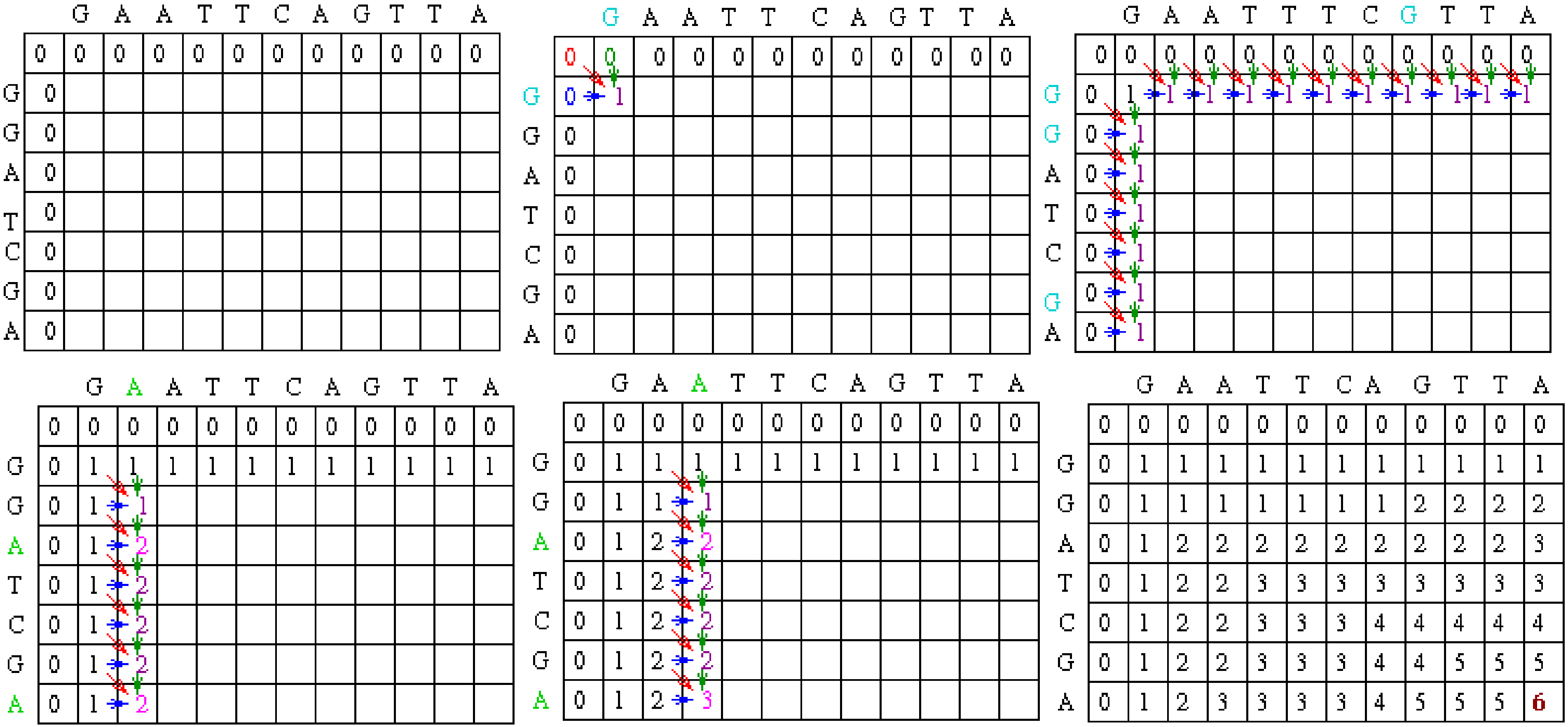
Recurrence Relation for Needleman-Wunsch-Sellers

- S[I, J] = MAXIMUM {

$$\begin{aligned} &S[I-1, J-1] + \delta(V[I], W[J]),\\ &S[I-1, J] + \delta(V[I], —),\\ &S[I, J-1] + \delta(—, W[J])\\ &\} \end{aligned}$$

Global Alignment: An example

```
V: G A A T T C A G T T A
W: G G A T C G A
```



		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0											
G	0											
A	0											
T	0											
C	0											
G	0											
A	0											

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1										
G	0											
A	0											
T	0											
C	0											
G	0											
A	0											

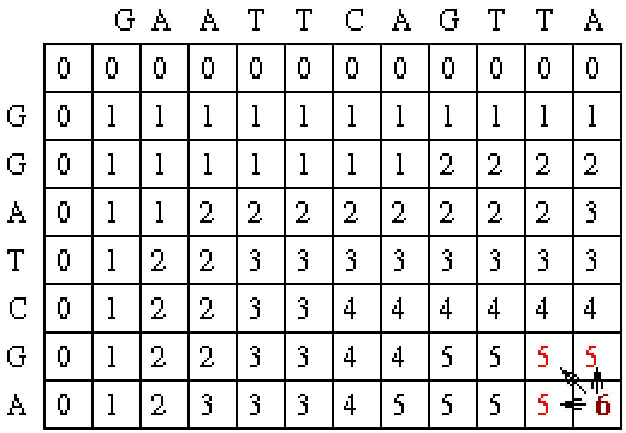
		G	A	A	T	T	T	C	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1										
A	0	1										
T	0	1										
C	0	1										
G	0	1										
A	0	1										

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1									
A	0	1	2									
T	0	1	2									
C	0	1	2									
G	0	1	2									
A	0	1	2									

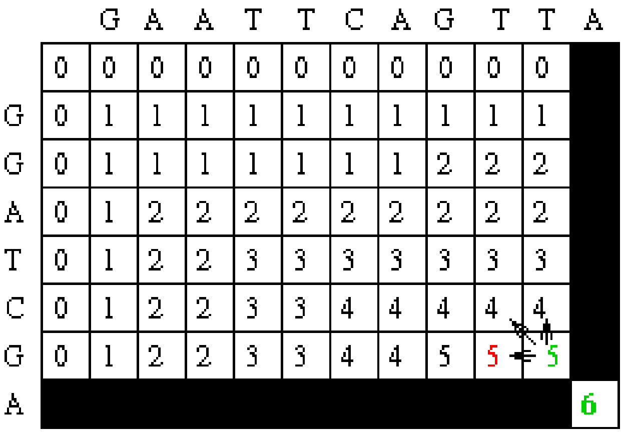
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1								
A	0	1	2	2								
T	0	1	2	2								
C	0	1	2	2								
G	0	1	2	2								
A	0	1	2	3								

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	3	4	4	4	4	4
G	0	1	2	2	3	3	3	4	4	5	5	5
A	0	1	2	3	3	3	3	4	5	5	5	6

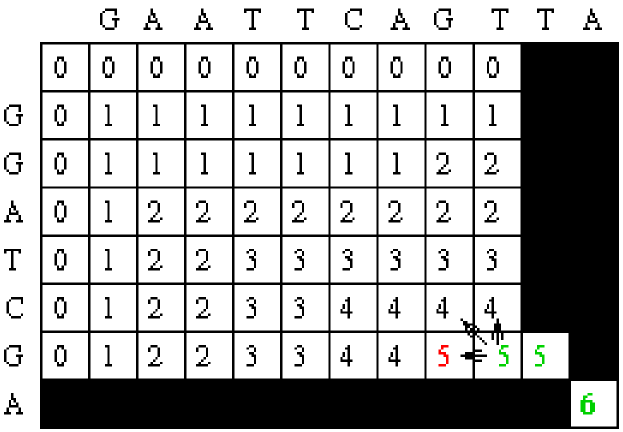
Traceback



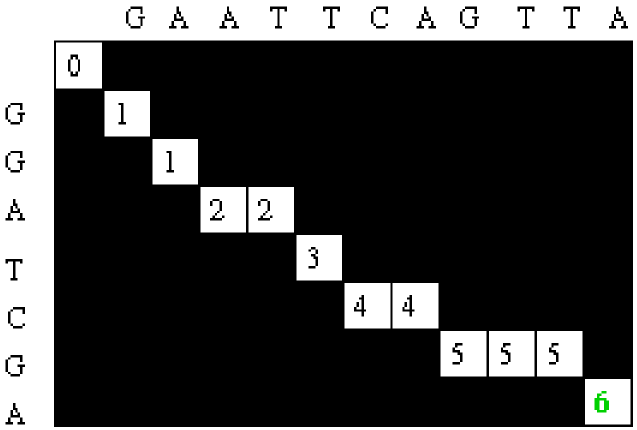
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6



		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	
G	0	1	1	1	1	1	1	1	1	1	1	
G	0	1	1	1	1	1	1	1	2	2	2	
A	0	1	2	2	2	2	2	2	2	2	2	
T	0	1	2	2	3	3	3	3	3	3	3	
C	0	1	2	2	3	3	4	4	4	4	4	
G	0	1	2	2	3	3	4	4	5	5	5	
A												6



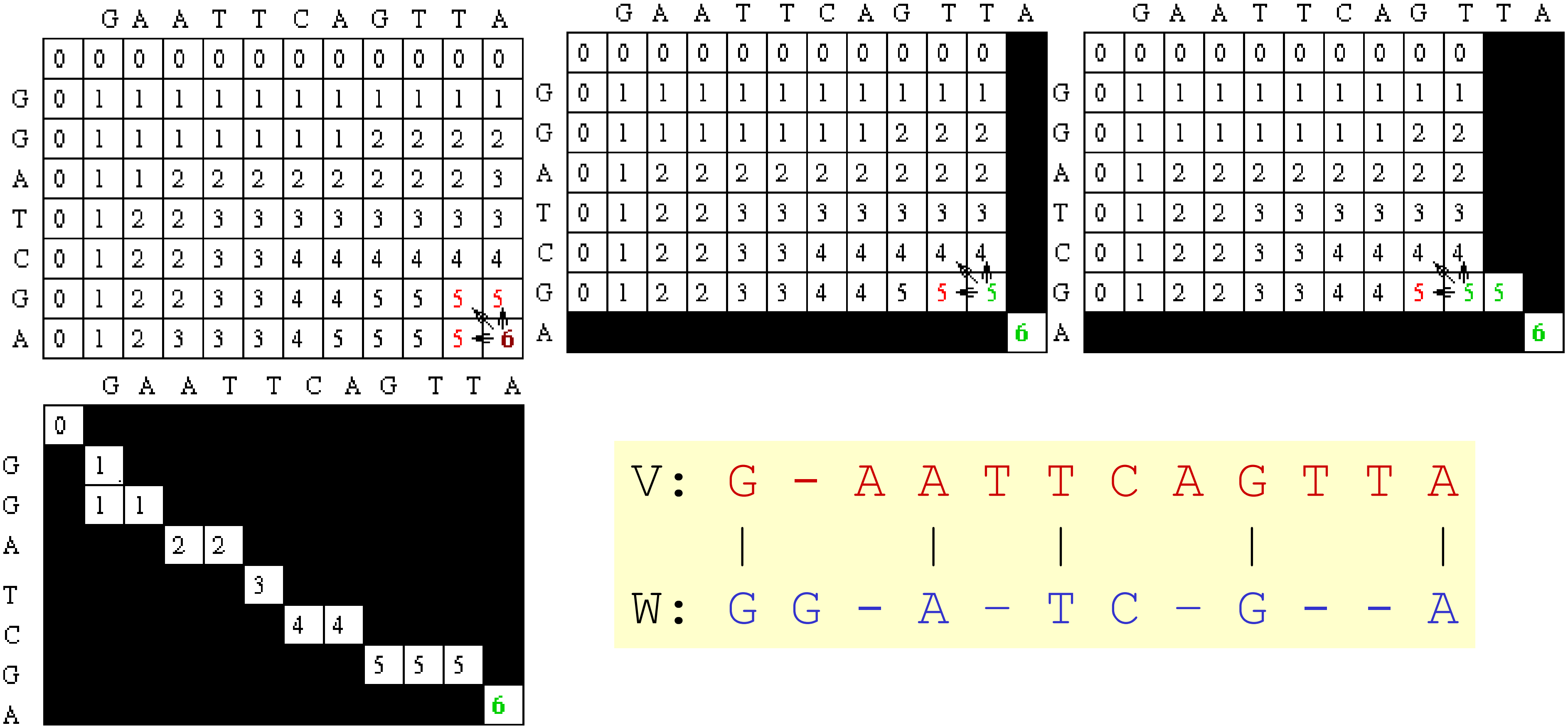
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0		
G	0	1	1	1	1	1	1	1	1	1		
G	0	1	1	1	1	1	1	1	2	2		
A	0	1	2	2	2	2	2	2	2	2		
T	0	1	2	2	3	3	3	3	3	3		
C	0	1	2	2	3	3	4	4	4	4		
G	0	1	2	2	3	3	4	4	5	5	5	
A												6



		G	A	A	T	T	C	A	G	T	T	A
	0											
G		1										
G			1									
A				2	2							
T						3						
C							4	4				
G									5	5	5	
A												6

```
V: G A A T T C A G T T A
   |   |   | |   |     |
W: G G A - T C - G - - A
```

Alternative Traceback



		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	
G	0	1	1	1	1	1	1	1	1	1	1	
G	0	1	1	1	1	1	1	1	2	2	2	
A	0	1	2	2	2	2	2	2	2	2	2	
T	0	1	2	2	3	3	3	3	3	3	3	
C	0	1	2	2	3	3	4	4	4	4	4	
G	0	1	2	2	3	3	4	4	5	5	5	
A												6

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0		
G	0	1	1	1	1	1	1	1	1	1		
G	0	1	1	1	1	1	1	1	2	2		
A	0	1	2	2	2	2	2	2	2	2		
T	0	1	2	2	3	3	3	3	3	3		
C	0	1	2	2	3	3	4	4	4	4		
G	0	1	2	2	3	3	4	4	5	5	5	
A												6

		G	A	A	T	T	C	A	G	T	T	A
	0											
G		1										
G		1	1									
A				2	2							
T						3						
C							4	4				
G									5	5	5	
A												6

```
V: G - A A T T C A G T T A
   |     |   |     |     |
W: G G - A - T C - G - - A
```

Improved Traceback

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	×1	←1	←1	←1	←1	←1	←1	×1	←1	←1	←1
G	0	×1	↑1	↑1	↑1	↑1	↑1	↑1	×2	←2	←2	←2
A	0	↑1	↑1	×2	←2	←2	←2	×2	↑2	↑2	↑2	×3
T	0	↑1	←2	↑2	×3	×3	←3	←3	←3	×3	×3	↑3
C	0	↑1	↑2	↑2	↑3	↑3	×4	←4	←4	←4	←4	←4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	×5	←5	←5	←5
A	0	↑1	↑2	×3	↑3	↑3	↑4	×5	↑5	↑5	↑5	×6

```
V: G A - A T T C A G T T A
   |     |   | |   |     |
W: G - G A - T C - G - - A
```

Subproblems

- Optimally align V[1..I] and W[1..J] for every possible values of I and J.
- Having optimally aligned
  - V[1..I-1] and W[1..J-1]
  - V[1..I] and W[1..J-1]
  - V[1..I-1] and W[1, J]

It is possible to optimally align V[1..I] and W[1..J]

Time Complexity

- O(mn),

  where m = length of V,

  and n = length of W.

Generalizations of Similarity Function

- Mismatch Penalty = α
- Spaces (Insertions/Deletions, InDels) = β
- Affine Gap Penalties:

  (Gap open, Gap extension) = (γ,δ)
- Weighted Mismatch = $\Phi(a,b)$
- Weighted Matches = $\Omega(a)$

Alternative Scoring Schemes

		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
G	-2	× 1	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
G	-3	↑-1	× -1	← -3	← -4	← -5	← -6	← -7	× -5	← -7	← -8	← -9
A	-4	↑-2	× 0	× 0	← -2	← -3	← -4	← -5	← -6	← -7	← -8	× -7
T	-5	↑-3	↑ -2	↑-2	× 1	← -1	← -2	← -3	← -4	← -5	← -6	← -7
C	-6	↑-4	↑ -3	↑-3	↑-1	× -1	× 0	← -2	← -3	← -4	← -5	← -6
G	-7	↑-5	↑-4	↑-4	↑-2	↑-3	↑-2	× -2	× -1	← -3	← -4	← -5
A	-8	↑-6	↑-5	↑-5	↑-3	↑-4	↑-3	× -1	↑-3	× -3	× -5	× -3

Match +1
Mismatch –2
Gap (-2, -1)

```
V: G A A T T C A G T T A
   |   | |   |   |     |
W: G G A T - C - G - - A
```

Local Sequence Alignment

- Example: comparing long stretches of anonymous DNA; aligning proteins that share only some motifs or domains.
- Smith-Waterman Algorithm

Recurrence Relations (Global vs Local Alignments)

- $$S[I, J] = \text{MAXIMUM} \left\{ \begin{array}{l} S[I-1, J-1] + \delta(V[I], W[J]), \\ S[I-1, J] + \delta(V[I], —), \\ S[I, J-1] + \delta(—, W[J]) \end{array} \right\}$$

Global Alignment

- $$S[I, J] = \text{MAXIMUM} \left\{ \begin{array}{l} 0, \\ S[I-1, J-1] + \delta(V[I], W[J]), \\ S[I-1, J] + \delta(V[I], —), \\ S[I, J-1] + \delta(—, W[J]) \end{array} \right\}$$

Local Alignment

Local Alignment: Example

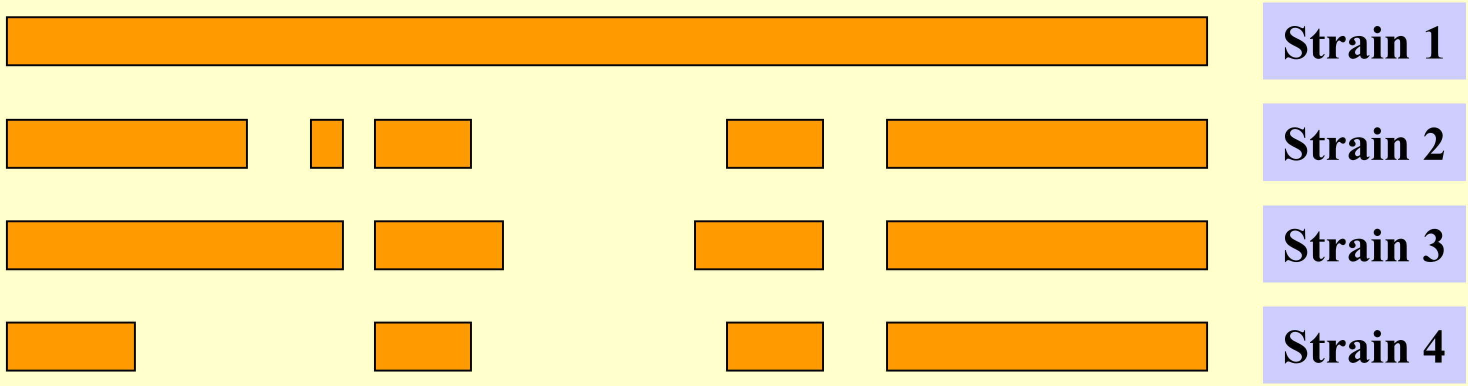
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	× 1	0	0	0	0	0	0	0	0	0	0
G	0	× 1	← 0	0	0	0	0	0	× 1	0	0	0
A	0	0	× 2	× 1	0	0	0	× 1	0	0	0	× 1
T	0	0	↑ 0	× 1	× 2	← 1	0	0	0	× 1	× 1	0
C	0	0	0	0	↑ 0	× 0	× 2	0	0	0	0	0
G	0	0	0	0	0	0	0	0	× 1	0	0	0
A	0	0	× 1	× 1	0	0	0	× 1	0	0	0	× 1

Match +1
Mismatch –1
Gap (-1, -1)

```
V: - G A A T T C A G T T A
     |   | |   |
W: G G - A T - C - G - - A
```

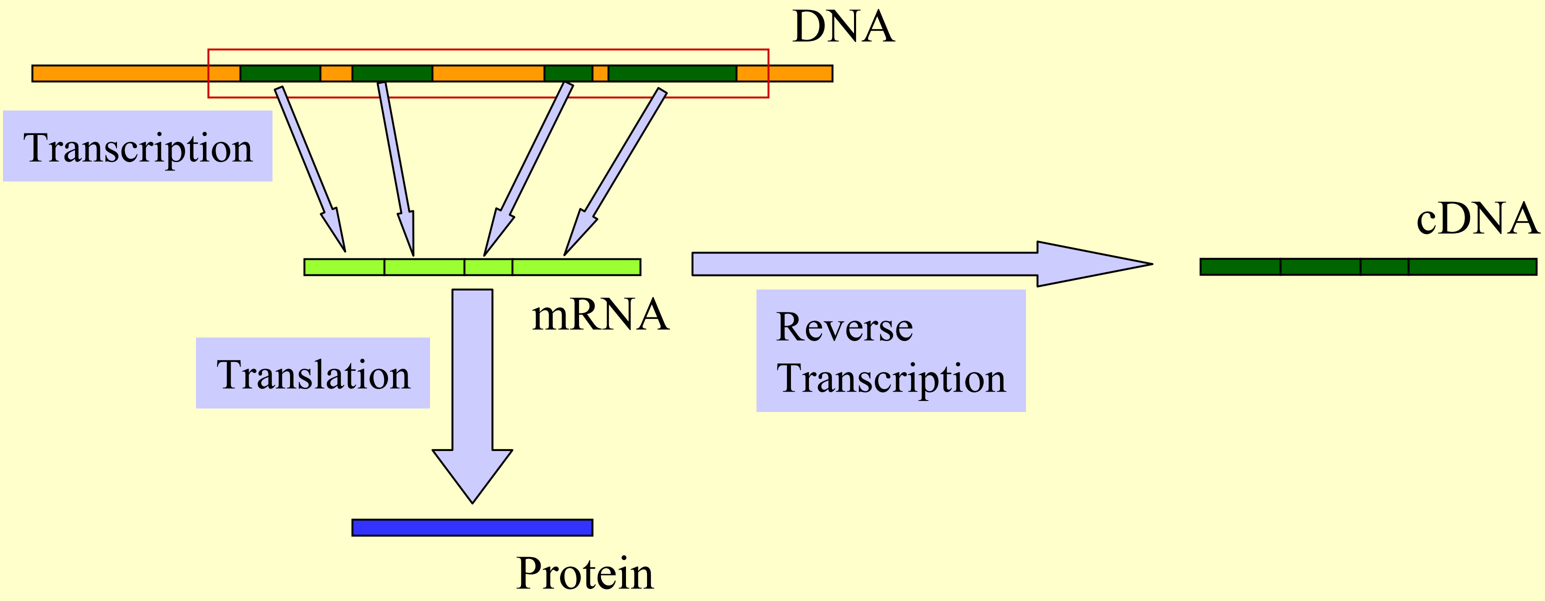
Why Gaps?

- Example: Finding the gene site for a given (eukaryotic) cDNA requires "gaps".
- Example: HIV-virus strains



What is cDNA?

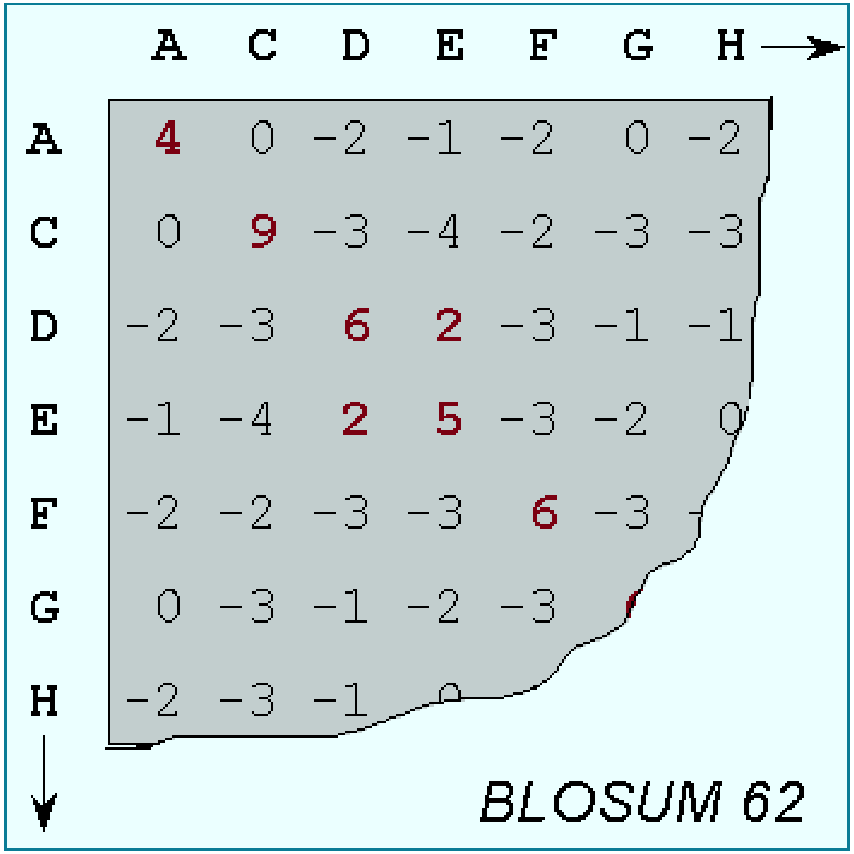
- cDNA = Copy DNA



Properties of Smith-Waterman Algorithm

- How to find all regions of “high similarity”?
  - Find all entries above a threshold score and traceback.
- What if: Matches = 1 & Mismatches/spaces = 0?
  - Longest Common Subsequence Problem
- What if: Matches = 1 & Mismatches/spaces = -∝?
  - Longest Common Substring Problem
- What if the average entry is positive?
  - Global Alignment

How to score mismatches?



	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	-
G	0	-3	-1	-2	-3	6	
H	-2	-3	-1	0			

BLOSUM 62

BLOSUM n Substitution Matrices

- For each amino acid pair a, b
  - For each BLOCK
    - Align all proteins in the BLOCK
    - Eliminate proteins that are more than n% identical
    - Count F(a), F(b), F(a,b)
    - Compute Log-odds Ratio

$$\log\left(\frac{F(a,b)}{F(a)F(b)}\right)$$