Sequence Comparison: Dynamic Programming

Genome 373

Genomic Informatics

Elhanan Borenstein

A quick review: Challenges

- Find the best *global* alignment of two sequences
- Find the best *global* alignment of multiple sequences
- Find the best *local (partial)* alignment of two sequences
- Find the best match (alignment) of a given sequence in a longer dataset of sequences

A quick review: Global Alignment

<u>Global Alignment Mission:</u>

Find the best global alignment between two sequences.

e.g., Find the best alignment of GAATC and CATAC:

GAATC CATAC	GAAT-C C-ATAC	-GAAT-C C-A-TAC	GAAT-C C-ATAC
GAATC- CA-TAC	GAAT-C CA-TAC	GA-ATC CATA-C	GAAT-C CA-TAC

- "Correct" alignment vs. "best" alignment

A quick review: Global Alignment

<u>Global Alignment Mission:</u>

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: Scoring aligned bases

- **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

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

Exhaustive search

- *Align the two sequences: GAATC and CATAC*

GAATC CATAC	GAAT-C C-ATAC	-GAAT-C C-A-TAC	GAAT-C C-ATAC
GAATC- CA-TAC	GAAT-C CA-TAC	GA-ATC CATA-C	GAAT-C CA-TAC

<u>Simple (exhaustive search) algorithm</u>

1) *Construct all possible alignments*
2) *Use the substitution matrix and gap penalty to score each alignment*
3) *Pick the alignment with the best score*

Exhaustive search

- *Align the two sequences: GAATC and CATAC*

GAATC CATAC	GAAT-C C-ATAC	-GAAT-C C-A-TAC	GAAT-C C-ATAC
GAATC- CA-TAC	GAAT-C CA-TAC	GA-ATC CATA-C	GAAT-C CA-TAC

<u>Simple (exhaustive search) algorithm</u>

1) *Construct all possible alignments*
2) *Use the substitution matrix and gap penalty to score each alignment*
3) *Pick the alignment with the best score*

Computational Complexity & the Big *O* Notation

Mission:

Find the best global alignment of two sequences.

A "search" algorithm for finding the alignment with the best score

A method for scoring alignments

?

1. More efficient search
2. A recipe

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

```
GAAT-C
CA-TAC
```

d=−4

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

Mission:

Find the best global alignment of two sequences.

A "search" algorithm for finding the alignment with the best score

A method for scoring alignments

?

The Needleman–Wunsch Algorithm

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

```
GAAT-C
CA-TAC
```

d=−4

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

The Needleman–Wunsch Algorithm

- An algorithm for **global alignment** on two sequences
- A **Dynamic Programming (DP)** approach
  - 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.

DP matrix

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

j ⇨ 0 1 2 3 etc.

i ⇩

		G	A	A	T	C
C						
A						
T						
A						
C						

0
1
2
3
4
5

DP matrix

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

i \ j		0	1	2	3 etc.			
			G	A	A	T	C	
0								
1	C							
2	A							
3	T							
4	A							
5	C							

initial row and column

DP matrix

Best alignment of GA to CA

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

i \ j		0	1	2	3 etc.		
			G	A	A	T	C
0							
1	C						
2	A			5			
3	T						
4	A						
5	C						

Which value are we interested in?

The value at (**i**, **j**) is the <u>score of the best alignment</u> of the first **i** characters of one sequence versus the first **j** characters of the other sequence.

DP matrix

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

i ⇩ \ j ⇨		0	1	2	3 etc.		
			G	A	A	T	C
0							
1	C						
2	A			5			
3	T						
4	A						
5	C						

The score of the <u>best alignment</u> of the two sequences.

Moving in the DP matrix

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			5			
T						
A						
C						

DP matrix

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			5			
T						
A						
C						

DP matrix

Best alignment of GA and CA

A	
–	

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			5			
T						
A						
C						

Moving horizontally in the matrix introduces a gap in the sequence along the left edge.

DP matrix

Best alignment of GA and CA

A	
–	

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			5	1		
T						
A						
C						

Moving horizontally in the matrix introduces a gap in the sequence along the left edge.

DP matrix

Best alignment of GA and CA

–	
T	

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			5			
T			1			
A						
C						

Moving vertically in the matrix introduces a gap in the sequence along the top edge.

DP matrix

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			5			
T						
A						
C						

Best alignment of GA and CA

A
T

DP matrix

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			5			
T				0		
A						
C						

Moving diagonally in the matrix aligns two residues

So …. Can I now fill this matrix?

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			7			
A						
C						

So …. Can I now fill this matrix?

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			7	3		
A			3	17		
C						

Initialization

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						

Initialization

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
	0					
C						
A						
T						
A						
C						

Introducing a gap

G
–

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
	0	-4				
C						
A						
T						
A						
C						

Introducing a gap

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
	0	-4				
C	-4					
A						
T						
A						
C						

Complete first row and column

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

CATAC

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

What about **i=1, j=1**

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

i \ j		0	1	2	3 etc.		
			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	?				
2	A	-8					
3	T	-12					
4	A	-16					
5	C	-20					

Three ways to get to **i=1, j=1**

G-
-C

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

i \ j		0	1	2	3 etc.		
			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	-8				
2	A	-8					
3	T	-12					
4	A	-16					
5	C	-20					

Three ways to get to **i=1, j=1**

-G
C-

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

i \ j		0	1	2	3 etc.		
			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	-8				
2	A	-8					
3	T	-12					
4	A	-16					
5	C	-20					

G
C

Three ways to get to **i=1, j=1**

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

j ⇨ 0 1 2 3 etc.

i ⇩

			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	-5				
2	A	-8					
3	T	-12					
4	A	-16					
5	C	-20					

Accept the highest scoring of the three

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

Then simply repeat the same rule progressively across the matrix

DP matrix

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

~~G–~~ ~~CA~~ -5+-4=-9

–G CA -4+0=-4

~~– –G~~ ~~CA–~~ -8+-4=-12

DP matrix

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

0 ~~-4~~ ~~-4~~

~~G–~~
~~CA~~

-5+-4=-9

–G
CA

-4+0=-4

~~– –G~~
~~CA–~~

-8+-4=-12

DP matrix

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

0 ~~-4~~

-4 ~~-4~~

DP matrix

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

DP matrix

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

DP matrix

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

Traceback

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
	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?

Traceback

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
	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?

Just follow the arrows back - this is called the **traceback**

-G-A
CATA

Full 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

		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			?

Continue and find the optimal global alignment, and its score.

Full 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: $d=-4$

		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**

Full 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

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4				2	-6
A	-8					-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

Best alignment starts at bottom right and follows traceback arrows to top left

GA–ATC
CATA–C

One best traceback

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
	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

Another best traceback

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
	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

GA−ATC
CATA−C

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
	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

Multiple solutions

```
GA-ATC
CATA-C

GAAT-C
CA-TAC

GAAT-C
C-ATAC

GAAT-C
-CATAC
```

- When a program returns a single sequence alignment, it may not be the **only** best alignment but it is guaranteed to be one of them.

- In our example, all of the alignments at the left have equal scores.

What’s the complexity of this algorithm?

Practice problem:

Find a best pairwise alignment of GAATC and AATTC

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
	0					
A						
A						
T						
T						
C						

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					

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

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

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

take the max of these three