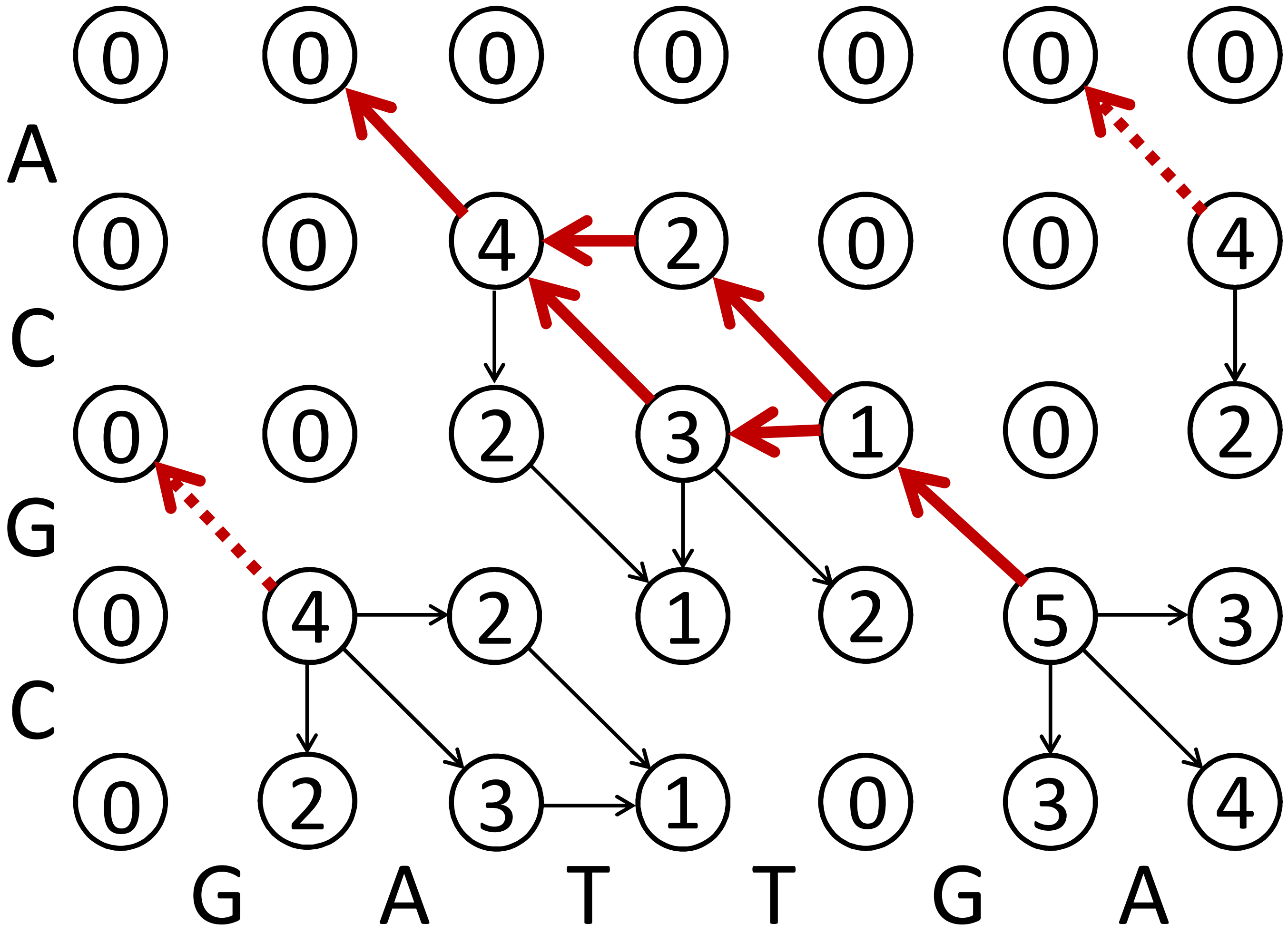
Locally Optimal Subalignments

Optimal subalignments:

```
AC–G        A–CG
ATTG  and   ATTG
```

Additional, locally optimal subalignments:

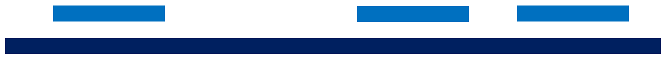
```
G        A
G  and   A
```



	G	A	T	T	G	A
0	0	0	0	0	0	0
A 0	0	4	2	0	0	4
C 0	0	2	3	1	0	2
G 0	4	2	1	2	5	3
C 0	2	3	1	0	3	4

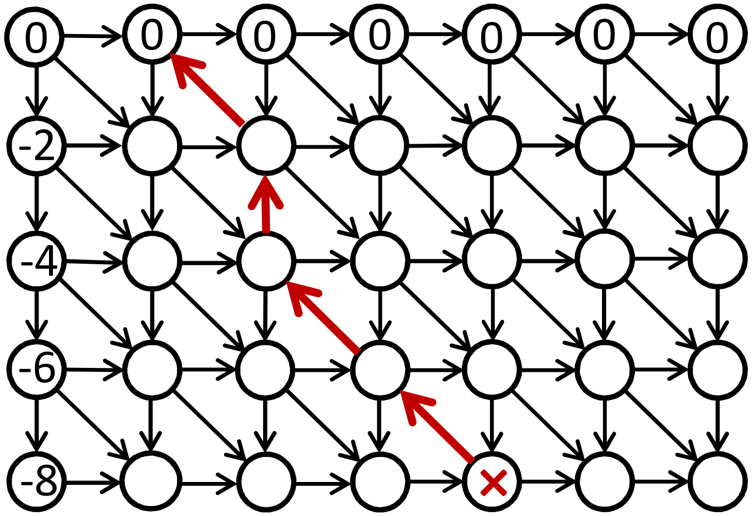
<u>Scores:</u> Match +4 Mismatch -1 Gap -2

Semi-Global Alignment



Biological problem: Find approximate matches to a given pattern within a large sequence. For example, seek promoters within a DNA sequence, or a copies of a domain within a protein sequence.

Solution: *Semi-global alignment*. Needleman-Wunsch algorithm with two modifications: 1) Penalize end gaps in the pattern, but not in the long sequence; 2) Trace back from the highest scoring node along the long edge of the path graph.



Erickson, B.W. & Sellers, P.H. (1983) "Recognition of patterns in genetic sequences." In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, D. Sankoff & J.B. Kruskal (eds.), pp. 55-91, Addison-Wesley, Reading, MA.