

Sequence comparison: Local alignment

Genome 559: Introduction to Statistical
and Computational Genomics
Prof. James H. Thomas

http://faculty.washington.edu/jht/GS559_2017/

Review - global alignment

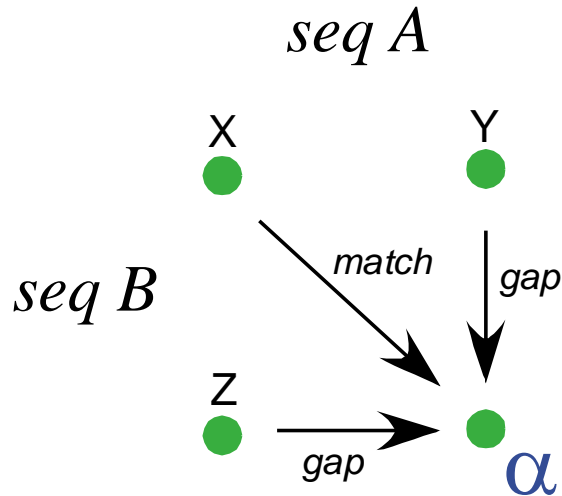
		G	A	A	T	C
	0	-4	-8	-12	-16	-20
U	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

Fill DP matrix from upper left to lower right,
trace back alignment from lower right corner.

start
traceback

FYI - informal inductive proof of best alignment path

Consider the last step in the best alignment path to node α below from its adjacent nodes, where X , Y , and Z are scores of the best alignments up to those nodes. We can reach node α by three possible paths: an A-B match, a gap in sequence A or a gap in sequence B:

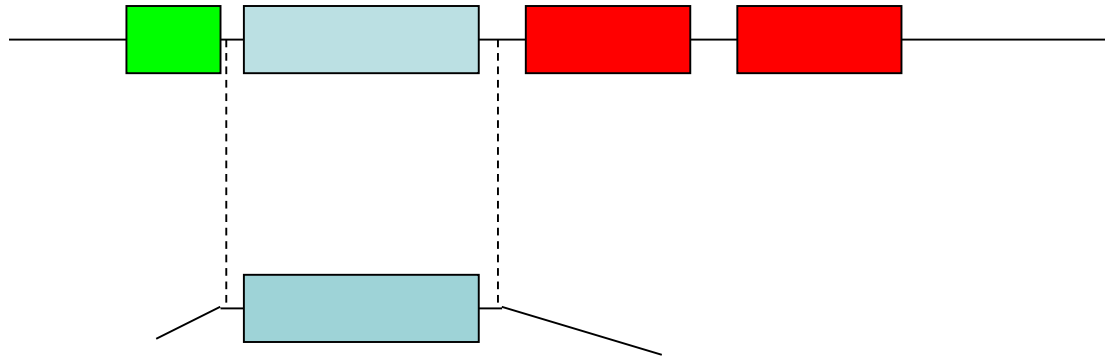


The best-scoring path to α is the maximum of:

- $X + \text{match}$
- $Y + \text{gap}$
- $Z + \text{gap}$

BUT the best paths to X , Y , and Z are analogously the max of their three upstream possibilities, etc. Inductively QED.

Local alignment

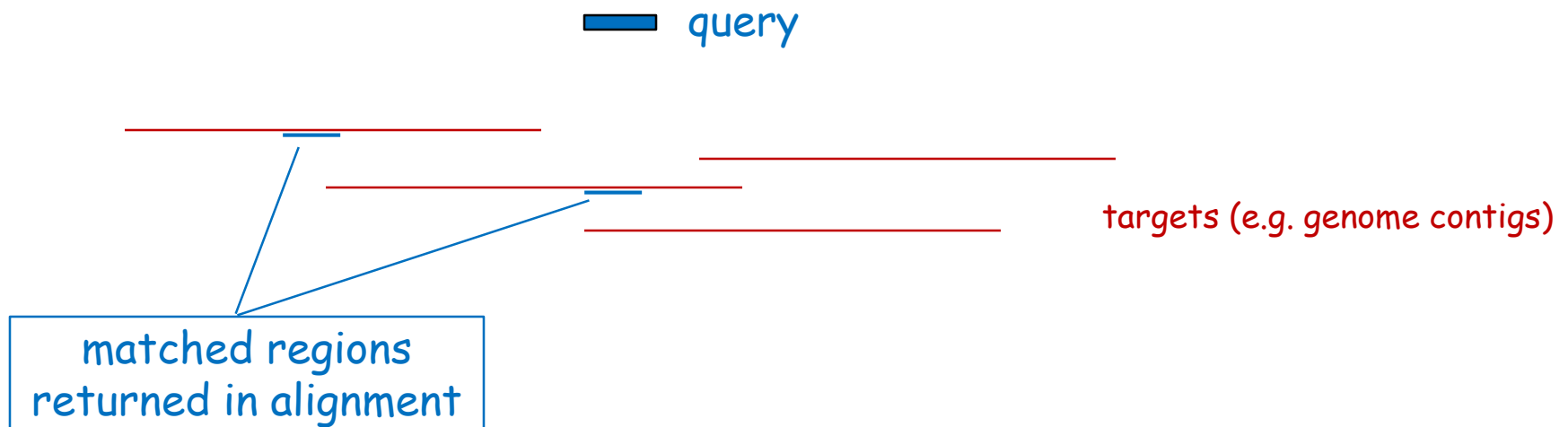


- A single-domain protein may be similar to only one region within a multi-domain protein.
- A DNA query may align to a small part of a genome.
- An alignment that spans the complete length of both sequences may be undesirable.

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.



Review - 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.

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

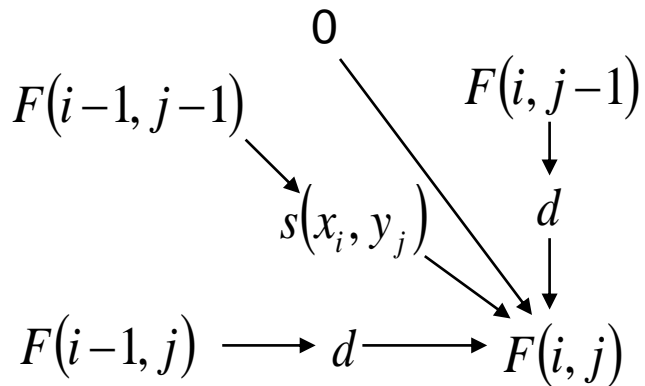
$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ 0 \end{cases}$$

for local 

A (very) 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

$d = -5$



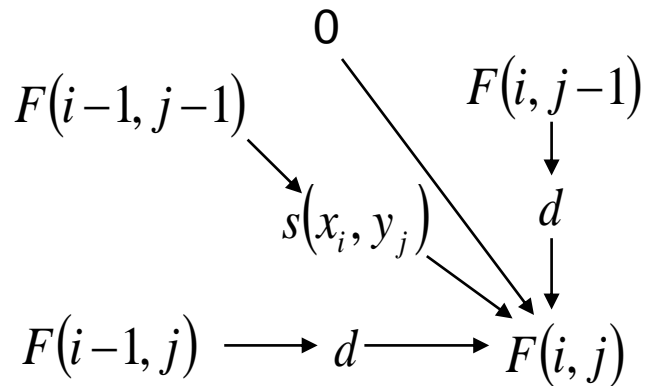
initialize the same way
as for global alignment

		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

$d = -5$

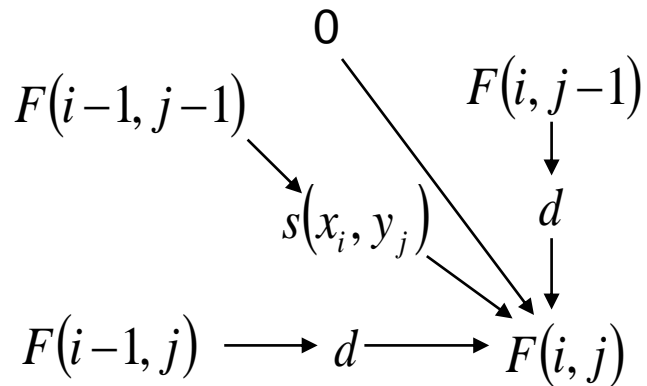


		A	A	G
	0			
A	?			
G	?			
C	?			

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

$d = -5$

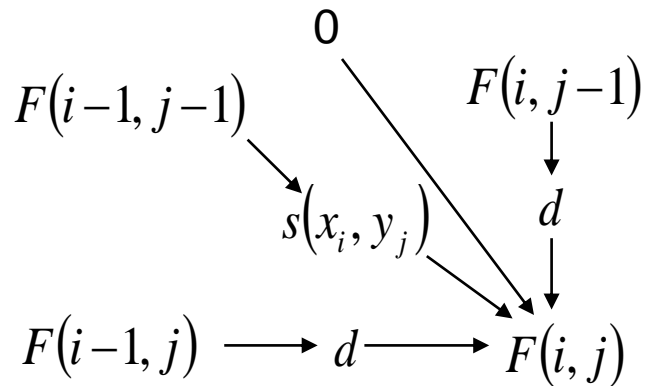


		A	A	G
	0	0		
A	0	?		
G	0			
C	0			

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

$d = -5$



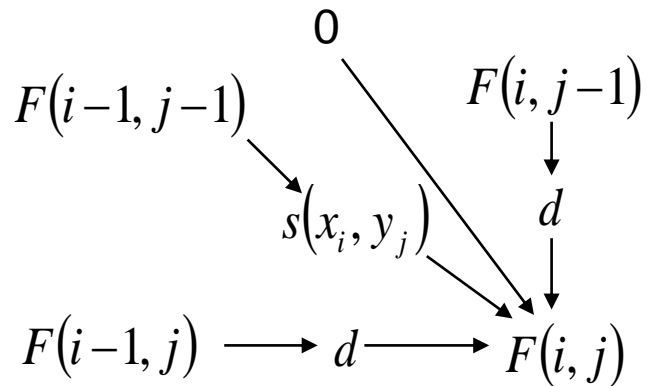
		A	A	G
	0	0		
A	0	2	-5	
G	0	-5	0	
C	0			

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

$d = -5$

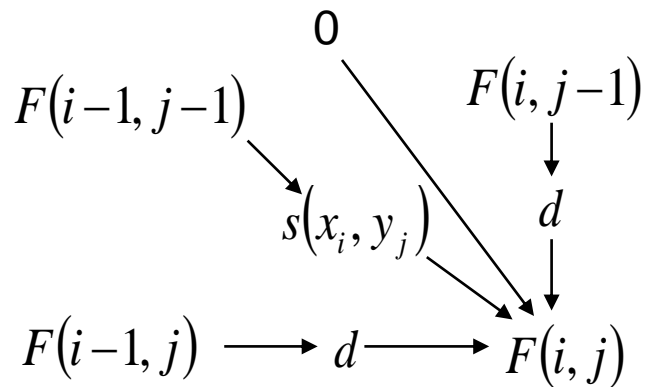


		A	A	G
	0	0		
A	0	2		
G	0			
C	0			

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

$d = -5$

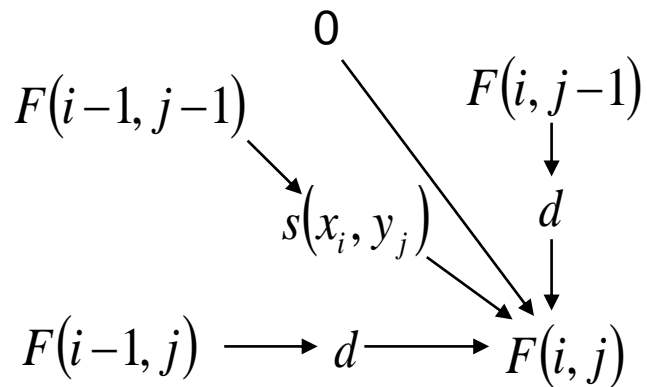


		A	A	G
	0	0		
A	0	2		
G	0	?		
C	0	?		

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

$d = -5$

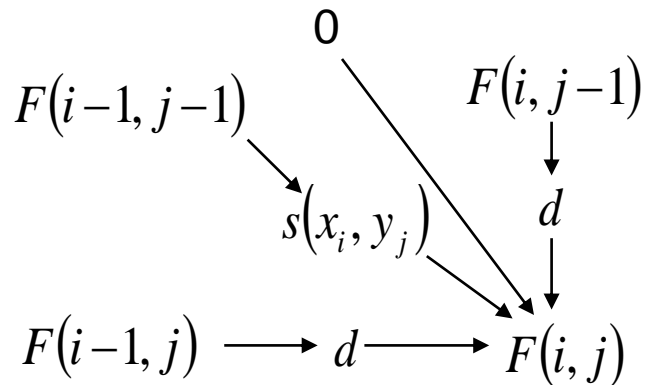


		A	A	G
	0	0		
A	0	2		
G	0	-5	-3	
C	0	-5	0	

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

$d = -5$



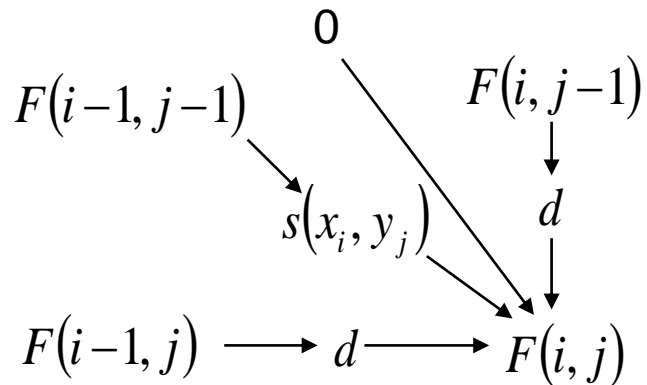
		A	A	G
	0	0		
A	0	2		
G	0	0		
C	0	0		

(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

$d = -5$



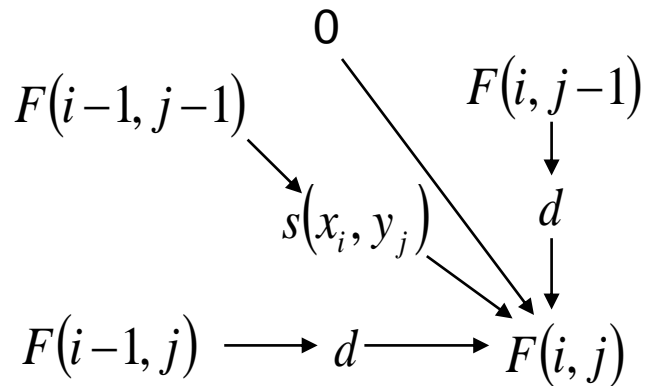
		A	A	G
	0	0	0	
A	0	2	?	
G	0	0	?	
C	0	0	?	

(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

$d = -5$

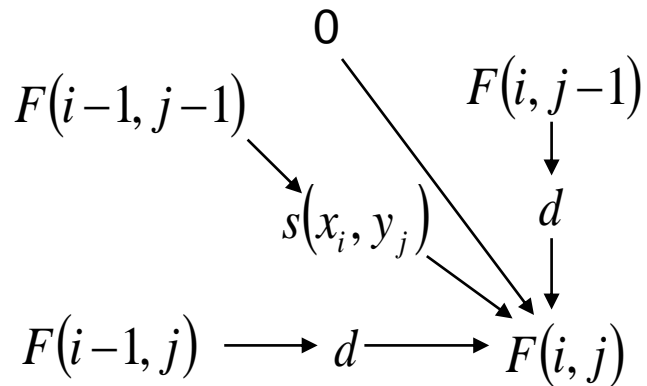


		A	A	G
	0	0	0	
A	0	2	2	
G	0	0	0	
C	0	0	0	

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

$d = -5$

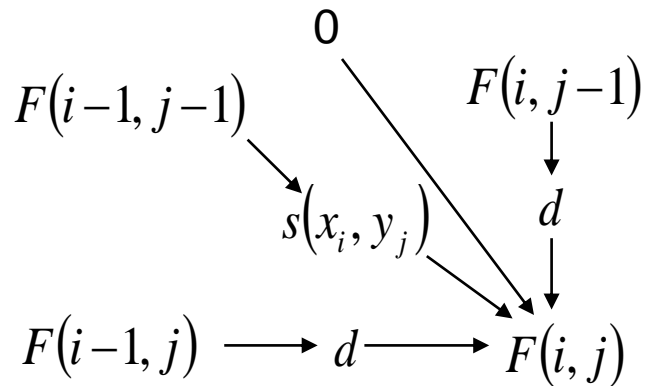


		A	A	G
	0	0	0	0
A	0	2	2	?
G	0	0	0	?
C	0	0	0	?

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
G	0	0	0	4
C	0	0	0	0

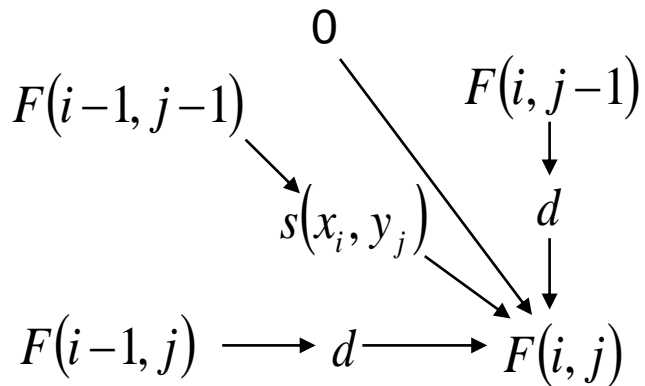
Traceback

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

$d = -5$

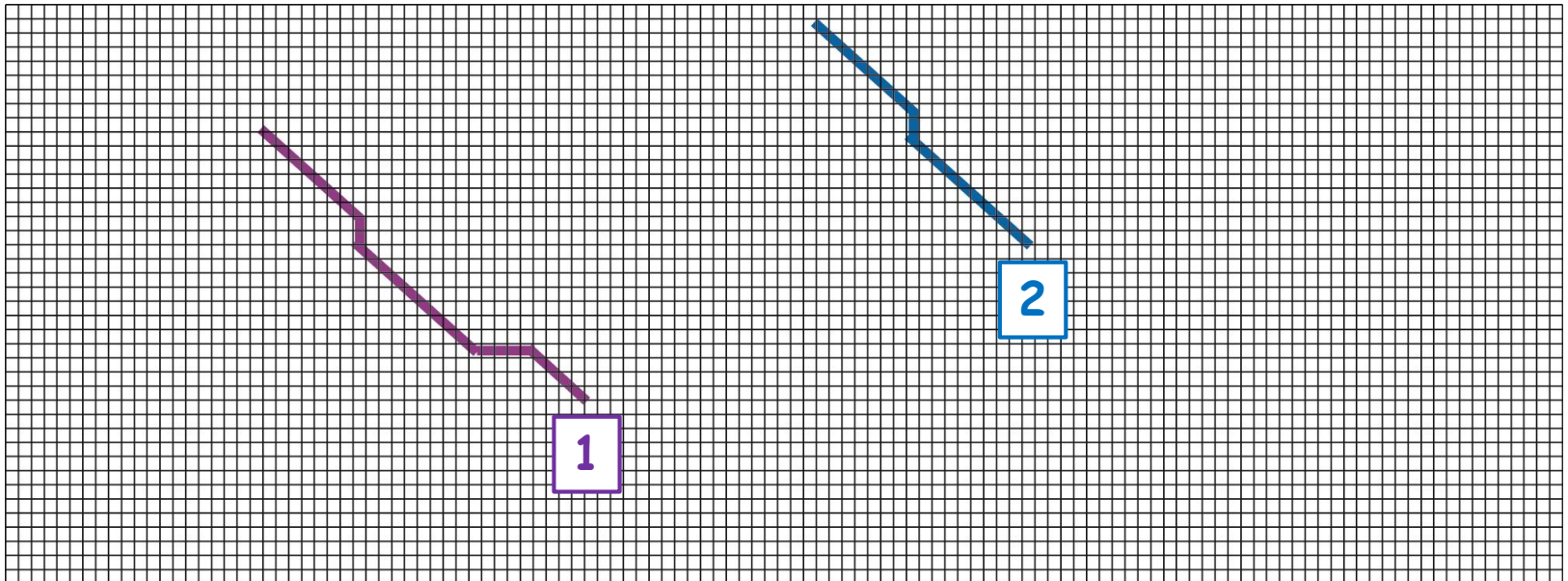
		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0



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*.

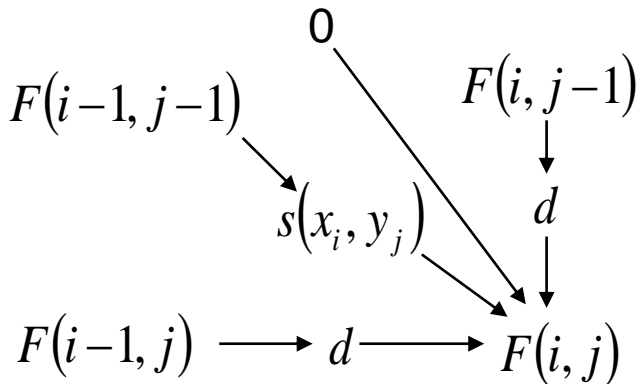
(some) specific uses for alignments

- 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)
- repeat-mask a genome sequence (find matches to a database of known repeats)
- 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

DP matrix

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

Traceback matrix

You don't actually need first row and column

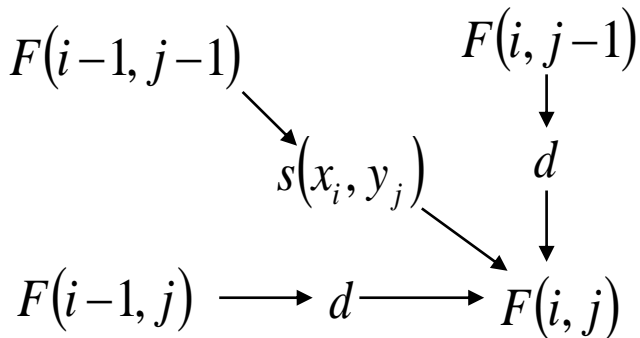
(-10)	(-10)	(-10)	(-10)
(-10)	-10	-10	0
(-10)	0	0	-10
(-10)	0	0	-10
(-10)	-10	-10	0
(-10)	-10	-10	0
(-10)	-10	-10	-10

0 = diagonal, -1 = gap left, +1 = gap top, -10 = no alignment

Problem - find 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

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			

(contrast with the best local alignment)