"When I was working on my Ph.D., I developed a computer algorithm to look for rapid changes in populations' DNA. Our DNA changes constantly over generations, but if certain changes spread through a population more quickly than others, they are probably the beneficial results of natural selection. This is the protection we give ourselves to survive."
-- Pardis Sabeti

Alignment type

S_1 = FTFTALILLAVAV
S_2 = FTALLLAAV

- Global
  - the whole of each sequence is included, end to end

```
FTFTALILLAVAV
F--TAL-LLA-AV
```

- Local
  - only the best matching parts of each sequence

```
FTALILLA
FTAL-LLA
```

- Glocal
  - global in query (small), local in reference (big)

```
FTFTALILL-AVAV
  FTAL-LLAAV
```

How many possibilities?

```
GAATC      GAAT-C     -GAAT-C
CATAC      C-ATAC     C-A-TAC

GAATC-     GAAT-C     GA-ATC
CA-TAC     CA-TAC     CATA-C
```

- How many different alignments of two sequences of length n exist?

Number of Possible Alignments

- given sequences of length m and n
- assume we don't count as distinct `C-` / `-G` and `-C` / `G-`
- we can have as few as 0 and as many as $\min\{m, n\}$ matched positions
- therefore the number of possible alignments is given by

$$\sum_{k=0}^{\min\{m,n\}} \binom{n}{k}\binom{m}{k} = \binom{n+m}{n}$$

Note that because gaps are ambiguous, I'm using the matched positions to define a distinct alignment. The number of positions that match will *always* be the same number (k) in each sequence. So each alignment chooses k positions to match from the original sequences.

Number of Possible Alignments

- there are $\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$

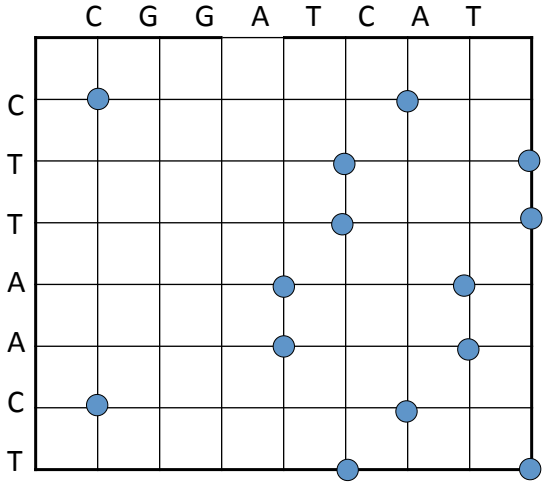
  possible global alignments for 2 sequences of length *n*

- e.g. two sequences of length 100 have $\approx 10^{77}$ possible alignments
- this is way TOO MANY TO ENUMERATE!!!
- but we can use *dynamic programming* to find an optimal alignment efficiently

Most basic comparison: identity

Sequence A：CTTAACT

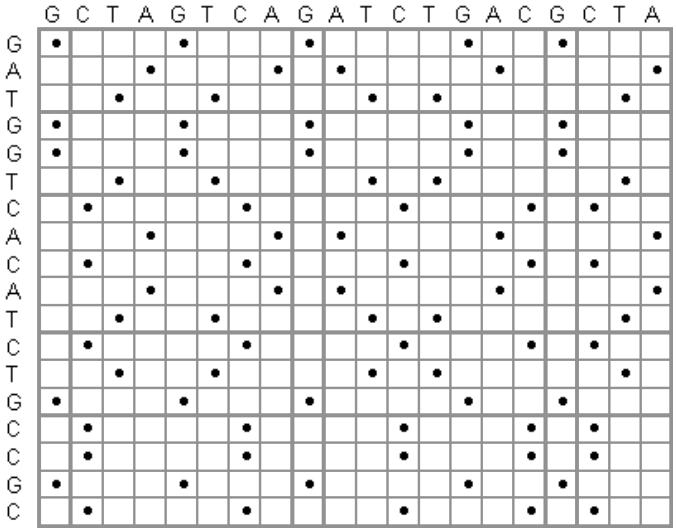
Sequence B：CGGATCAT



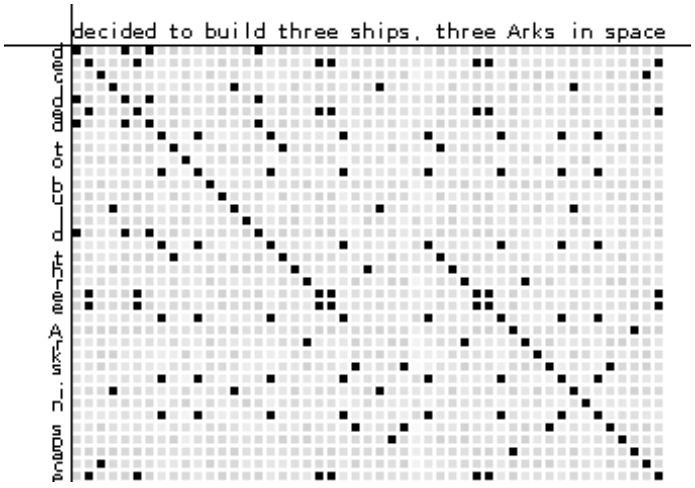
Nucleic Acids Dot Plots –

http://arbl.cvmbs.colostate.edu/molkit/dnadot/index.html

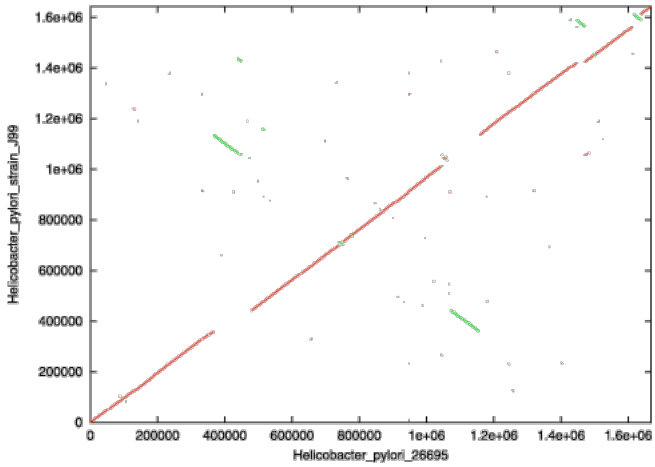
Your eye is a good pattern detector



You begin to see “diagonals” when there is a long string of matches.



Be sure to note axis labels!



Dynamic programming

- A systematic means of calculating the BEST alignment given a particular scoring scheme.

- Yes, it's a weird name.
- DP is closely related to recursion and to mathematical induction.
- We can prove that the resulting score is optimal.

Dynamic Programming

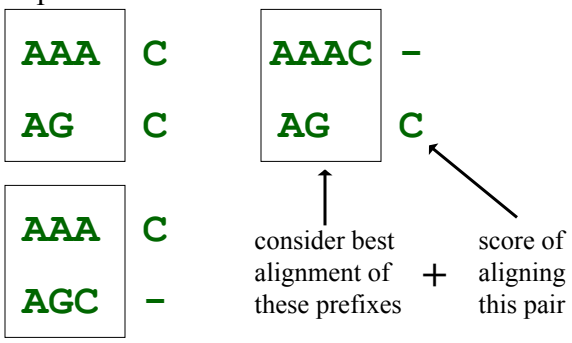
- Breaks the problem into overlapping subproblems
- Uses memorization to keep solutions to subproblems we've already seen.
- works for either DNA or protein sequences, although the substitution matrices used differ
- finds an optimal alignment

Three legal moves

- A match pairs the next two characters in each sequence.
- An insertion introduces a gap in the sequence along the top edge.
- A deletion introduces a gap in the sequence along the left edge.

Dynamic Programming Idea

- consider last step in computing alignment of **AAAC** with **AGC**
- three possible options; in each we'll choose a different pairing for end of alignment, and add this to the best alignment of previous characters



Global alignment algorithm: *Needleman-Wunsch*.

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

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

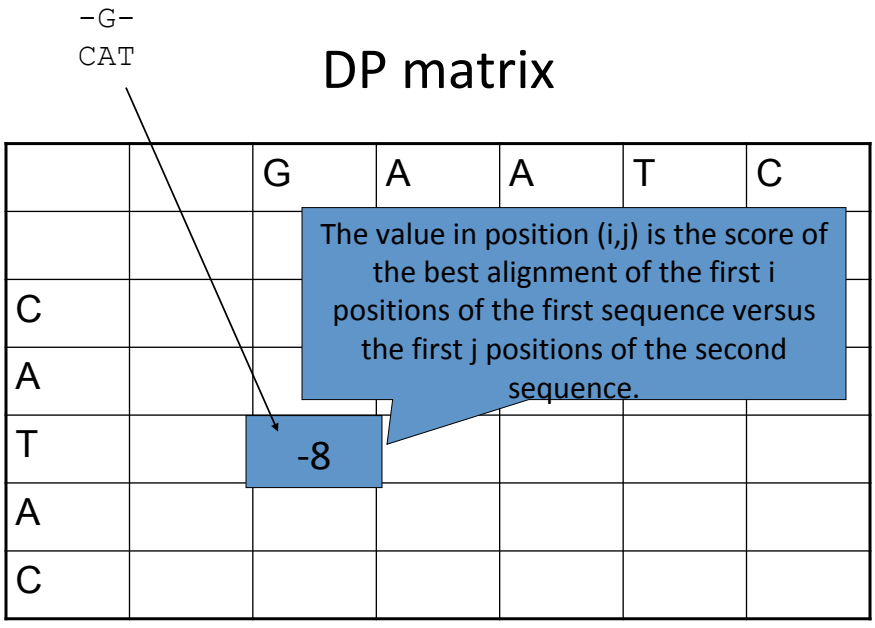
$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) + d \\ F(i,j-1) + d \end{cases}$$

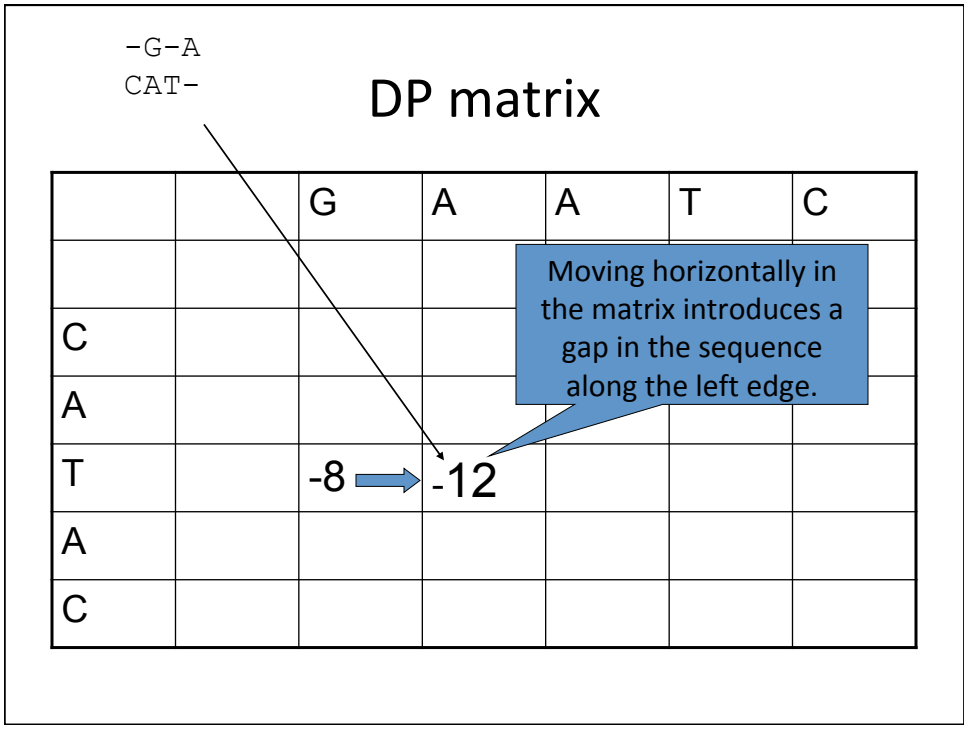
Lets first consider the simple scoring scheme …

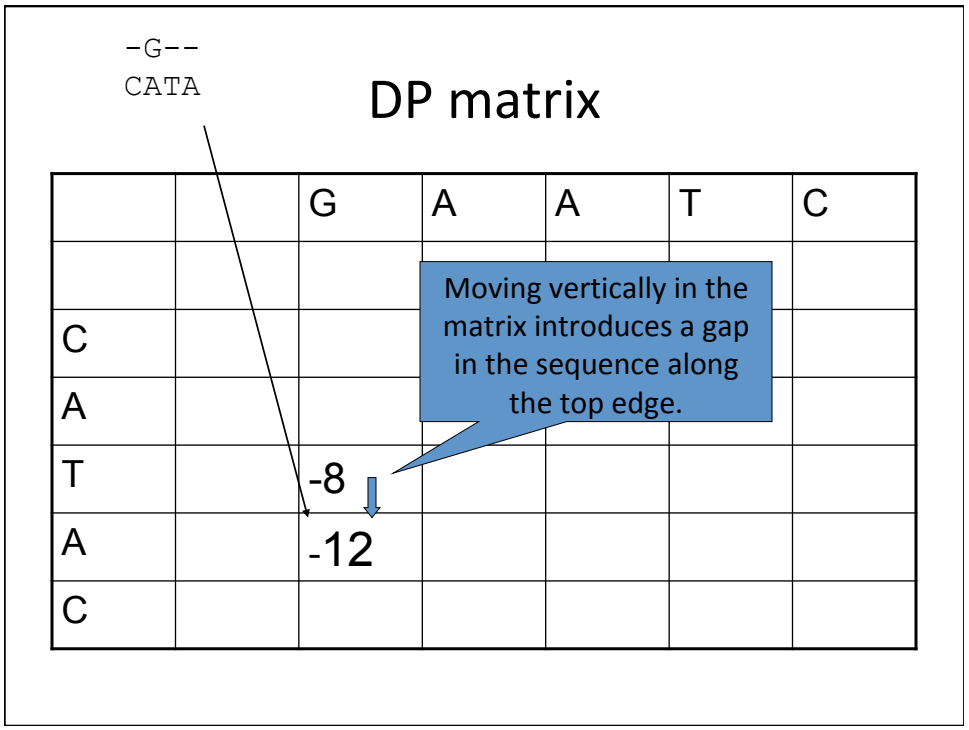
- Problem: find the best pairwise alignment of `GAATC` and `CATAC`.
- Use a linear gap penalty of -4.
- Use the following substitution matrix:

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

DP matrix




-G-A
CAT-
DP matrix
G
A
A
T
C
C
A
T
A
C
-8
-12
Moving horizontally in the matrix introduces a gap in the sequence along the left edge.


-G--
CATA
DP matrix
G
A
A
T
C
C
A
T
A
C
-8
-12
Moving vertically in the matrix introduces a gap in the sequence along the top edge.

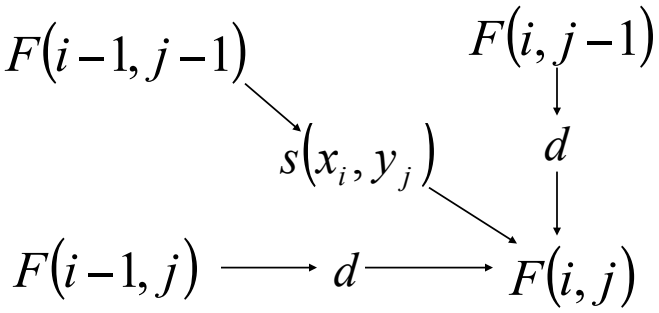
DP matrix

-G-A
CATA

		G	A	A	T	C
C						
A						
T		-8				
A			2			
C						

Moving diagonally in the matrix matches both characters in the two sequences.

DP in equation form



Initialization

		G	A	A	T	C
	0					
C						
A						
T						
A						
C						

Introducing a gap

G
–

		G	A	A	T	C
	0 →	-4				
C						
A						
T						
A						
C						

DP matrix

		G	A	A	T	C
	0 →	-4				
C	-4					
A						
T						
A						
C						

DP matrix

		G	A	A	T	C
	0 →	-4				
C	-4 →	-8				
A						
T						
A						
C						

G
C

DP matrix

		G	A	A	T	C
	0 →	-4				
C	-4	-5				
A						
T						
A						
C						

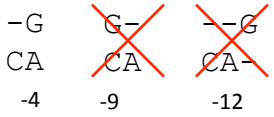
\-----
CATAC

DP matrix

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

DP matrix

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



DP matrix

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

0

DP matrix

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

DP matrix

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

DP matrix

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

DP matrix

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4 ↓	-5 →	-9			
A	-8 ↓	-4 ↘	5 ↓			
T	-12 ↓	-8 ↓	1			
A	-16 ↓	-12 ↓	2			
C	-20	-16	-2			

What is the alignment associated with this entry?

DP matrix

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4 ↓	-5 →	-9			
A	-8 ↓	-4 ↓	5 ↓			
T	-12 ↓	-8 ↓	1 ↓			
A	-16 ↓	-12 ↓	**2**			
C	-20	-16	-2			

-G-A
CATA

DP matrix

		G	A	A	T	C
	0 →	-4 →	-8 →	-12 →	-16 →	-20
C	-4 ↓	-5 →	-9			
A	-8 ↓	-4 ↓	5 ↓			
T	-12 ↓	-8 ↓	1			
A	-16 ↓	-12 ↓	2 ↓			
C	-20	-16	-2			**?**

Find the optimal alignment, and its score.

DP matrix

		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**

Traceback

- Start from the lower right corner and trace back to the upper left.
- Each arrow introduces one character at the end of each aligned sequence.
- A <u>horizontal</u> move puts a gap in the <u>left</u> sequence.
- A <u>vertical</u> move puts a gap in the <u>top</u> sequence.
- A diagonal move uses one character from each sequence.

GA-ATC
CATA-C

DP matrix

		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**

GAAT-C
CA-TAC

DP matrix

		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**

GAAT-C
C-ATAC

DP matrix

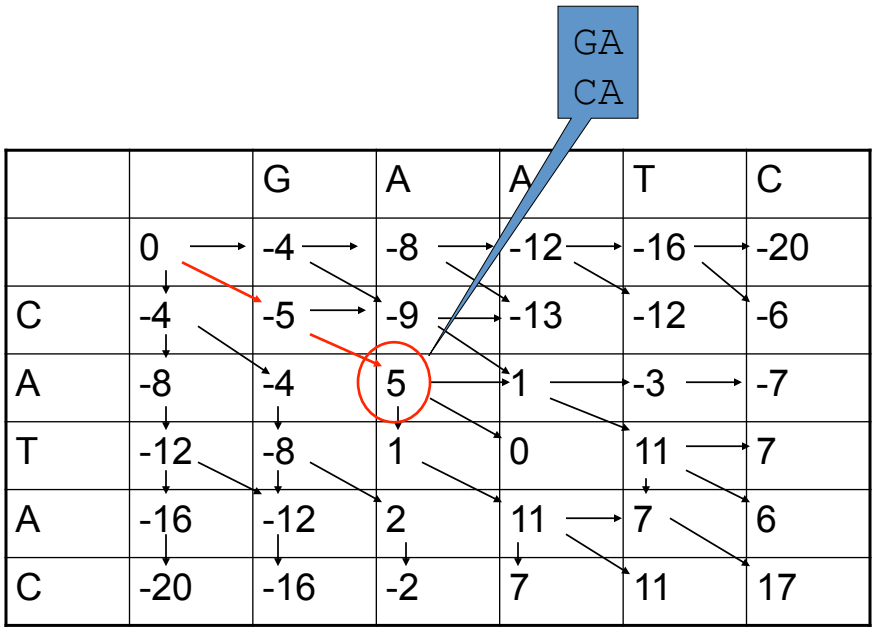
		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	**17**

GAAT-C
-CATAC

DP matrix

		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**

Write down the alignment corresponding to the circled score.



Multiple solutions

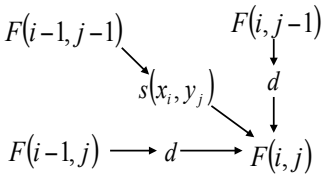
```
GA-ATC
CATA-C

GAAT-C
CA-TAC

GAAT-C
C-ATAC

GAAT-C
-CATAC
```

- When a program returns a sequence alignment, it may not be the **only** best alignment.

Needleman-Wunsch

Scoring Scheme:

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



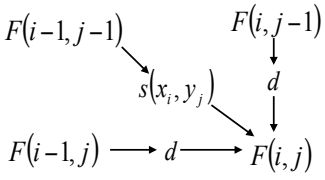
Find the optimal alignment of AAG and AGC.
Use a gap penalty of d=-5.

		A	A	G
A				
G				
C				

Needleman-Wunsch

Scoring Scheme:

	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 alignment of AAG and AGC.
Use a gap penalty of d=-5.

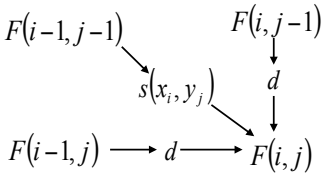
		A	A	G
	0			
A				
G				
C				

Needleman-Wunsch

	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 alignment of AAG and AGC.
Use a gap penalty of d=-5.

		A	A	G
	0 →	-5 →	-10 →	-15
A	-5			
G	-10			
C	-15			

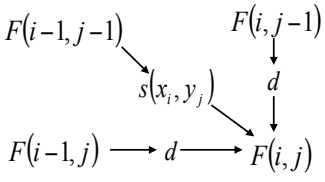


Needleman-Wunsch

	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 alignment of AAG and AGC.
Use a gap penalty of d=-5.

		A	A	G
	0 →	-5 →	-10 →	-15
A	-5	2 →	-3 →	-8
G	-10	-3	-3	-1
C	-15	-8	-8	**-6**



Needleman-Wunsch

Find the optimal alignment of AAG and AGC.
Use a gap penalty of d=-5.

- Start from the lower right corner and trace back to the upper left.
- Each arrow introduces one character at the end of each aligned sequence.
- A horizontal move puts a gap in the left sequence.
- A vertical move puts a gap in the top sequence.
- A diagonal move uses one character from each sequence.

		A	A	G
	0 →	-5		
A		2 →	-3	
G				-1
C				**-6**

Needleman-Wunsch

Find the optimal alignment of AAG and AGC.
Use a gap penalty of d=-5.

- Start from the lower right corner and trace back to the upper left.
- Each arrow introduces one character at the end of each aligned sequence.
- A horizontal move puts a gap in the left sequence.
- A vertical move puts a gap in the top sequence.
- A diagonal move uses one character from each sequence.

		A	A	G
	0 →	-5		
A		2 →	-3	
G				-1
C				**-6**

```
AAG-    AAG-
-AGC    A-GC
```