Sequence Comparison: Local Alignment

Genome 373

Genomic Informatics

Elhanan Borenstein

Review: Global Alignment

- Three Possible Moves:
  - A diagonal move aligns a character from each sequence.
  - A horizontal move aligns a gap in the seq along the left edge
  - A vertical move aligns a gap in the seq along the top edge.

- The move you keep is the best scoring of the three.

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4	-5				
A	-8	?				
T	-12					
A	-16					
C	-20					

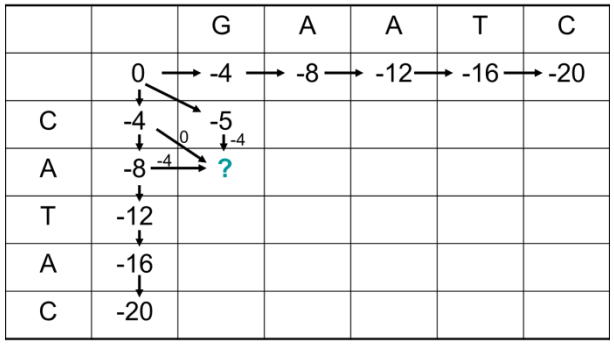
Review: Global Alignment

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Fill DP matrix from upper left to lower right.
Traceback alignment from lower right corner.

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	**17**

DP in equation form



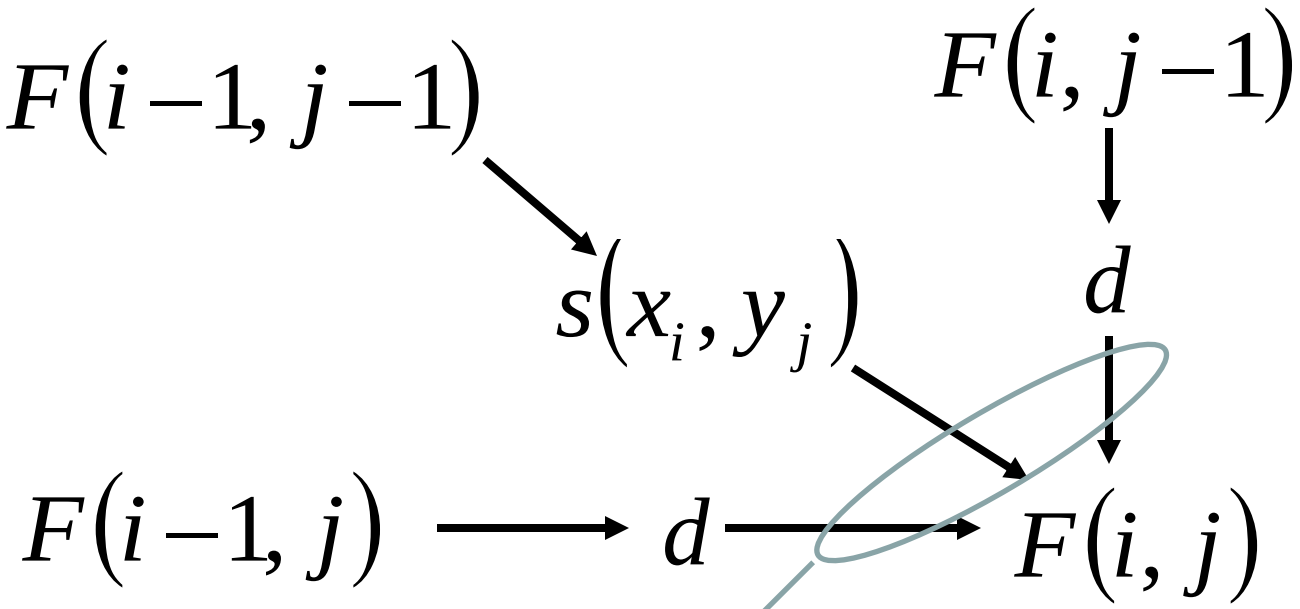
- Align sequence ***x*** and ***y***.
- ***F*** is the DP matrix; ***s*** is the substitution matrix; ***d*** is the linear gap penalty.

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases}$$

DP equation graphically

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5				
A	-8	?				
T	-12					
A	-16					
C	-20					



take the max of these three

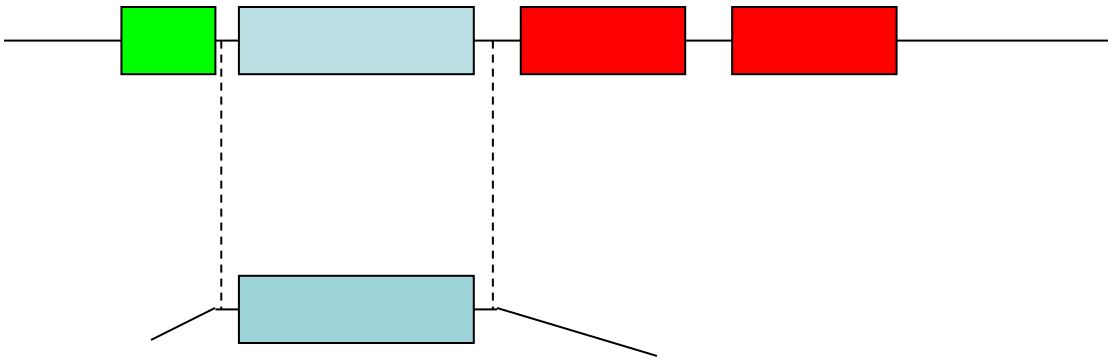
Local alignment

Mission:

Find best partial alignment between two sequences.

Why?

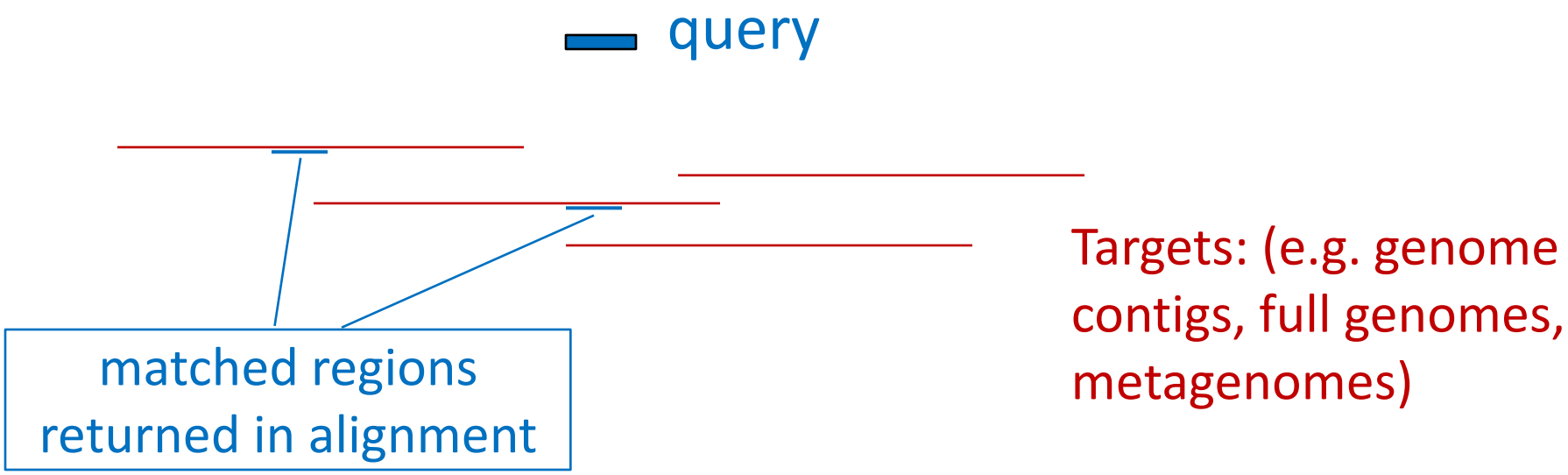
Local alignment



- A single-domain protein may be similar only to one region within a multi-domain protein.
- A DNA query may align to a small part of a genome/genomes/metagenomes.
- An alignment that spans the complete length of both sequences may be undesirable.

BLAST does local alignments

- Typical search has a short query against long targets.
- The alignments returned show only the well-aligned match region of both query and target.



Remember: Global alignment DP

- Align sequence x and y.
- F is the DP matrix; s is the substitution matrix; d is the linear gap penalty.

$$F(0,0) = 0$$

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases}$$

Local alignment DP

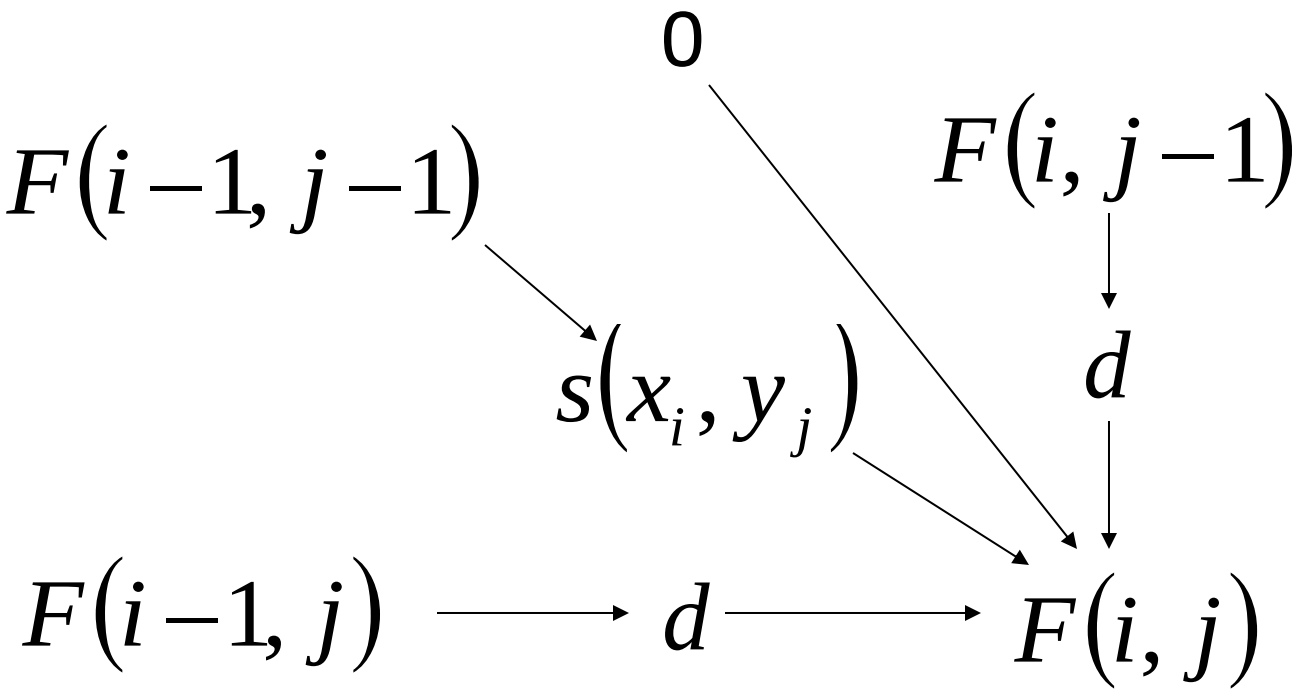
- Align sequence x and y.
- F is the DP matrix; s is the substitution matrix; d is the linear gap penalty.

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ 0 \end{cases}$$

(corresponds to start of alignment)

Local DP in equation form



keep max of these four values

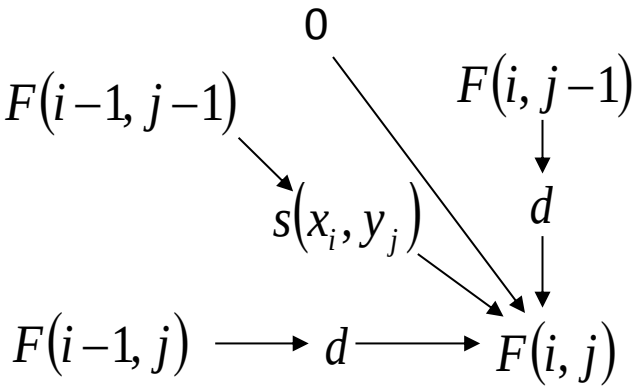
A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5

initialize the same way as for global alignment

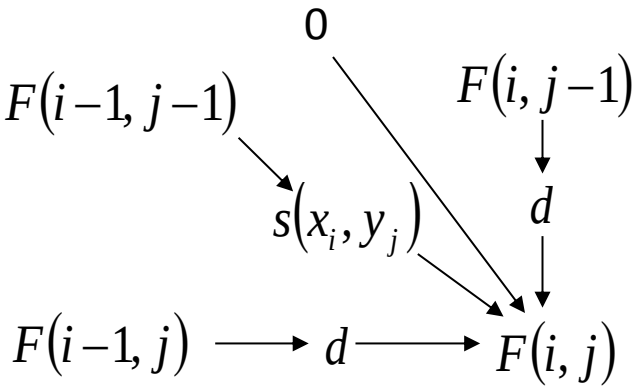
		A	A	G
	0			
A				
G				
C				



A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5

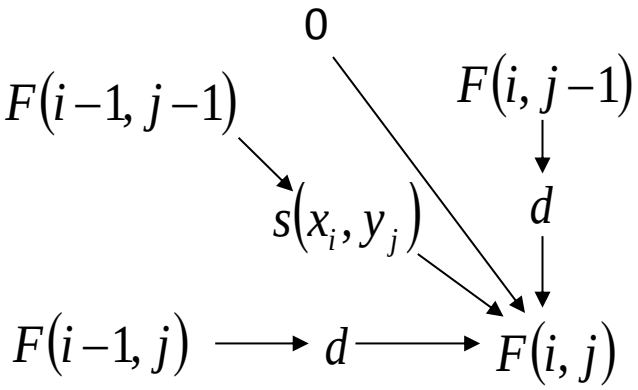


		A	A	G
	0	?	?	?
A	?			
G	?			
C	?			

A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5

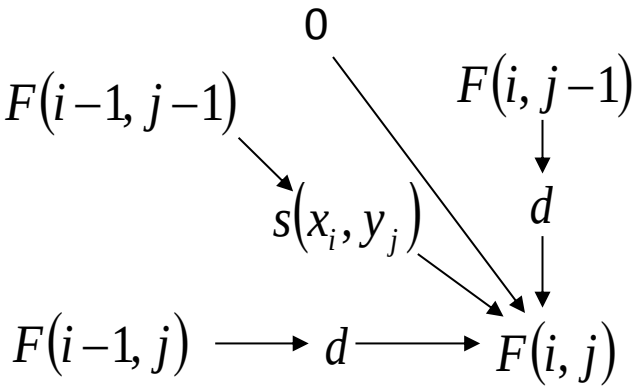


		A	A	G
	0	0	0	0
A	0	?		
G	0			
C	0			

A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5



		A	A	G
	0	0	0	0
A	0	2	-5	
G	0	-5	0	
C	0			

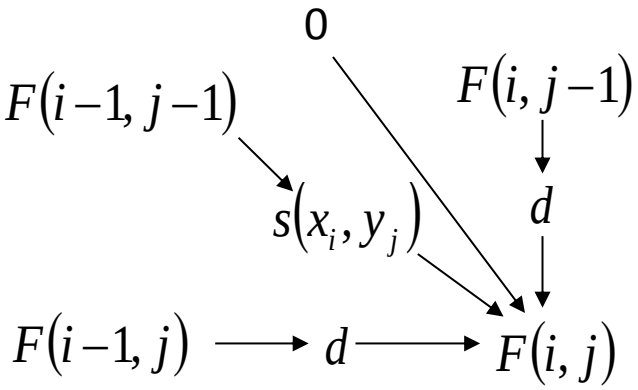
A
A

A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5

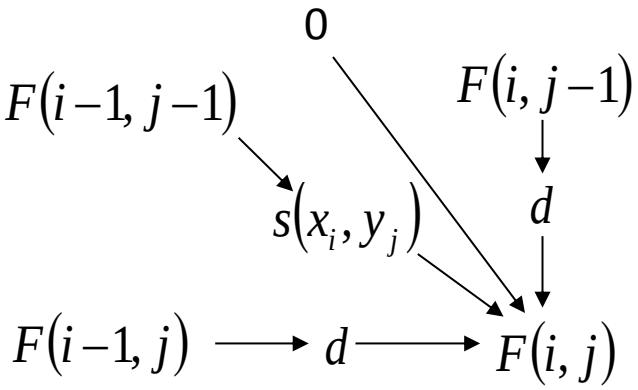
		A	A	G
	0	0	0	0
A	0	2		
G	0			
C	0			



A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5

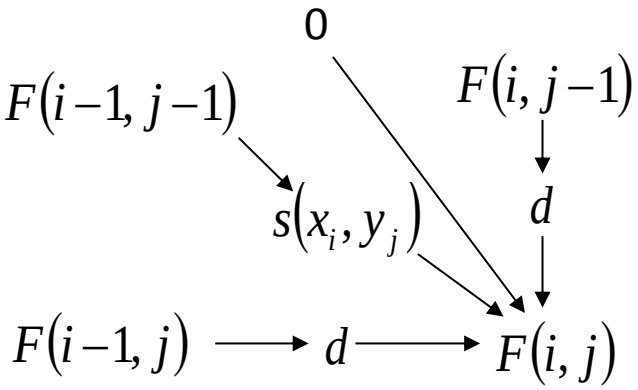


		A	A	G
	0	0	0	0
A	0	2		
G	0	?		
C	0	?		

A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5



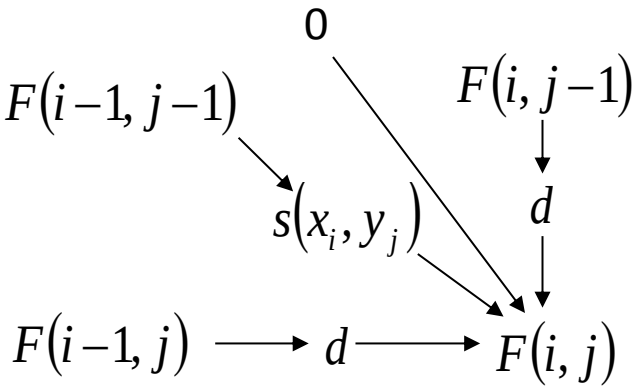
		A	A	G
	0	0	0	0
A	0	2		
G	0	0		
C	0	?		

(signify no preceding alignment with no arrow)

A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5

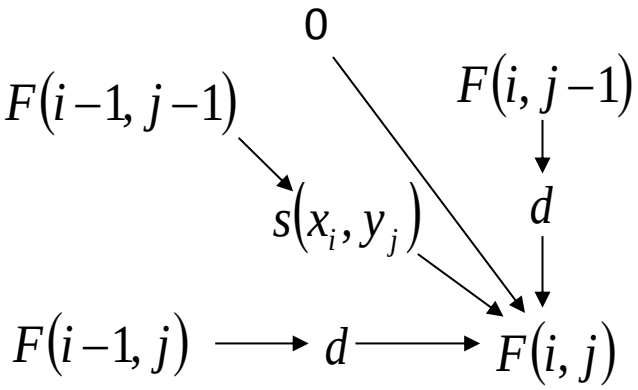


		A	A	G
	0	0	0	0
A	0	2	?	
G	0	0	?	
C	0	0	?	

A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5

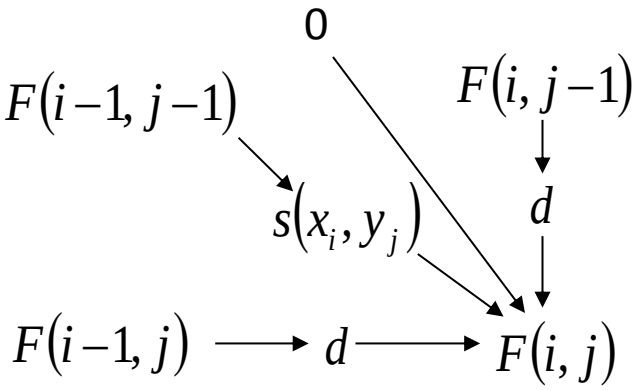


		A	A	G
	0	0	0	0
A	0	2	2	
G	0	0	0	
C	0	0	0	

A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5



		A	A	G
	0	0	0	0
A	0	2	2	?
G	0	0	0	?
C	0	0	0	?

A simple example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5

		A	A	G
	0	0	0	0
A	0	2	2	0
	0	0	0	4
	0	0	0	0

But … how do we traceback?

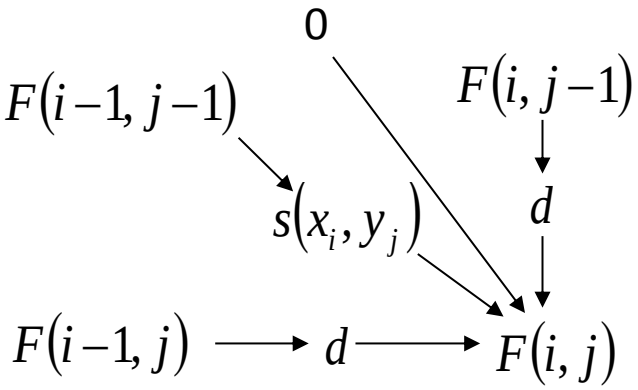
Traceback

AG
AG

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

d = -5

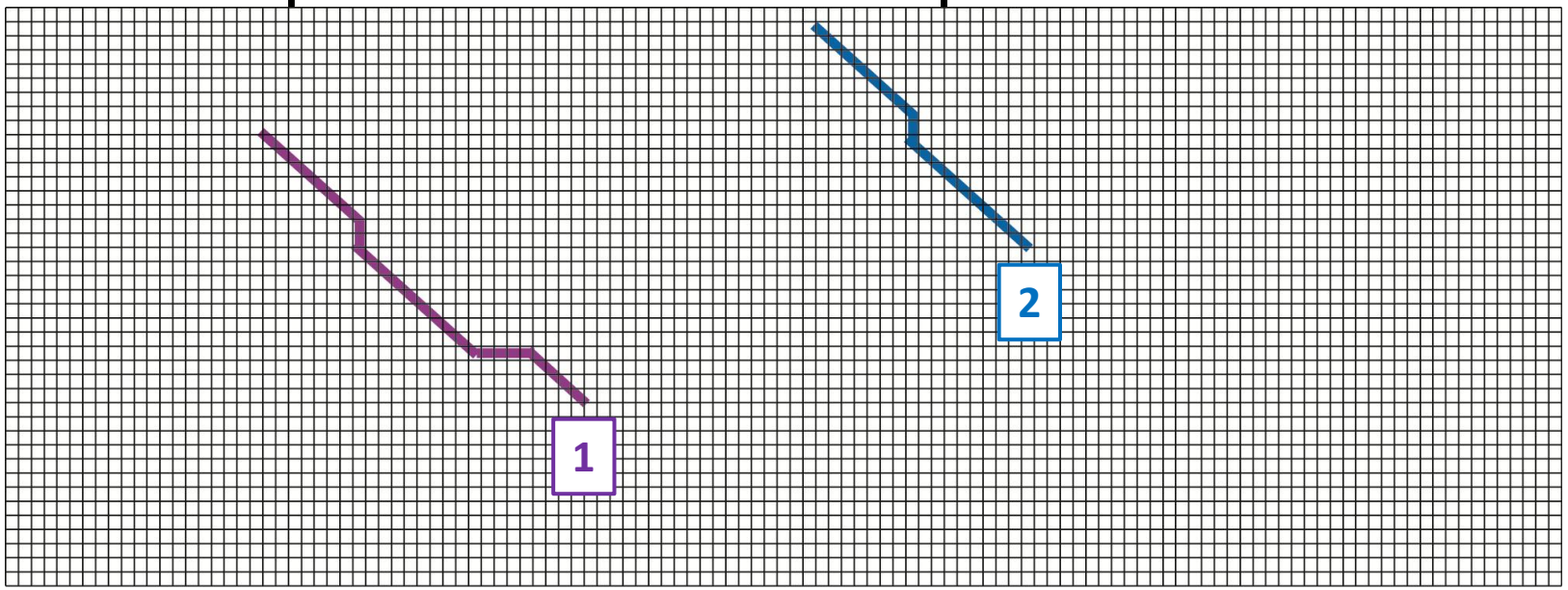
		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0



Start traceback at highest score anywhere in matrix, follow arrows back until you reach 0

Multiple local alignments

- Traceback from highest score, setting each DP matrix score along traceback to zero.
- Now traceback from the remaining highest score, etc.
- The alignments may or may not include the same parts of the two sequences.



Local alignment

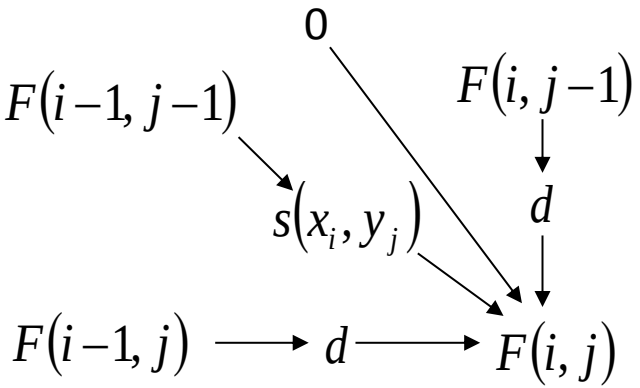
- Two differences from global alignment:
  - If a DP score is negative, replace with 0.
  - Traceback from the highest score in the matrix and continue until you reach 0.
- Global alignment algorithm: *Needleman-Wunsch*.
- Local alignment algorithm: *Smith-Waterman*.

(Some) Specific Uses for Alignments

- Make a pairwise or multiple alignment (duh)
- Test whether two sequences share a common ancestor (i.e. are significantly related)
- Find matches to a sequence in a large database
- Build a sequence tree (phylogenetic tree)
- Make a genome assembly (find overlaps of sequence reads)
- Map sequence reads to a reference genome

Another example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2



Find the optimal **local** alignment of AAG and GAAGGC.

Use a gap penalty of d = -5.

		A	A	G
	0	0	0	0
G	0	0	0	2
A	0	2	2	0
A	0	2	4	0
G	0	0	0	6
G	0	0	0	2
C	0	0	0	0

Traceback

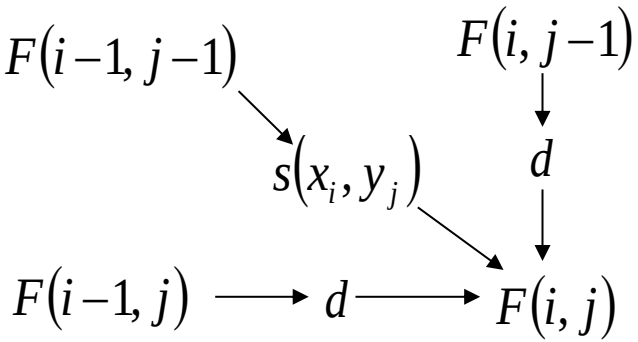
		A	A	G
	0	0	0	0
G	0	0	0	2
A	0	2	2	0
A	0	2	4	0
G	0	0	0	6
G	0	0	0	2
C	0	0	0	0

AAG

AAG

Compare with the Best GLOBAL Alignment

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2



(contrast with the best local alignment)

Find the optimal **Global** alignment of AAG and GAAGGC.

Use a gap penalty of d = -5.

		A	A	G
	0	-5	-10	-15
G	-5			
A	-10			
A	-15			
G	-20			
G	-25			
C	-30			