Sequence Comparison: Local Alignment

Genome 373

Genomic Informatics

Elhanan Borenstein

A quick review: Global Alignment

Global Alignment Mission:

Find the best global alignment between two sequences.

An algorithm for finding the alignment with the best score

A method for scoring alignments

A quick review: Global Alignment

Global Alignment Mission:

Find the best global alignment between two sequences.

An algorithm for finding the alignment with the best score

A method for scoring alignments

?

- **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

Purine	A	G
Pyrimidine	C	T

- **Gap penalty:**
  - Linear gap penalty
  - Affine gap penalty

GAAT-C
CA-TAC

d=-4

-5 + 10 + -4 + 10 + -4 + 10 = 17

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					

0 -4 -4 0

Review: Global Alignment

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

Gap penalty: $d = -4$

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

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

- 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					

$$F(i-1, j-1) \rightarrow s(x_i, y_j)$$

$$F(i, j-1) \rightarrow d$$

$$s(x_i, y_j) \rightarrow F(i, j)$$

$$d \rightarrow F(i, j)$$

$$F(i-1, j) \rightarrow d \rightarrow F(i, j)$$

take the max
of these three

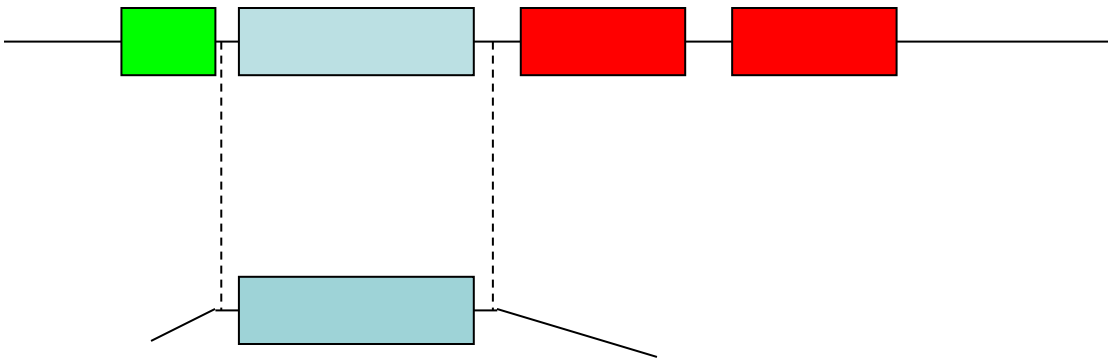
Local alignment

<u>Mission:</u>

Find best <u>partial</u> alignment between two sequences.

Why?

Local alignment



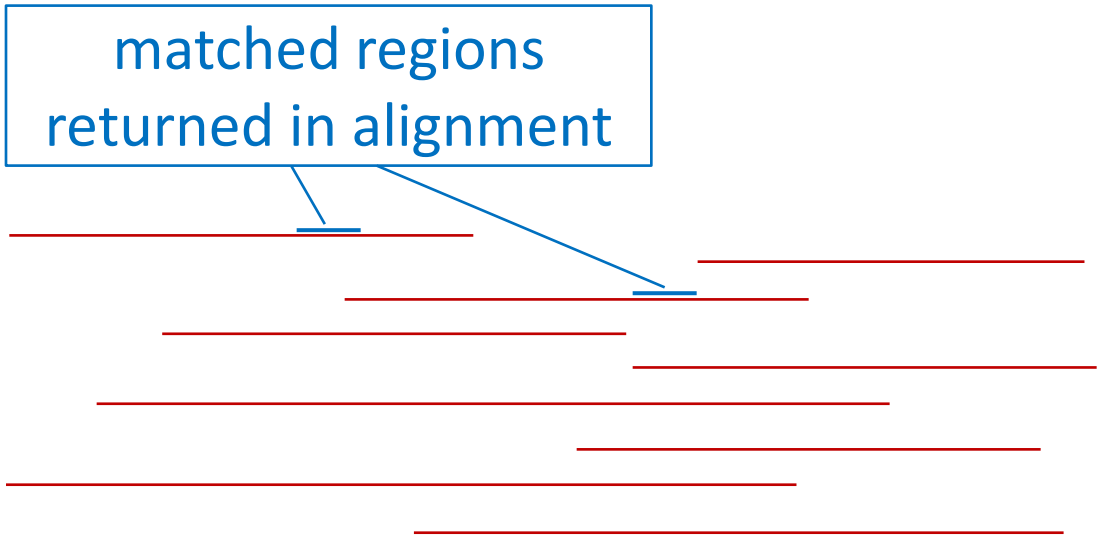
- A single-**domain** protein may be similar only to one region within a multi-domain protein.
- A DNA/RNA **query** may align to a small part of a genome/genomes/metagenomes.
- An alignment that spans the complete length of both sequences may be undesirable.

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.

Query:

Targets: (e.g. genome contigs, full genomes, metagenomes)



How can we modify the *Needleman-Wunsch* DP algorithm (for finding global alignment) such that it will find instead the best **local** alignment??

		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

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

Gap penalty: d=−4

```
G  A  A  T  -  C
-  C  A  T  A  C
-4 -5 10 10 -4 10 = 17
```

		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

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

Gap penalty: d=−4

```
G  A  A  T  -  C
-  C  A  T  A  C
-4 -5 10 10 -4 10 = 17
→-4→-9→ 1→11→ 7→17
```

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)

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				

$$
\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1, j-1) \arrow[r] & s(x_i, y_j) \arrow[dr] & F(i, j-1) \arrow[d] & \\
 & & d \arrow[d] & \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j) &
\end{tikzcd}
$$

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	?			

$$
\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] & \\
 & s(x_i, y_j) \arrow[dr] & d \arrow[d] & \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j) &
\end{tikzcd}
$$

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			

$$
\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] & \\
 & s(x_i, y_j) \arrow[dr] & d \arrow[d] & \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j) &
\end{tikzcd}
$$

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			

$$
\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] & \\
 & s(x_i, y_j) \arrow[dr] & d \arrow[d] & \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j) &
\end{tikzcd}
$$

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	
		-5	0	
G	0			
C	0			

$$\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1, j-1) \arrow[r] & s(x_i, y_j) \arrow[dr] & F(i, j-1) \arrow[d] & \\
 & & d \arrow[d] & \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j) &
\end{tikzcd}$$

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			

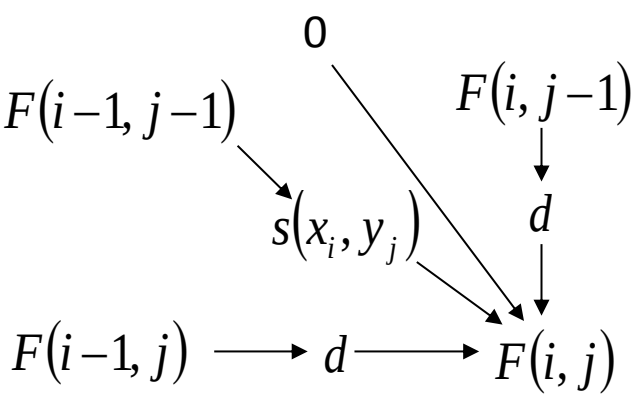
$$
\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] & \\
 & s(x_i, y_j) \arrow[dr] & d \arrow[d] & \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j) &
\end{tikzcd}
$$

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

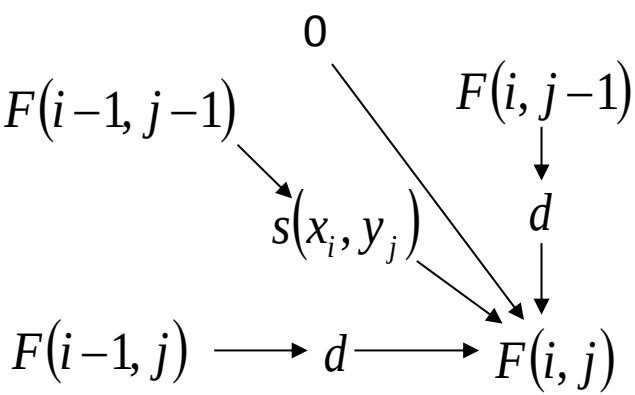
		A	A	G
	0	0	0	0
A	0	2		
G	0	?		
C	0	?		



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

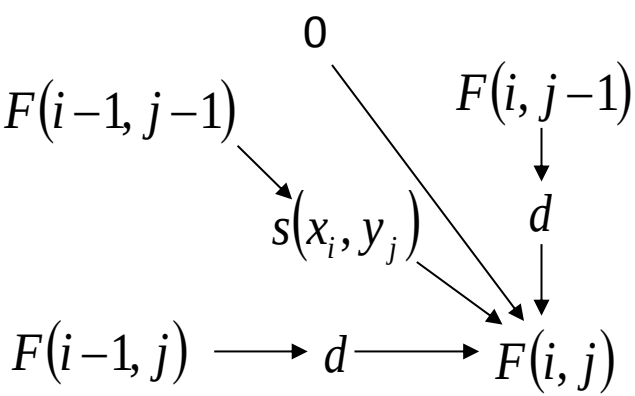


		A	A	G
	0	0	0	0
A	0	2		
G	0	-5 -3 -5 0		
C	0	?		

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	?	

$$\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1,j-1) \arrow[dr] & & F(i,j-1) \arrow[d] & \\
 & s(x_i,y_j) \arrow[dr] & d \arrow[d] & \\
F(i-1,j) \arrow[r] & d \arrow[r] & F(i,j) &
\end{tikzcd}$$

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	

$$\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] & \\
 & s(x_i, y_j) \arrow[dr] & d \arrow[d] & \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j) &
\end{tikzcd}$$

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	?

$$\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] & \\
 & s(x_i, y_j) \arrow[dr] & d \arrow[d] & \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j) &
\end{tikzcd}$$

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

$$
\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1, j-1) \arrow[r] & s(x_i, y_j) \arrow[dr] & F(i, j-1) \arrow[d] & \\
 & & d \arrow[d] & \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j) &
\end{tikzcd}
$$

What's different about the DP matrix

Global Alignment 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

Local Alignment DP Matrix

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

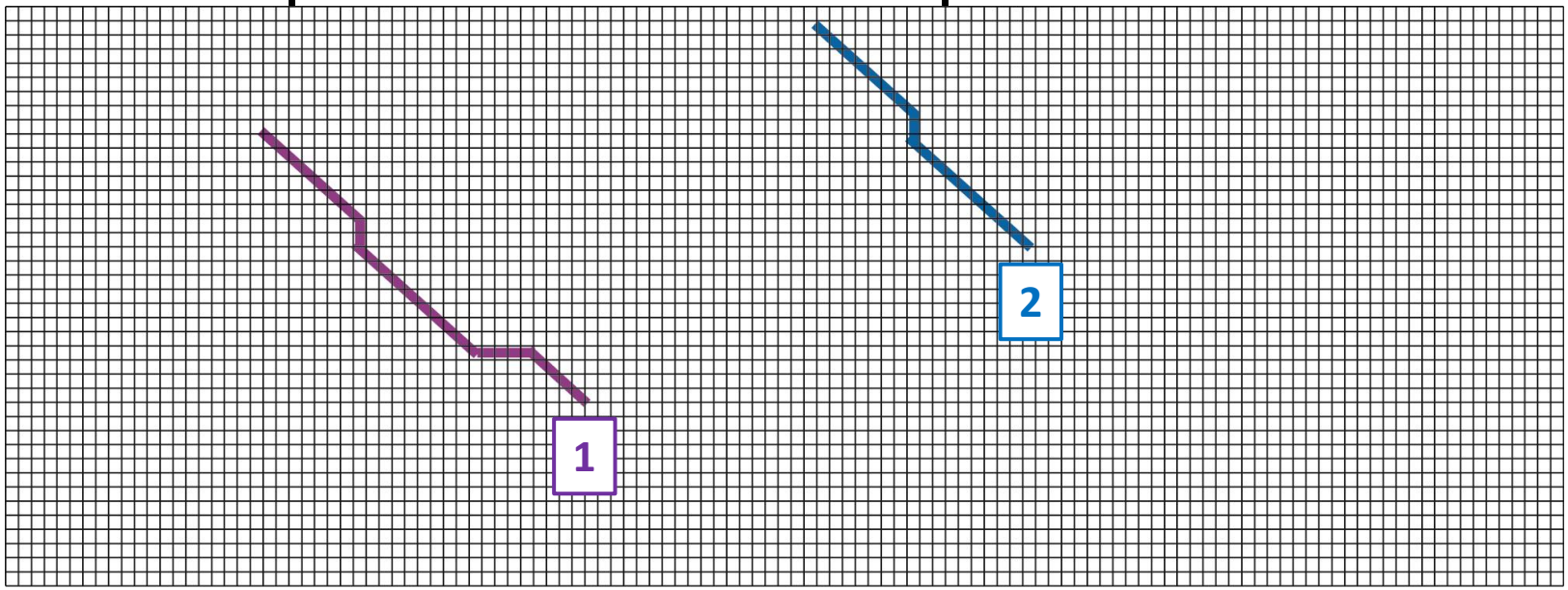
		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0

$$\begin{tikzcd}
 & 0 \arrow[dr] & & \\
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] & \\
 & s(x_i, y_j) \arrow[dr] & d \arrow[d] & \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j) &
\end{tikzcd}$$

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

Multiple local alignments

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



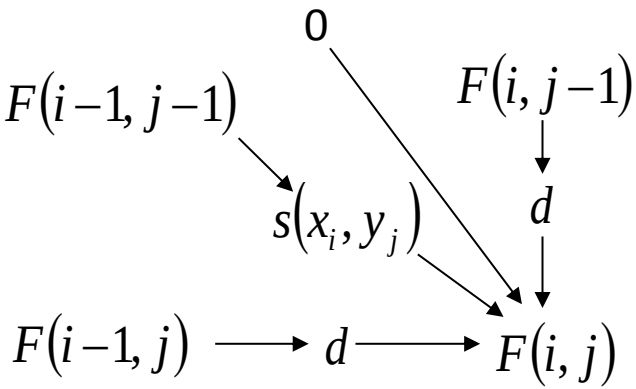
Local alignment

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

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

d = -5



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

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

d = -5

$$\begin{tikzcd}
F(i-1,j-1) \arrow[dr] & & F(i,j-1) \arrow[d] \\
& s(x_i,y_j) \arrow[dr] & d \arrow[d] \\
F(i-1,j) \arrow[r] & d \arrow[r] & F(i,j)
\end{tikzcd}$$

(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			

Summary

Global alignment algorithm:

Needleman-Wunsch.

		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**

Local alignment algorithm:

Smith-Waterman.

		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

Using sequence alignment to study evolution

Are these proteins related?

The intuitive answer:

```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
           L P      L   Y N    Y C     L
SEQ 2: QFFPLMPPAPYFILATDYENLPLVYSCTTFFWLF
```

score = -1 → NO?

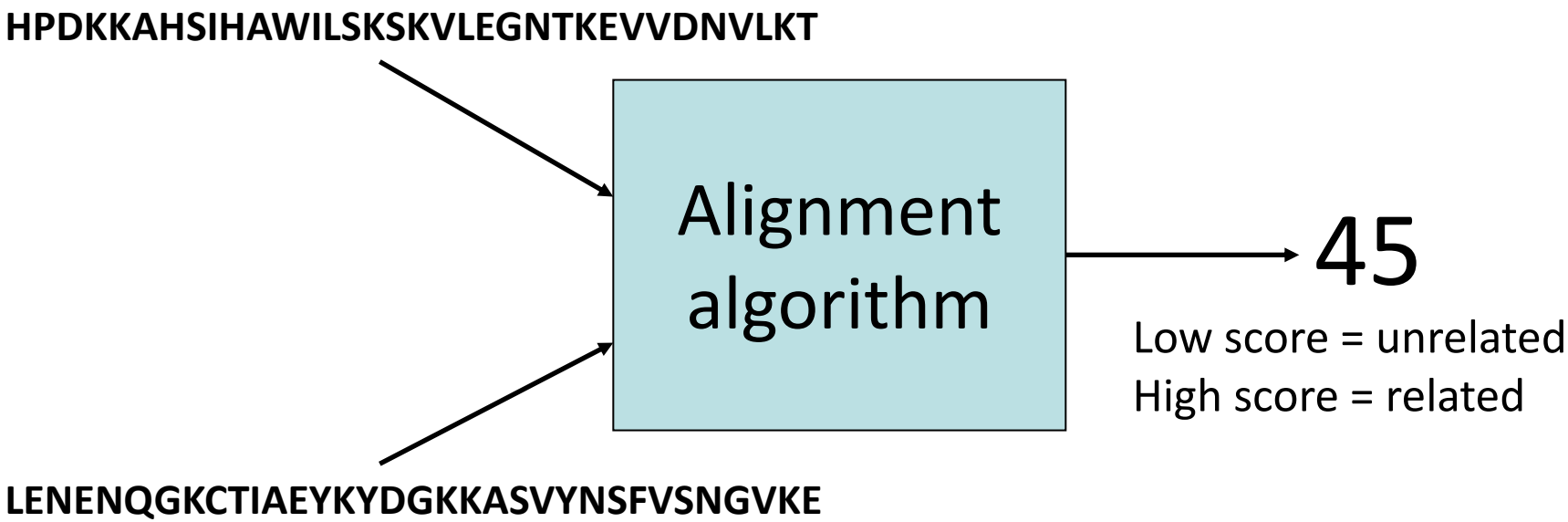
```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
           L P    W LDATYKN A  Y C     L
SEQ 2: QFFPLMPPAPYWILDATYKNLALVYSCTTFFWLF
```

score = 15 → PROBABLY?

```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
       RVV L PS   W LDATYKNYA  Y CDVTYKL
SEQ 2: RVVPLMPSAPYWILDATYKNYALVYSCDVTYKLF
```

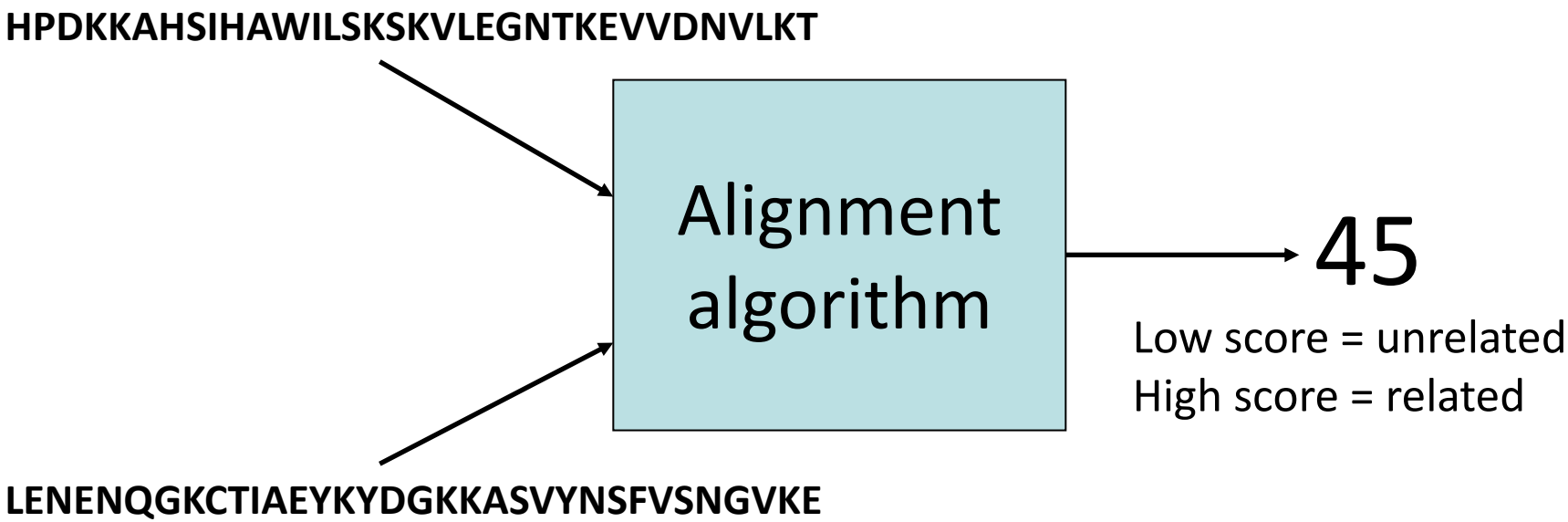
score = 37 → YES?

Significance of scores



But … how high is high enough?

Significance of scores



But … how high is high enough?

Subjective

Problem specific

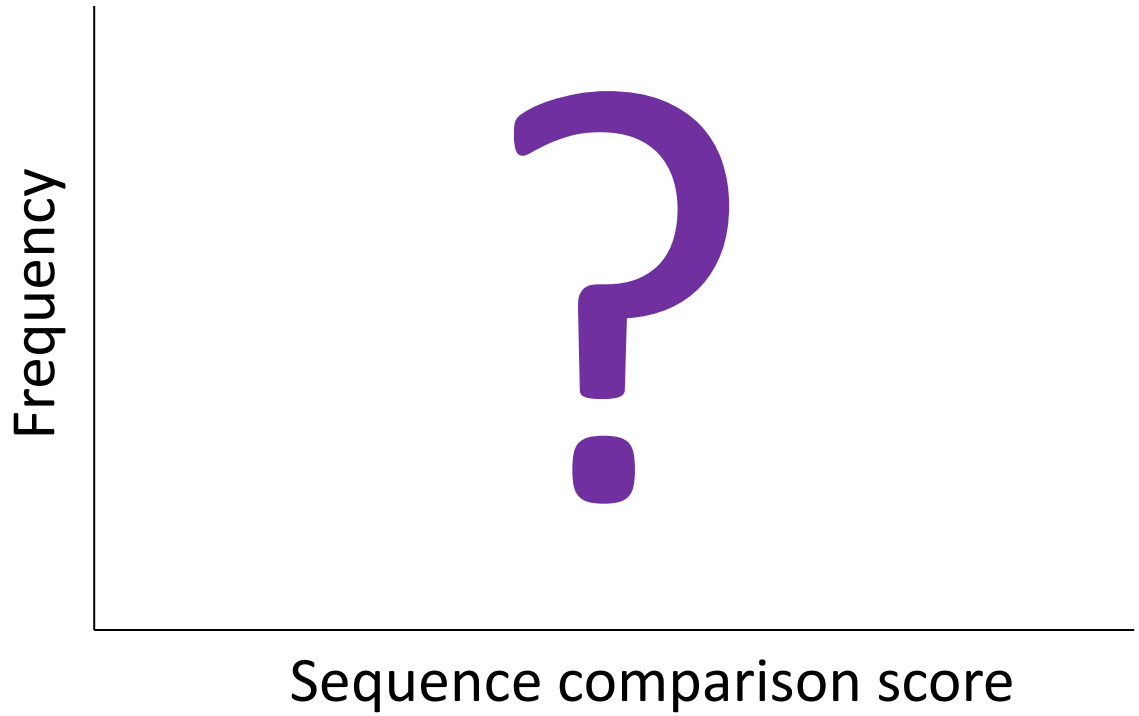
Parameter specific

The null hypothesis

- We want to know how **surprising** a given score is, …
  assuming that the two sequences are <u>not</u> related.
- This assumption is called the null hypothesis.
- The purpose of most statistical tests is to determine whether the observed result provides a reason to reject the null hypothesis.

- We want to characterize the distribution of scores from pairwise sequence alignments.

Sequence similarity score distribution



- Search a database of **unrelated** sequences using a given query sequence.
- What will be the form of the resulting distribution of pairwise alignment scores?