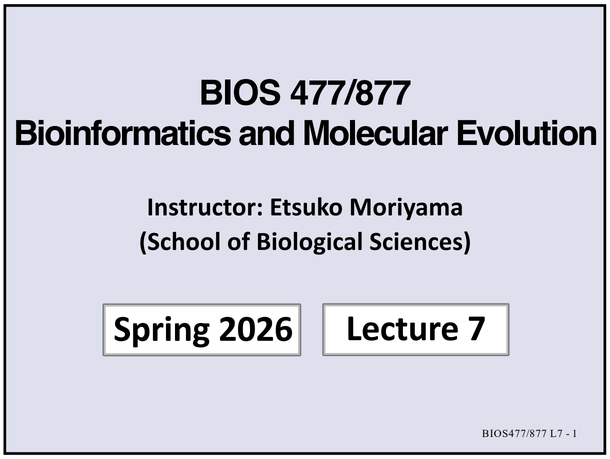
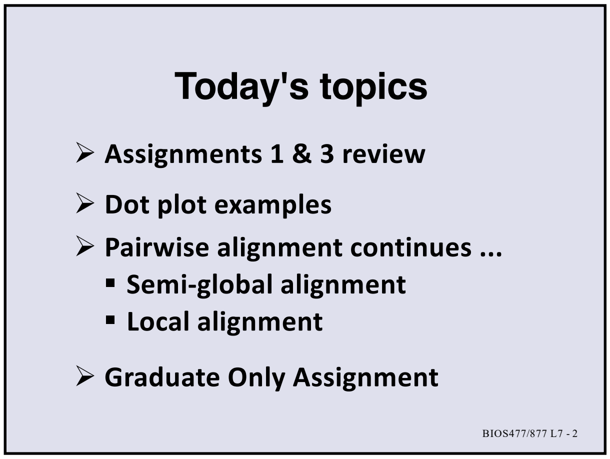


BIOS 477/877
Bioinformatics and Molecular Evolution

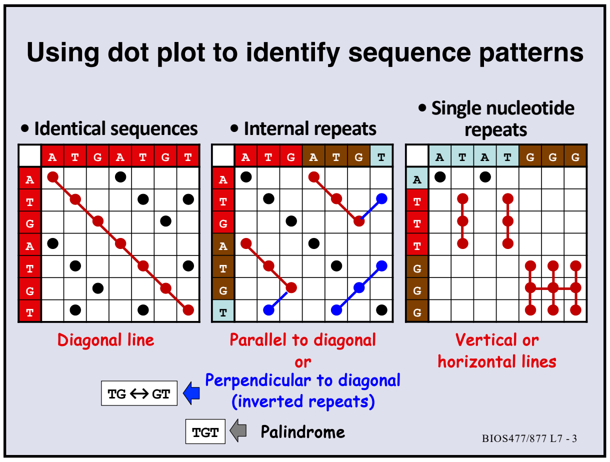
Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 **Lecture 7**



Today's topics

- Assignments 1 & 3 review
- Dot plot examples
- Pairwise alignment continues ...
  - Semi-global alignment
  - Local alignment
- Graduate Only Assignment



Using dot plot to identify sequence patterns

- Identical sequences
- Internal repeats
- Single nucleotide repeats

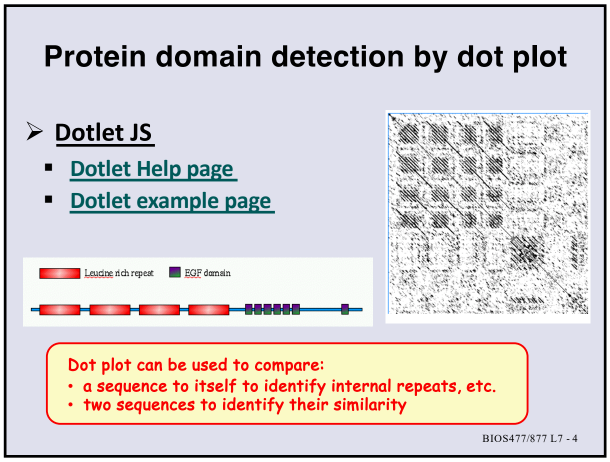
Diagonal line

Parallel to diagonal
or
Perpendicular to diagonal
(inverted repeats)

TG ↔ GT

TGT Palindrome

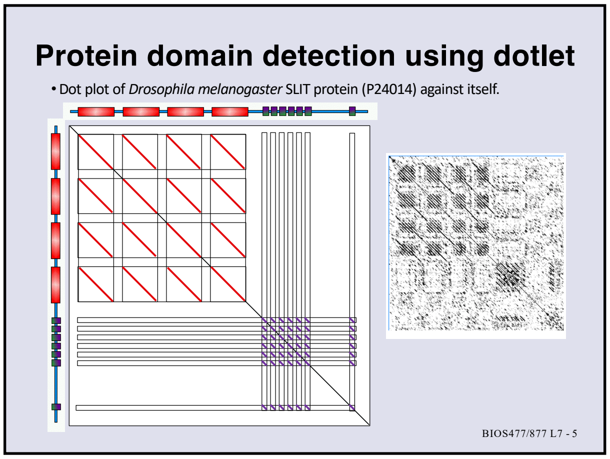
Vertical or
horizontal lines



Protein domain detection by dot plot

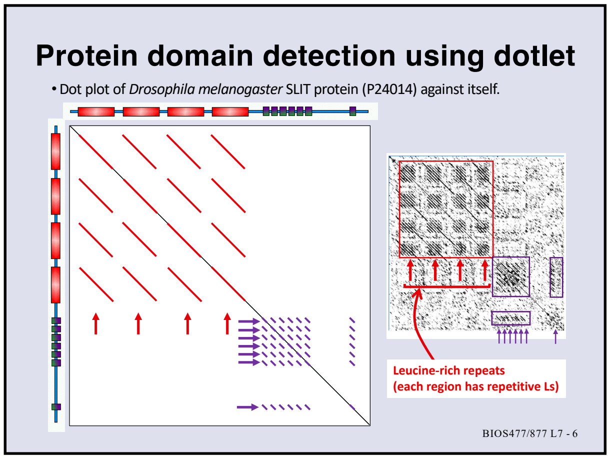
- Dotlet JS
  - Dotlet Help page
  - Dotlet example page

Dot plot can be used to compare:
- a sequence to itself to identify internal repeats, etc.
- two sequences to identify their similarity



Protein domain detection using dotlet

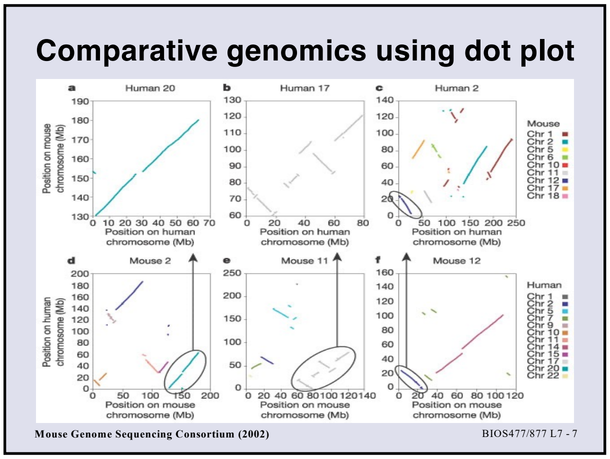
- Dot plot of *Drosophila melanogaster* SLIT protein (P24014) against itself.



Protein domain detection using dotlet

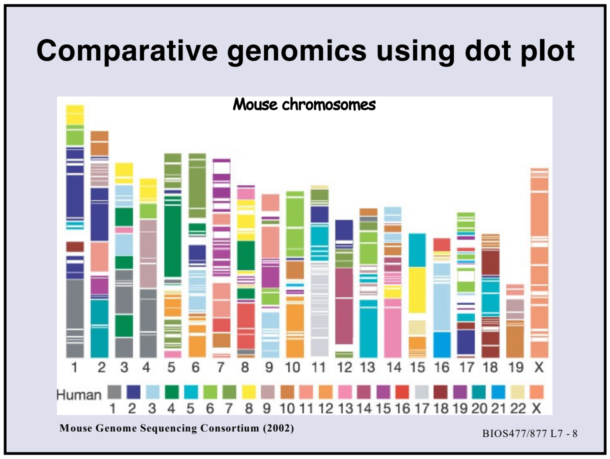
- Dot plot of *Drosophila melanogaster* SLIT protein (P24014) against itself.

Leucine-rich repeats
(each region has repetitive Ls)



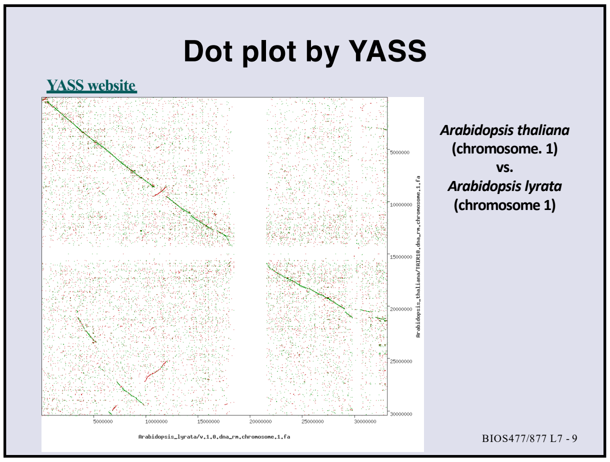
Comparative genomics using dot plot

Mouse Genome Sequencing Consortium (2002)



Comparative genomics using dot plot

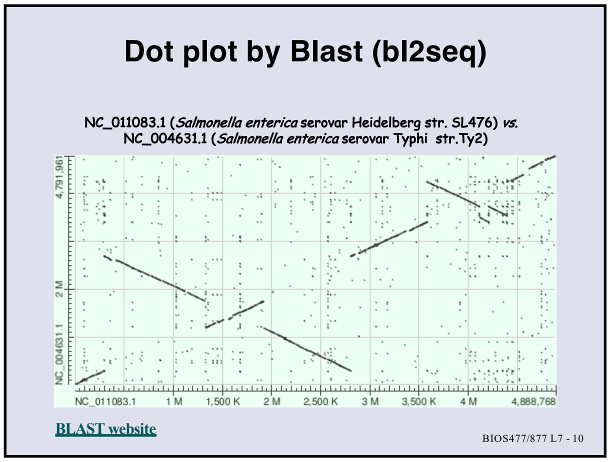
Mouse Genome Sequencing Consortium (2002)



Dot plot by YASS

YASS website

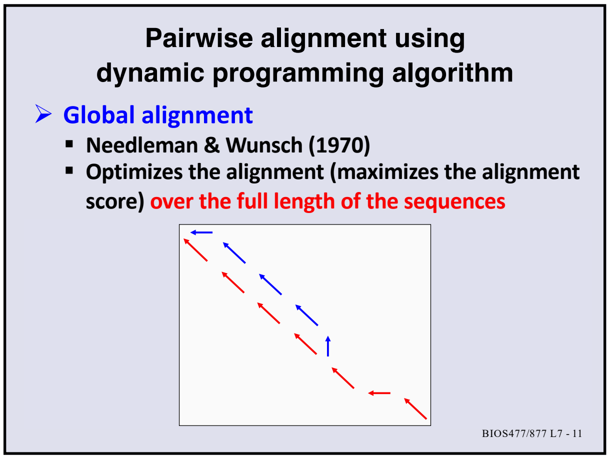
Arabidopsis thaliana (chromosome. 1) vs. *Arabidopsis lyrata* (chromosome 1)



Dot plot by Blast (bl2seq)

NC_011083.1 (*Salmonella enterica* serovar Heidelberg str. SL476) *vs.* NC_004631.1 (*Salmonella enterica* serovar Typhi str.Ty2)

BLAST website

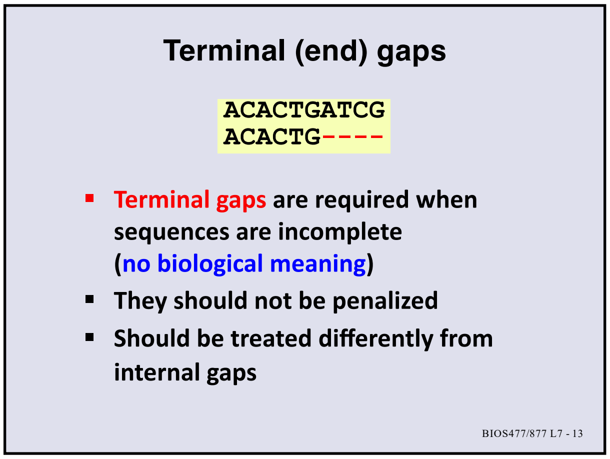


Pairwise alignment using dynamic programming algorithm

- **Global alignment**
  - Needleman & Wunsch (1970)
  - Optimizes the alignment (maximizes the alignment score) over the full length of the sequences

Pairwise alignment using dynamic programming algorithm

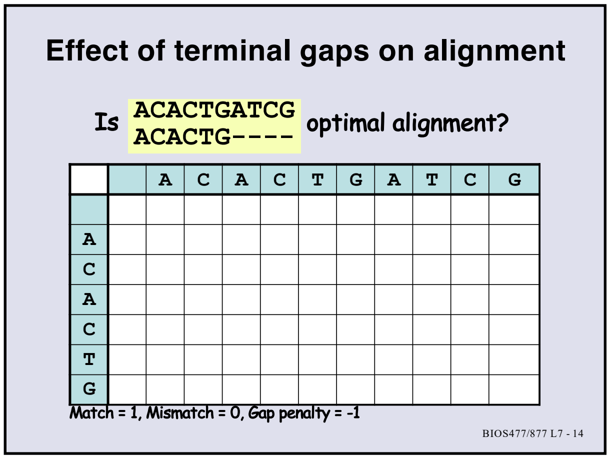
- **Global alignment**
  - Needleman & Wunsch (1970)
  - Optimizes the alignment (maximizes the alignment score) over the full length of the sequences
- **Semi-global alignment**
  - Global alignment with cost-free end gaps
- **Local alignment**
  - Smith & Waterman (1981)
  - Finds the best matching subsequences



Terminal (end) gaps

```
ACACTGATCG
ACACTG----
```

- Terminal gaps are required when sequences are incomplete (no biological meaning)
- They should not be penalized
- Should be treated differently from internal gaps

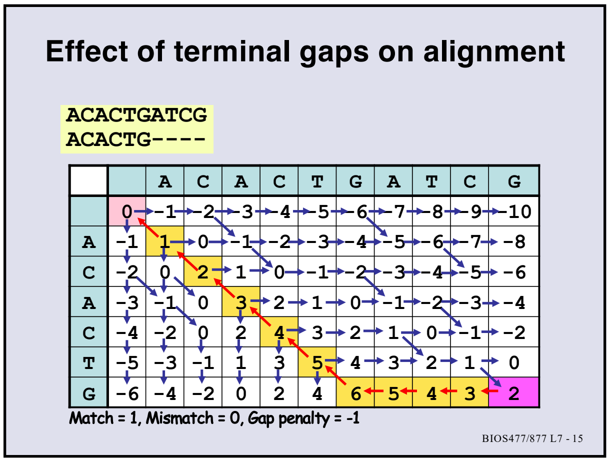


Effect of terminal gaps on alignment

Is
```
ACACTGATCG
ACACTG----
```
optimal alignment?

		A	C	A	C	T	G	A	T	C	G
A											
C											
A											
C											
T											
G											

Match = 1, Mismatch = 0, Gap penalty = -1

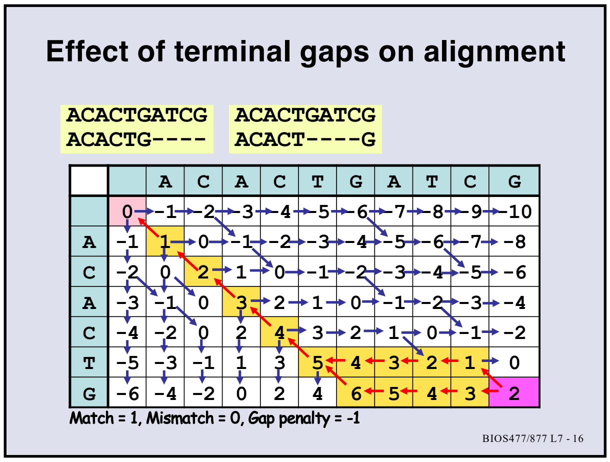


Effect of terminal gaps on alignment

```
ACACTGATCG
ACACTG----
```

		A	C	A	C	T	G	A	T	C	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8
C	-2	0	2	1	0	-1	-2	-3	-4	-5	-6
A	-3	-1	0	3	2	1	0	-1	-2	-3	-4
C	-4	-2	0	2	4	3	2	1	0	-1	-2
T	-5	-3	-1	1	3	5	4	3	2	1	0
G	-6	-4	-2	0	2	4	6	5	4	3	2

Match = 1, Mismatch = 0, Gap penalty = -1

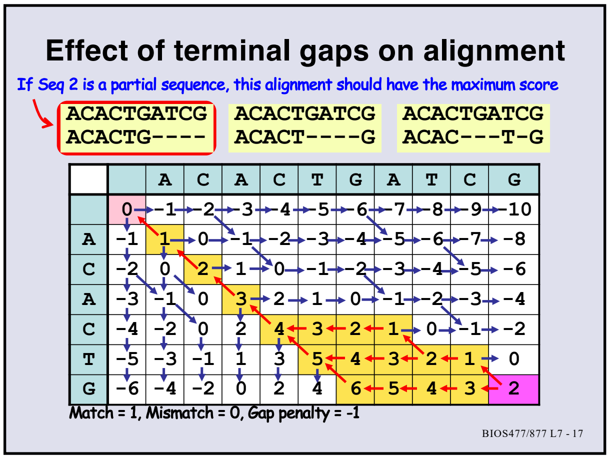


Effect of terminal gaps on alignment

```
ACACTGATCG   ACACTGATCG
ACACTG----   ACACT----G
```

		A	C	A	C	T	G	A	T	C	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8
C	-2	0	2	1	0	-1	-2	-3	-4	-5	-6
A	-3	-1	0	3	2	1	0	-1	-2	-3	-4
C	-4	-2	0	2	4	3	2	1	0	-1	-2
T	-5	-3	-1	1	3	5	4	3	2	1	0
G	-6	-4	-2	0	2	4	6	5	4	3	2

Match = 1, Mismatch = 0, Gap penalty = -1



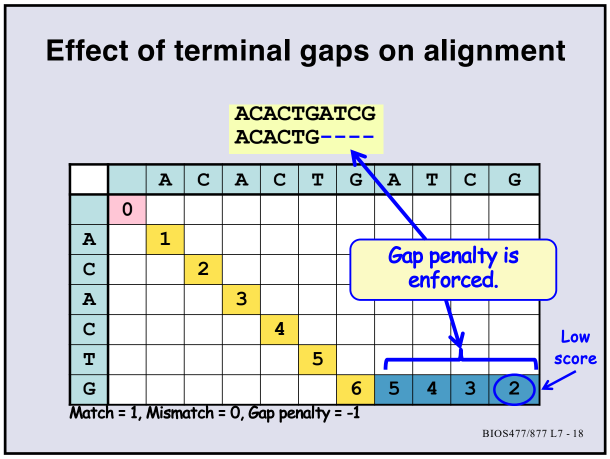
Effect of terminal gaps on alignment

If Seq 2 is a partial sequence, this alignment should have the maximum score

```
ACACTGATCG   ACACTGATCG   ACACTGATCG
ACACTG----   ACACT----G   ACAC---T-G
```

		A	C	A	C	T	G	A	T	C	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8
C	-2	0	2	1	0	-1	-2	-3	-4	-5	-6
A	-3	-1	0	3	2	1	0	-1	-2	-3	-4
C	-4	-2	0	2	4	3	2	1	0	-1	-2
T	-5	-3	-1	1	3	5	4	3	2	1	0
G	-6	-4	-2	0	2	4	6	5	4	3	2

Match = 1, Mismatch = 0, Gap penalty = -1



Effect of terminal gaps on alignment

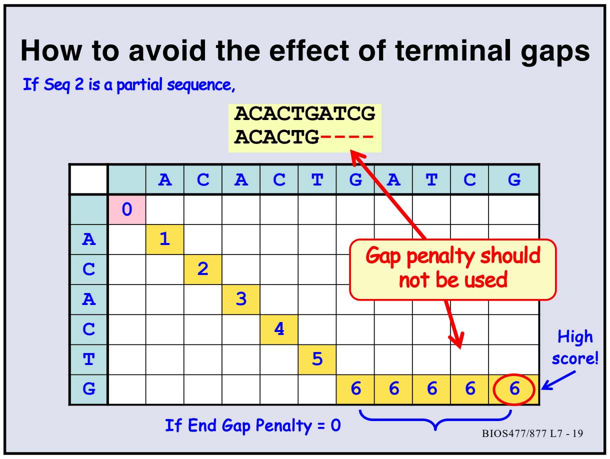
```
ACACTGATCG
ACACTG----
```

		A	C	A	C	T	G	A	T	C	G
	0										
A		1									
C			2								
A				3							
C					4						
T						5					
G							6	5	4	3	2

Gap penalty is enforced.

Low score

Match = 1, Mismatch = 0, Gap penalty = -1



How to avoid the effect of terminal gaps

If Seq 2 is a partial sequence,

```
ACACTGATCG
ACACTG----
```

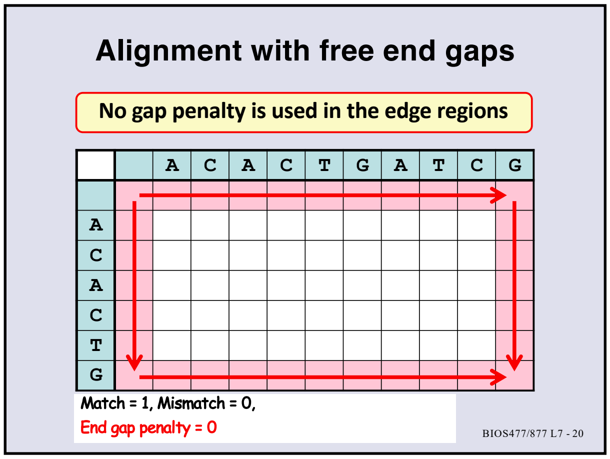
		A	C	A	C	T	G	A	T	C	G
	0										
A		1									
C			2								
A				3							
C					4						
T						5					
G							6	6	6	6	6

Gap penalty should not be used

High score!

If End Gap Penalty = 0

19

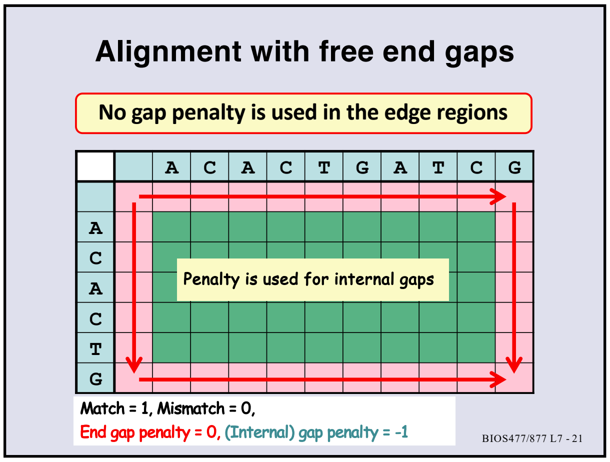


Alignment with free end gaps

No gap penalty is used in the edge regions

Match = 1, Mismatch = 0,
End gap penalty = 0

20



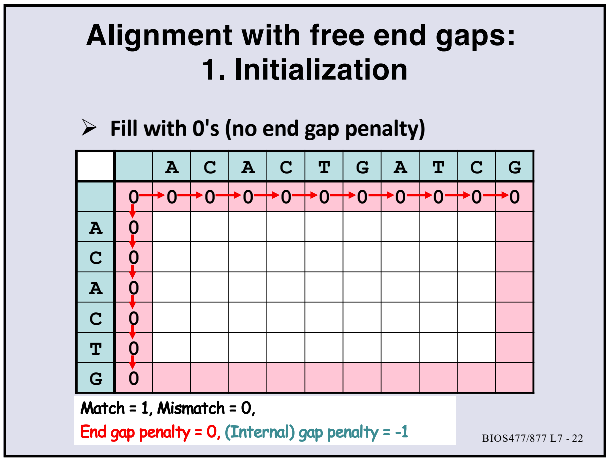
Alignment with free end gaps

No gap penalty is used in the edge regions

Penalty is used for internal gaps

Match = 1, Mismatch = 0,
End gap penalty = 0, (Internal) gap penalty = -1

21



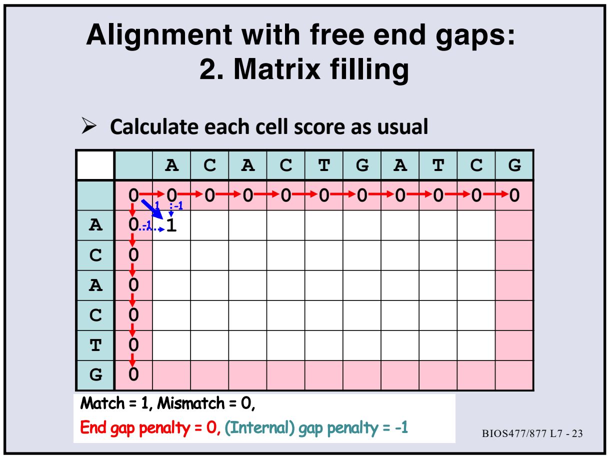
Alignment with free end gaps:
1. Initialization

- Fill with 0's (no end gap penalty)

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0										
C	0										
A	0										
C	0										
T	0										
G	0										

Match = 1, Mismatch = 0,
End gap penalty = 0, (Internal) gap penalty = -1

22



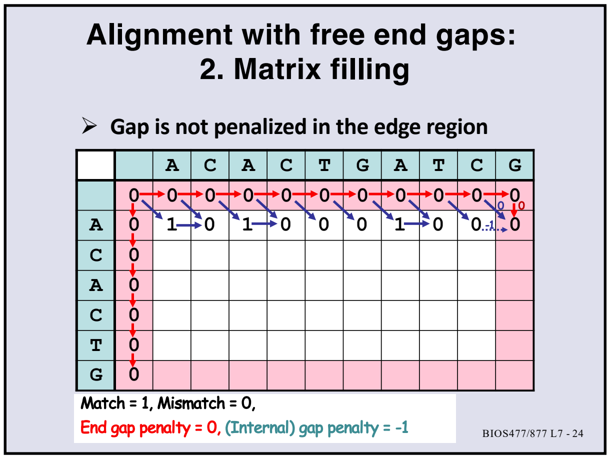
Alignment with free end gaps:
2. Matrix filling

- Calculate each cell score as usual

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1									
C	0										
A	0										
C	0										
T	0										
G	0										

Match = 1, Mismatch = 0,
End gap penalty = 0, (Internal) gap penalty = -1

23



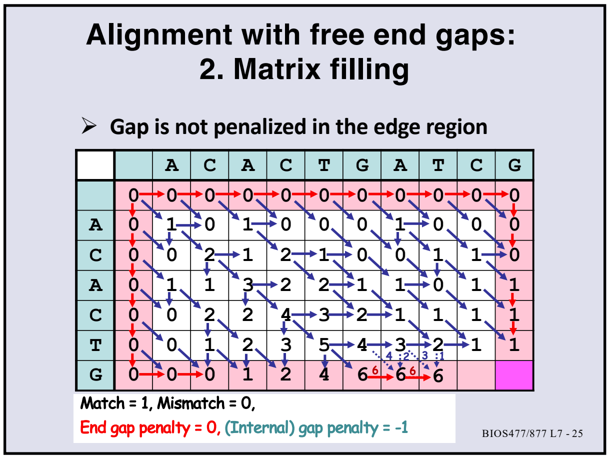
Alignment with free end gaps:
2. Matrix filling

- Gap is not penalized in the edge region

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0										
A	0										
C	0										
T	0										
G	0										

Match = 1, Mismatch = 0,
End gap penalty = 0, (Internal) gap penalty = -1

24

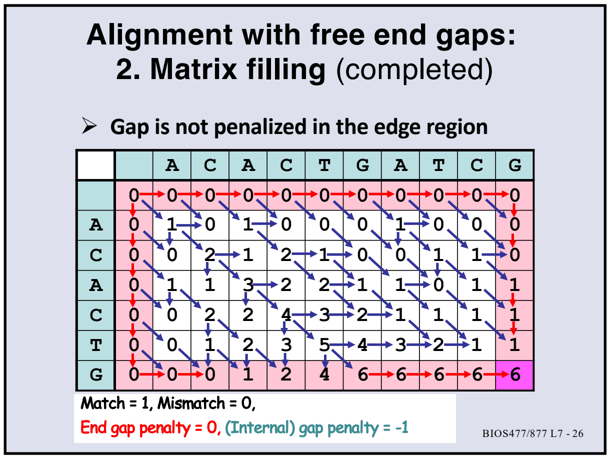


Alignment with free end gaps: 2. Matrix filling

- Gap is not penalized in the edge region

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0	0	2	1	2	1	0	0	1	1	0
A	0	1	1	3	2	2	1	1	0	1	1
C	0	0	2	2	4	3	2	1	1	1	1
T	0	0	1	2	3	5	4	3	2	1	1
G	0	0	0	1	2	4	6	6	6		

Match = 1, Mismatch = 0,
End gap penalty = 0, (Internal) gap penalty = -1

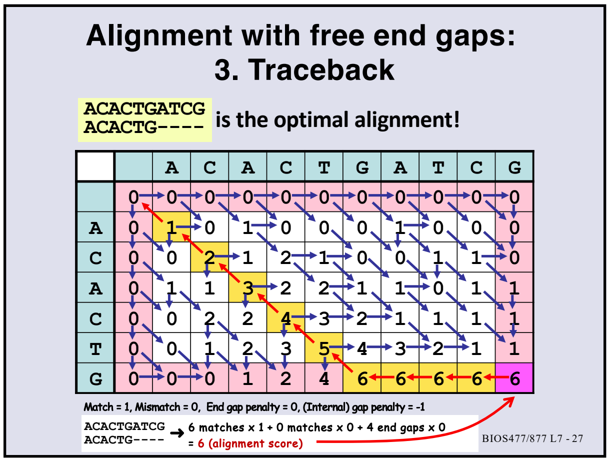


Alignment with free end gaps: 2. Matrix filling (completed)

- Gap is not penalized in the edge region

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0	0	2	1	2	1	0	0	1	1	0
A	0	1	1	3	2	2	1	1	0	1	1
C	0	0	2	2	4	3	2	1	1	1	1
T	0	0	1	2	3	5	4	3	2	1	1
G	0	0	0	1	2	4	6	6	6	6	6

Match = 1, Mismatch = 0,
End gap penalty = 0, (Internal) gap penalty = -1



Alignment with free end gaps: 3. Traceback

```
ACACTGATCG
ACACTG----
```

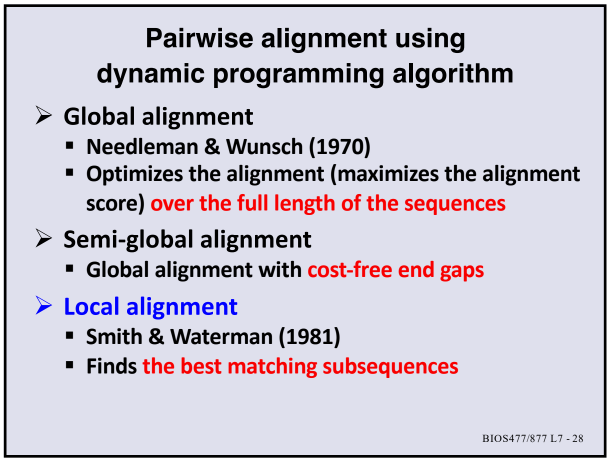
is the optimal alignment!

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0	0	2	1	2	1	0	0	1	1	0
A	0	1	1	3	2	2	1	1	0	1	1
C	0	0	2	2	4	3	2	1	1	1	1
T	0	0	1	2	3	5	4	3	2	1	1
G	0	0	0	1	2	4	6	6	6	6	6

Match = 1, Mismatch = 0, End gap penalty = 0, (Internal) gap penalty = -1

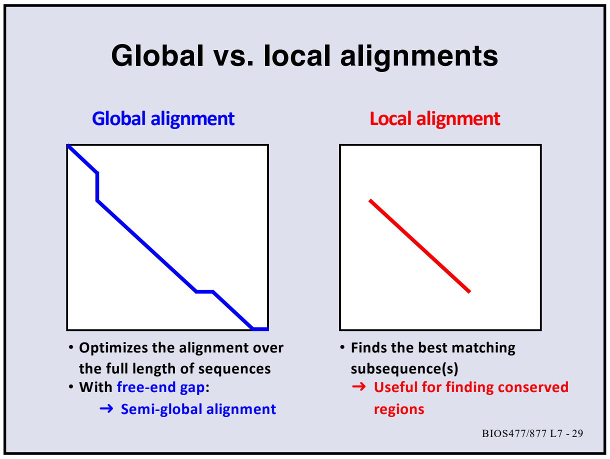
```
ACACTGATCG
ACACTG----
```

→ 6 matches x 1 + 0 matches x 0 + 4 end gaps x 0
= 6 (alignment score)



Pairwise alignment using dynamic programming algorithm

- Global alignment
  - Needleman & Wunsch (1970)
  - Optimizes the alignment (maximizes the alignment score) over the full length of the sequences
- Semi-global alignment
  - Global alignment with cost-free end gaps
- Local alignment
  - Smith & Waterman (1981)
  - Finds the best matching subsequences



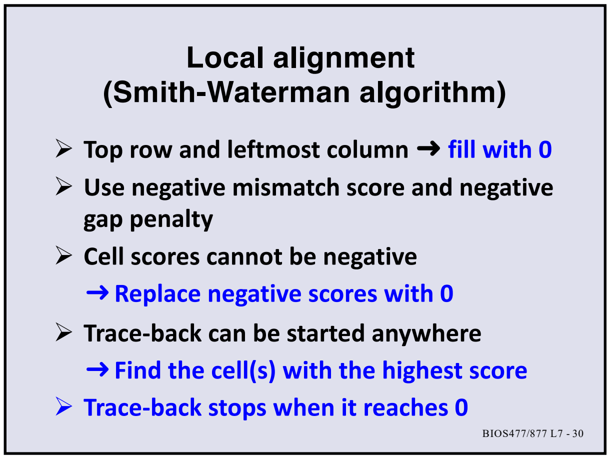
Global vs. local alignments

Global alignment

Local alignment

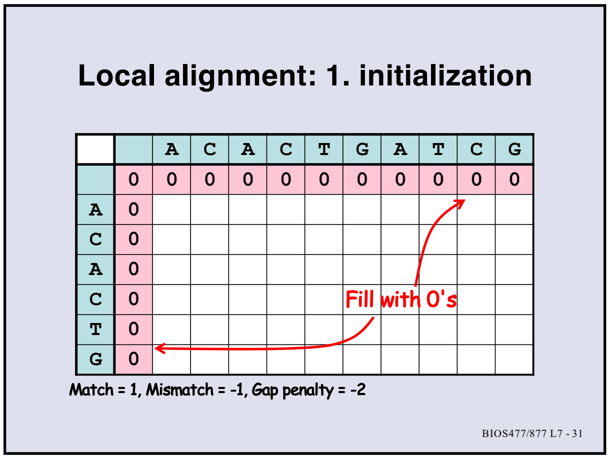
- Optimizes the alignment over the full length of sequences
- With free-end gap:
  → Semi-global alignment

- Finds the best matching subsequence(s)
  → Useful for finding conserved regions



Local alignment (Smith-Waterman algorithm)

- Top row and leftmost column → fill with 0
- Use negative mismatch score and negative gap penalty
- Cell scores cannot be negative
  → Replace negative scores with 0
- Trace-back can be started anywhere
  → Find the cell(s) with the highest score
- Trace-back stops when it reaches 0



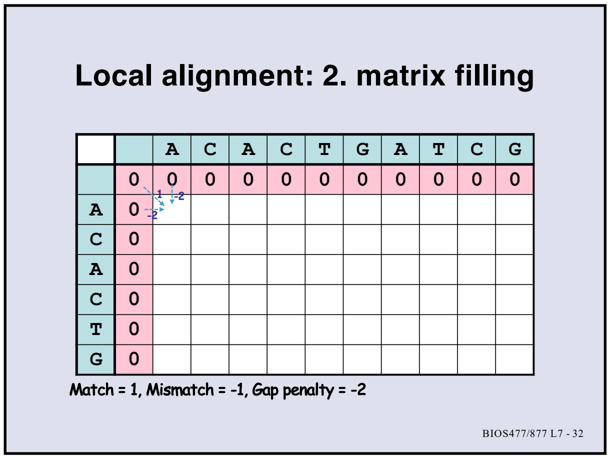
Local alignment: 1. initialization

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0										
C	0										
A	0										
C	0										
T	0										
G	0										

Fill with 0's

Match = 1, Mismatch = -1, Gap penalty = -2

31

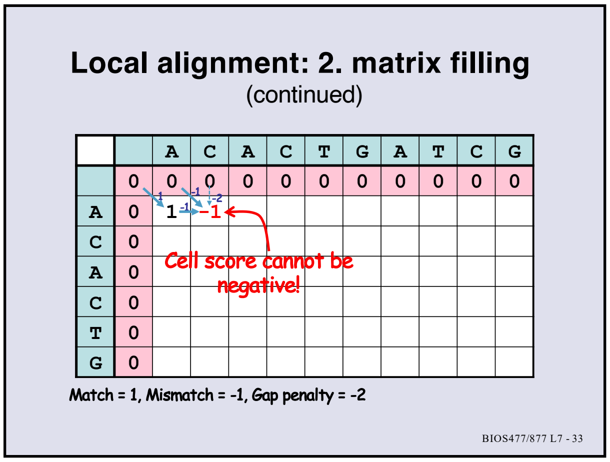


Local alignment: 2. matrix filling

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0										
C	0										
A	0										
C	0										
T	0										
G	0										

Match = 1, Mismatch = -1, Gap penalty = -2

32



Local alignment: 2. matrix filling

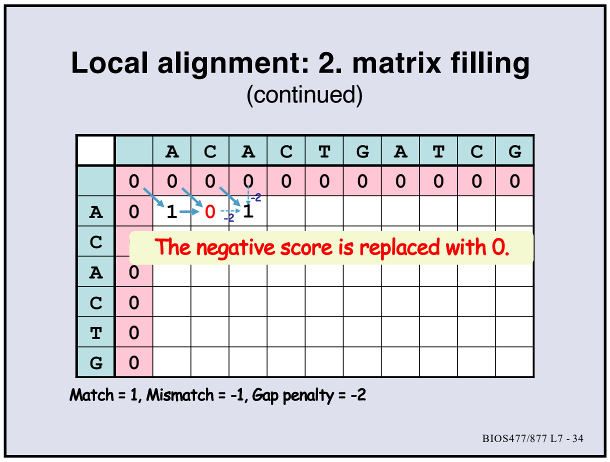
(continued)

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	-1								
C	0										
A	0										
C	0										
T	0										
G	0										

Cell score cannot be negative!

Match = 1, Mismatch = -1, Gap penalty = -2

33



Local alignment: 2. matrix filling

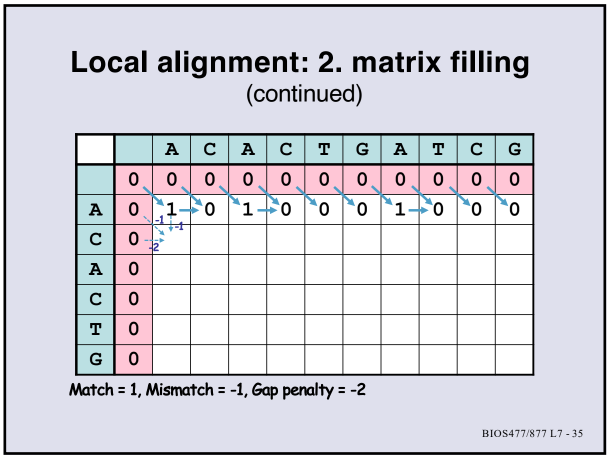
(continued)

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1							
C											
A	0										
C	0										
T	0										
G	0										

The negative score is replaced with 0.

Match = 1, Mismatch = -1, Gap penalty = -2

34



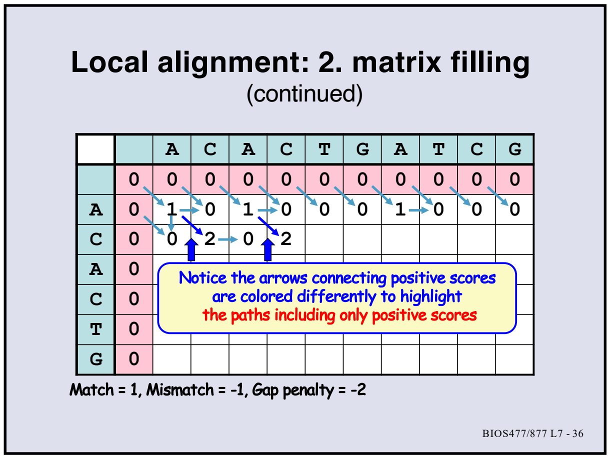
Local alignment: 2. matrix filling

(continued)

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0										
A	0										
C	0										
T	0										
G	0										

Match = 1, Mismatch = -1, Gap penalty = -2

35



Local alignment: 2. matrix filling

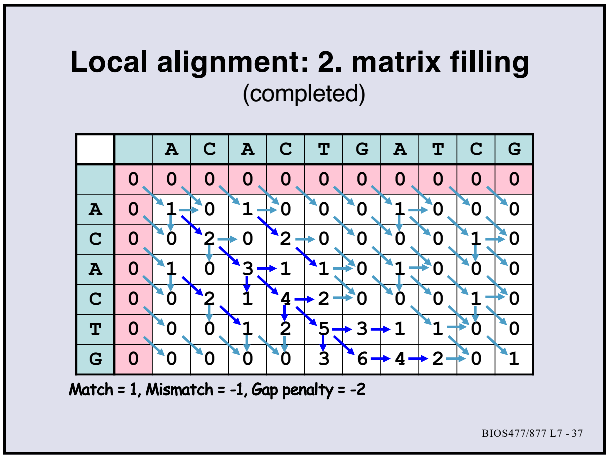
(continued)

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0	0	2	0	2						
A	0										
C	0										
T	0										
G	0										

Notice the arrows connecting positive scores
are colored differently to highlight
the paths including only positive scores

Match = 1, Mismatch = -1, Gap penalty = -2

36

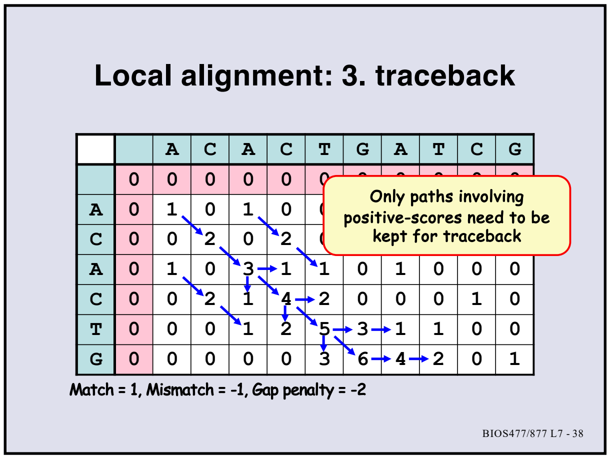


Local alignment: 2. matrix filling

(completed)

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0	0	2	0	2	0	0	0	0	1	0
A	0	1	0	3	1	1	0	1	0	0	0
C	0	0	2	1	4	2	0	0	0	1	0
T	0	0	0	1	2	5	3	1	1	0	0
G	0	0	0	0	0	3	6	4	2	0	1

Match = 1, Mismatch = -1, Gap penalty = -2

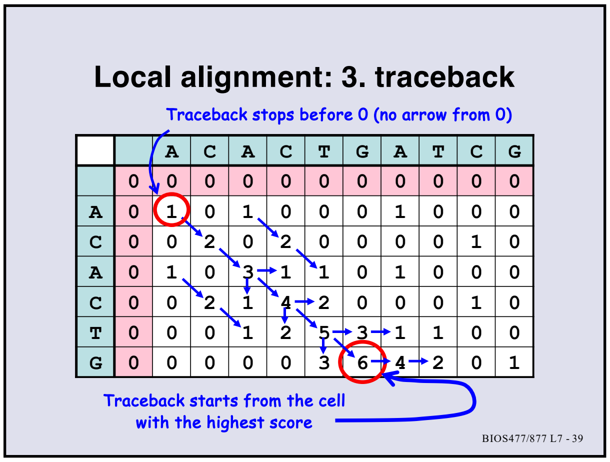


Local alignment: 3. traceback

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0					
A	0	1	0	1	0	0					
C	0	0	2	0	2	0					
A	0	1	0	3	1	1	0	1	0	0	0
C	0	0	2	1	4	2	0	0	0	1	0
T	0	0	0	1	2	5	3	1	1	0	0
G	0	0	0	0	0	3	6	4	2	0	1

Only paths involving positive-scores need to be kept for traceback

Match = 1, Mismatch = -1, Gap penalty = -2

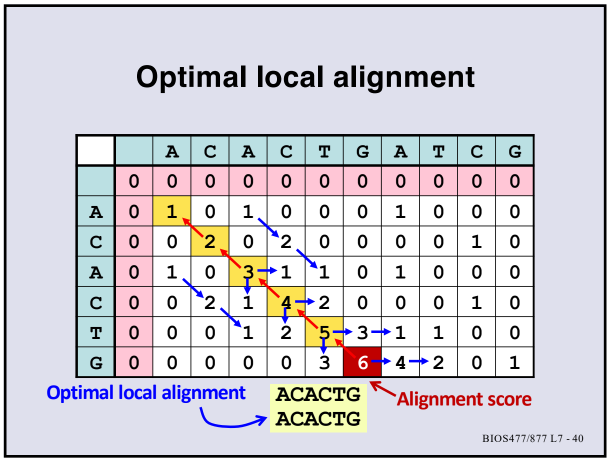


Local alignment: 3. traceback

Traceback stops before 0 (no arrow from 0)

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0	0	2	0	2	0	0	0	0	1	0
A	0	1	0	3	1	1	0	1	0	0	0
C	0	0	2	1	4	2	0	0	0	1	0
T	0	0	0	1	2	5	3	1	1	0	0
G	0	0	0	0	0	3	6	4	2	0	1

Traceback starts from the cell with the highest score



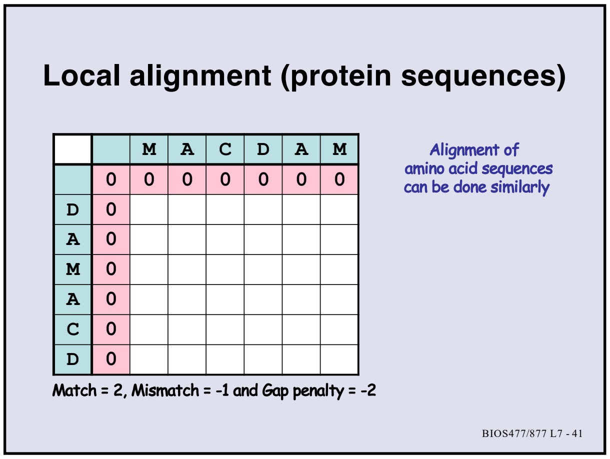
Optimal local alignment

		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0	0	2	0	2	0	0	0	0	1	0
A	0	1	0	3	1	1	0	1	0	0	0
C	0	0	2	1	4	2	0	0	0	1	0
T	0	0	0	1	2	5	3	1	1	0	0
G	0	0	0	0	0	3	6	4	2	0	1

Optimal local alignment

ACACTG
ACACTG

Alignment score

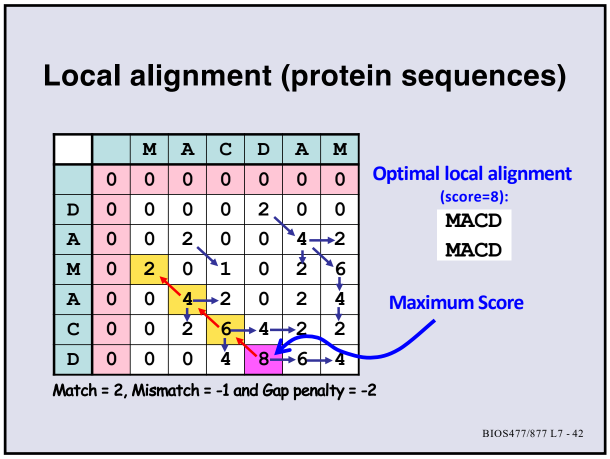


Local alignment (protein sequences)

		M	A	C	D	A	M
	0	0	0	0	0	0	0
D	0						
A	0						
M	0						
A	0						
C	0						
D	0						

Alignment of amino acid sequences can be done similarly

Match = 2, Mismatch = -1 and Gap penalty = -2



Local alignment (protein sequences)

		M	A	C	D	A	M
	0	0	0	0	0	0	0
D	0	0	0	0	2	0	0
A	0	0	2	0	0	4	2
M	0	2	0	1	0	2	6
A	0	0	4	2	0	2	4
C	0	0	2	6	4	2	2
D	0	0	0	4	8	6	4

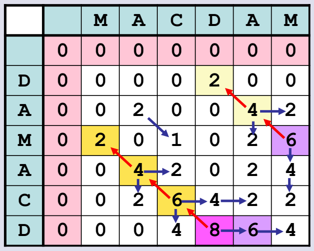
Optimal local alignment (score=8):

MACD
MACD

Maximum Score

Match = 2, Mismatch = -1 and Gap penalty = -2

Local alignment (protein sequences)



		M	A	C	D	A	M
	0	0	0	0	0	0	0
D	0	0	0	0	2	0	0
A	0	0	2	0	0	4	2
M	0	2	0	1	0	2	6
A	0	0	4	2	0	2	4
C	0	0	2	6	4	2	2
D	0	0	0	4	8	6	4

Match = 2, Mismatch = -1 and Gap penalty = -2

Optimal local alignment
(score: 8)

MACD
MACD

You can also find
Suboptimal local alignment
(score: 6)

DAM
DAM

MACDA
MACD−

Graduate Only Assignment

- Report 1: March 8
  - → Short proposal (2-3 pages)
- Report 2: May 3
  - → Final report (5-7 pages)
- Comparative analysis of bioinformatics methods
  - ○ *e.g.,* Multiple alignment methods:
    - Clustal W vs. MUSCLE vs. MAFFT
  - → Compare three methods
  - → Choose appropriate input datasets (3 or more)
  - → Read the Guidelines!
- Undergraduate students can do these assignments for bonus points (both reports need to be submitted)