Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

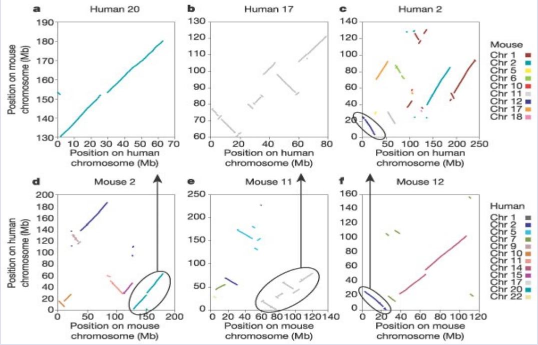
Lecture 7

TODAY'S TOPICS

- Assignments 1 & 3 review

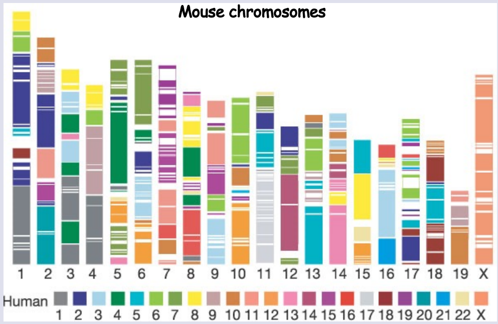
- ➢ Dotplot examples
- ➢ Pairwise Alignment continues ...
  - Semi-global alignment
  - Local alignment
- ➢ Graduate Only Assignment

Comparative genomics using dot plot



Mouse Genome Sequencing Consortium (2002) *Nature* 420: 520-562

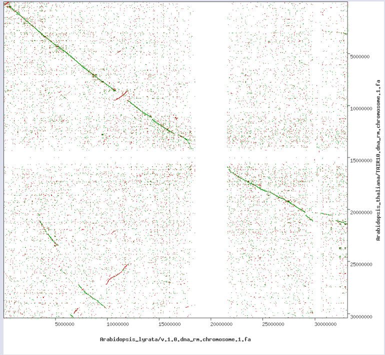
Comparative genomics using dot plot



Mouse Genome Sequencing Consortium (2002) *Nature* 420: 520-562

Dot plot by YASS

https://bioinfo.univ-lille.fr/yass/yass.php

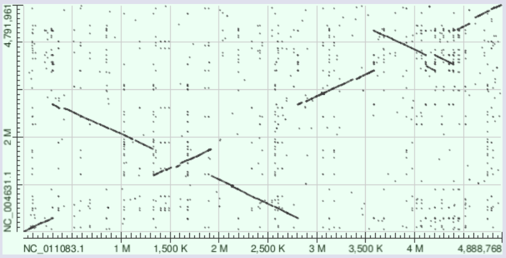


Arabidopsis thaliana
(chromosome. 1)
vs.
Arabidopsis lyrata
(chrromosome 1)

Dot plot by Blast (bl2seq)

https://blast.ncbi.nlm.nih.gov/Blast.cgi

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



Dynamic Programming Algorithm

- Three steps for sequence alignment

(Eddy, 2004)

1. Initialization
2. Matrix filling (Scoring)
3. Traceback (Alignment)

- Sequences a and b
- Simple scoring scheme:

→ $s(a_i,b_j) = 2$ ($a_i=b_j$, match)
→ $s(a_i,b_j) = -1$ ($a_i \neq b_j$, mismatch)
→ $w = -2$ (gap penalty)

	[a]	A	A	T	C	T	G	C
[b]								
A								
T								
T								
C								
T								
C								

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

Alignment including Terminal (End) Gaps

Is

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

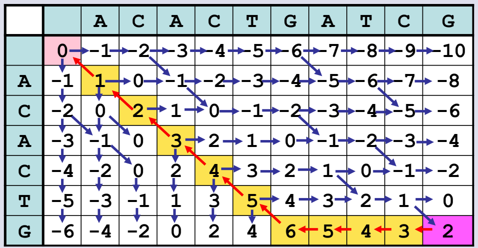
optimal alignment?

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

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

Alignment including Terminal (End) Gaps

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

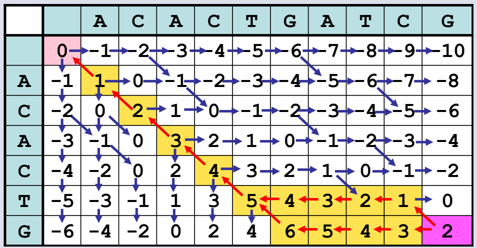


		A	C	A	C	T	G	A	T	C	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8
C	-2	0	2	1	0	-1	-2	-3	-4	-5	-6
A	-3	-1	0	3	2	1	0	-1	-2	-3	-4
C	-4	-2	0	2	4	3	2	1	0	-1	-2
T	-5	-3	-1	1	3	5	4	3	2	1	0
G	-6	-4	-2	0	2	4	6	5	4	3	2

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

Alignment including Terminal (End) Gaps

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



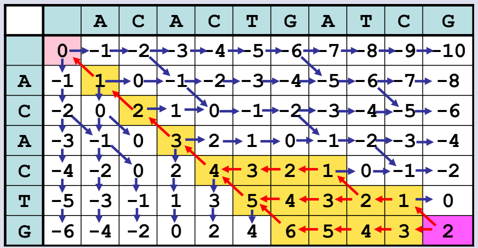
		A	C	A	C	T	G	A	T	C	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8
C	-2	0	2	1	0	-1	-2	-3	-4	-5	-6
A	-3	-1	0	3	2	1	0	-1	-2	-3	-4
C	-4	-2	0	2	4	3	2	1	0	-1	-2
T	-5	-3	-1	1	3	5	4	3	2	1	0
G	-6	-4	-2	0	2	4	6	5	4	3	2

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

Alignment including Terminal (End) Gaps

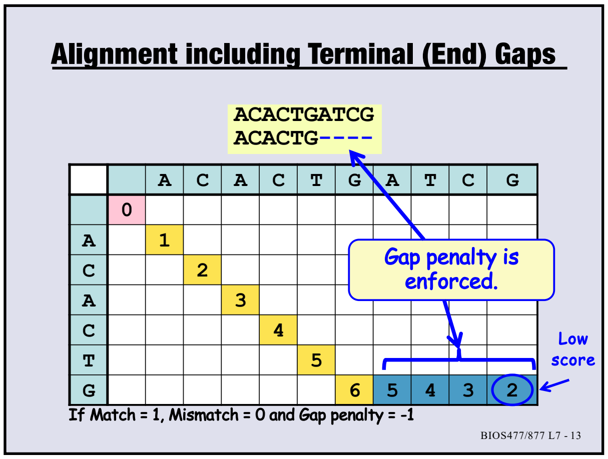
If Seq 2 is a partial sequence, this alignment should have the maximum score

```
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

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



Alignment including Terminal (End) Gaps

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

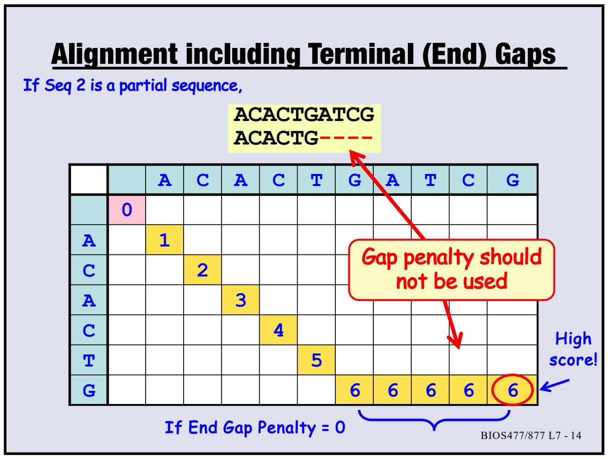
Gap penalty is enforced.

Low score

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

BIOS477/877 L7 - 13

13



Alignment including Terminal (End) 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

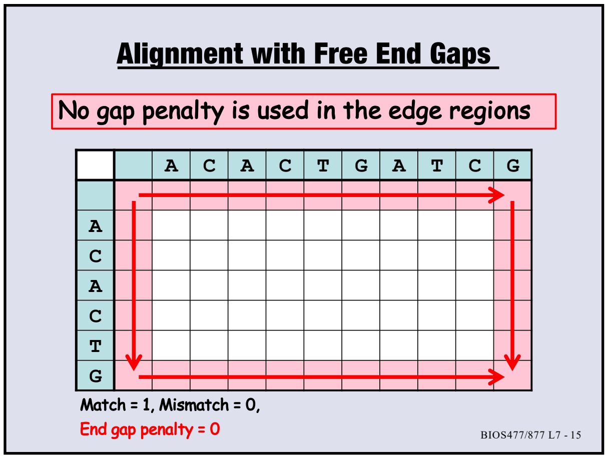
Gap penalty should not be used

High score!

If End Gap Penalty = 0

BIOS477/877 L7 - 14

14



Alignment with Free End Gaps

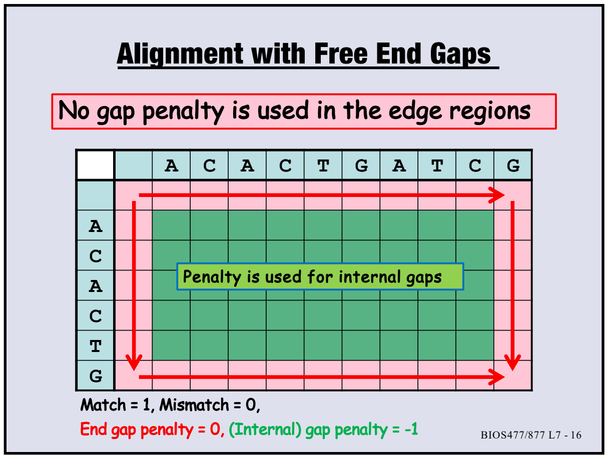
No gap penalty is used in the edge regions

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

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

BIOS477/877 L7 - 15

15



Alignment with Free End Gaps

No gap penalty is used in the edge regions

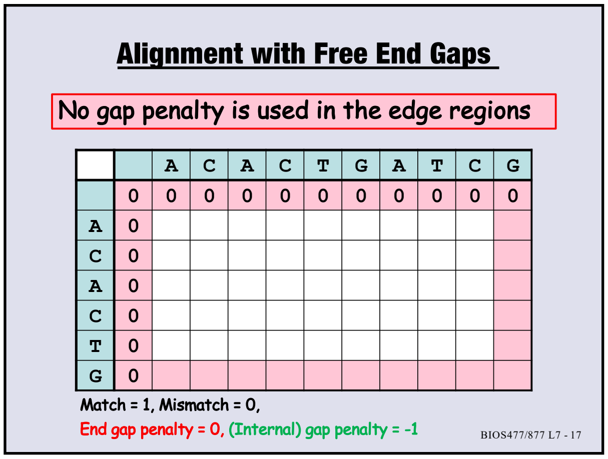
		A	C	A	C	T	G	A	T	C	G
A											
C											
A											
C											
T											
G											

Penalty is used for internal gaps

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

BIOS477/877 L7 - 16

16



Alignment with Free End Gaps

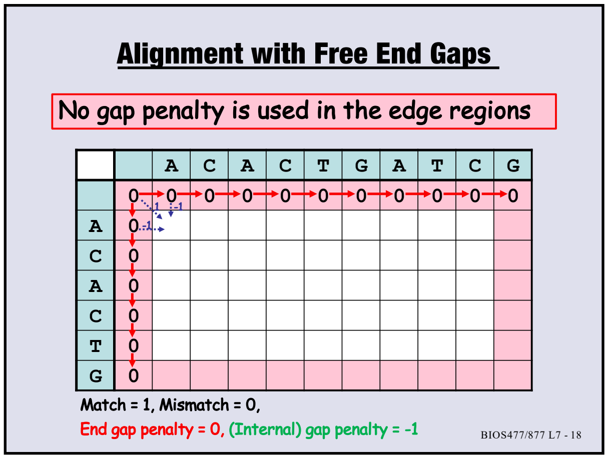
No gap penalty is used in the edge regions

		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

BIOS477/877 L7 - 17

17



Alignment with Free End Gaps

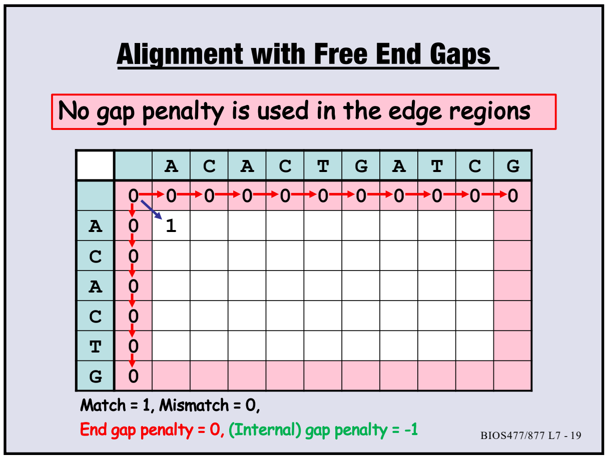
No gap penalty is used in the edge regions

		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

BIOS477/877 L7 - 18

18



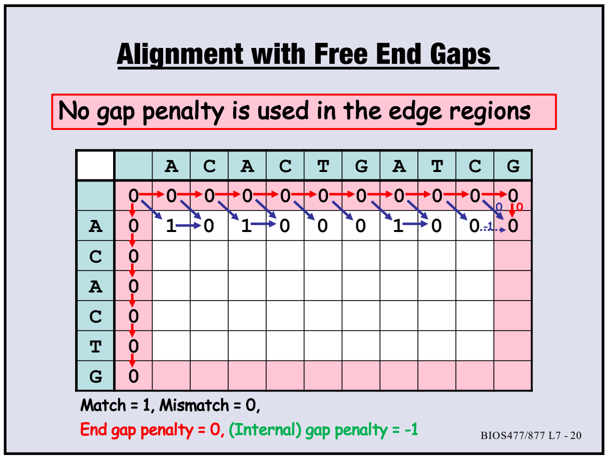
Alignment with Free End Gaps

No gap penalty is used in the edge regions

		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



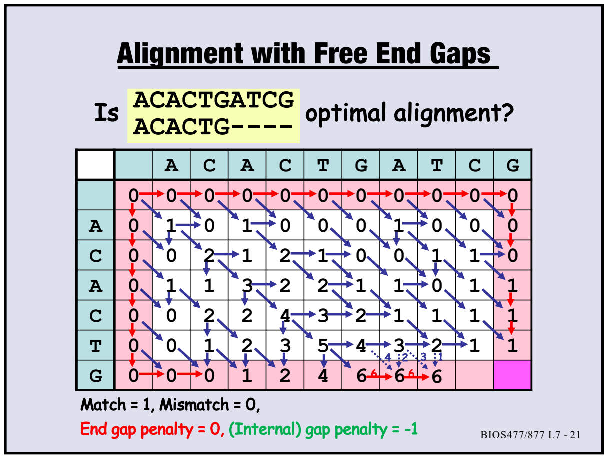
Alignment with Free End Gaps

No gap penalty is used in the edge regions

		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



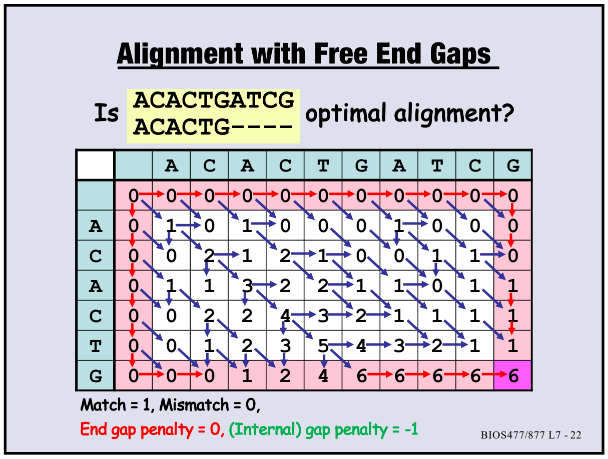
Alignment with Free End Gaps

Is ACACTGATCG
ACACTG---- 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		

Match = 1, Mismatch = 0,

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



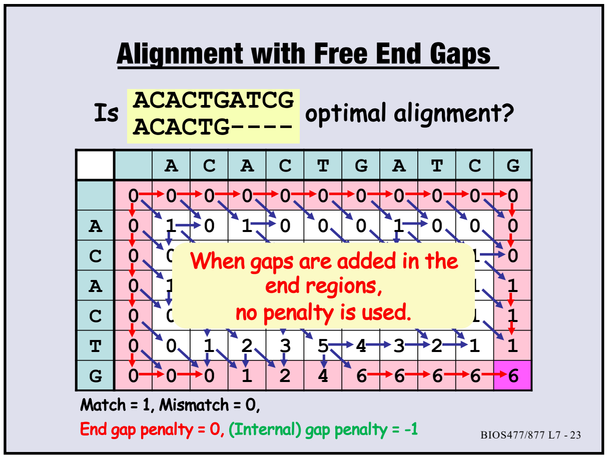
Alignment with Free End Gaps

Is ACACTGATCG
ACACTG---- 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



Alignment with Free End Gaps

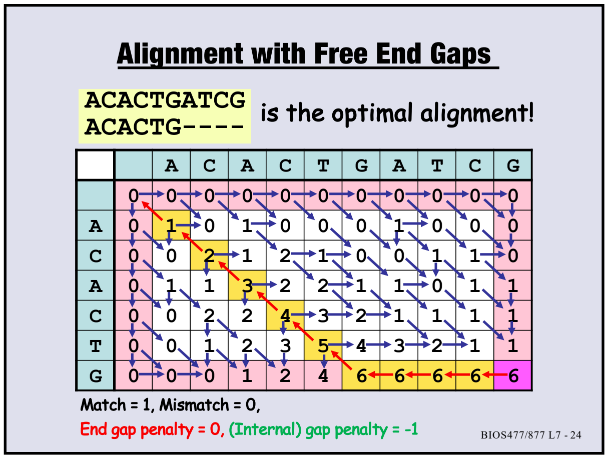
Is ACACTGATCG
ACACTG---- optimal alignment?

		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								1	0
A	0	1								1	1
C	0	0								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

When gaps are added in the
end regions,
no penalty is used.

Match = 1, Mismatch = 0,

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



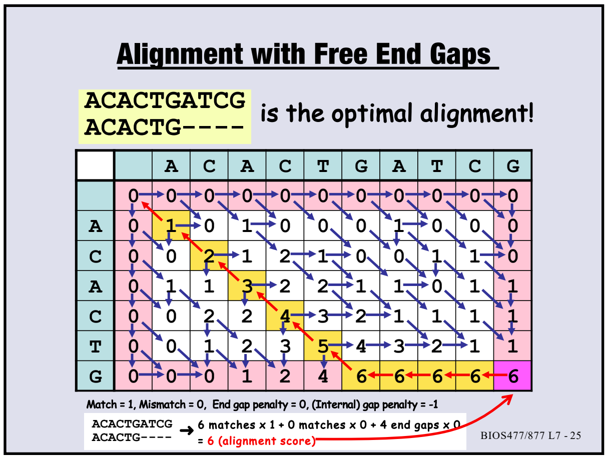
Alignment with Free End Gaps

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



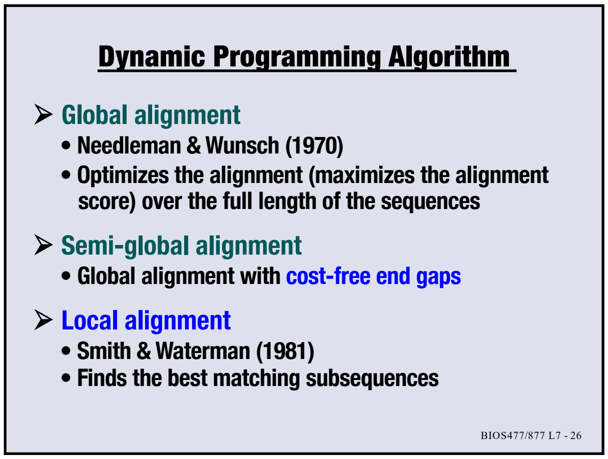
Alignment with Free End Gaps

```
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

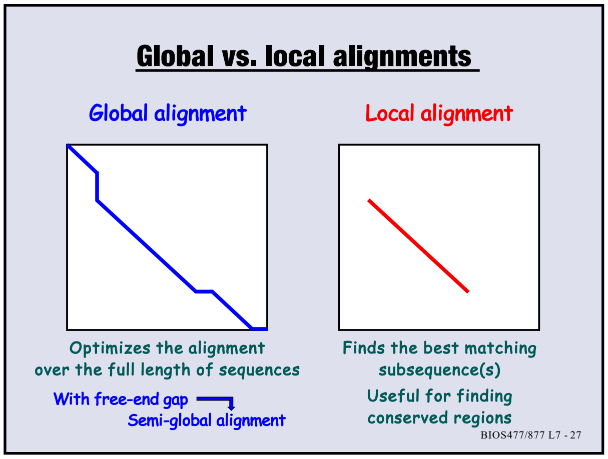
Match = 1, Mismatch = 0, End gap penalty = 0, (Internal) gap penalty = -1

```
ACACTGATCG
ACACTG----
```
→ 6 matches x 1 + 0 matches x 0 + 4 end gaps x 0 = 6 (alignment score)



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

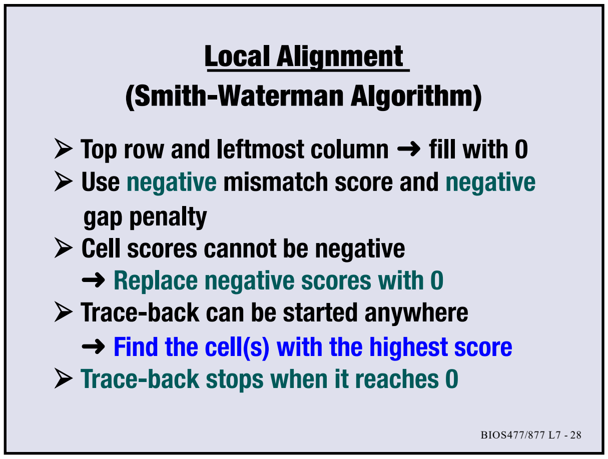
Optimizes the alignment over the full length of sequences

With free-end gap → Semi-global alignment

Local alignment

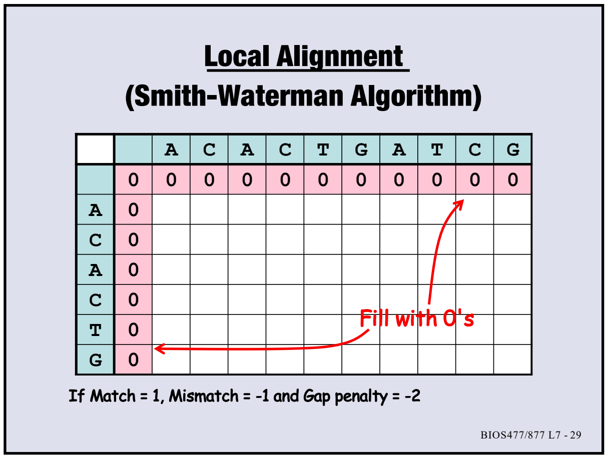
Finds the best matching subsequence(s)

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

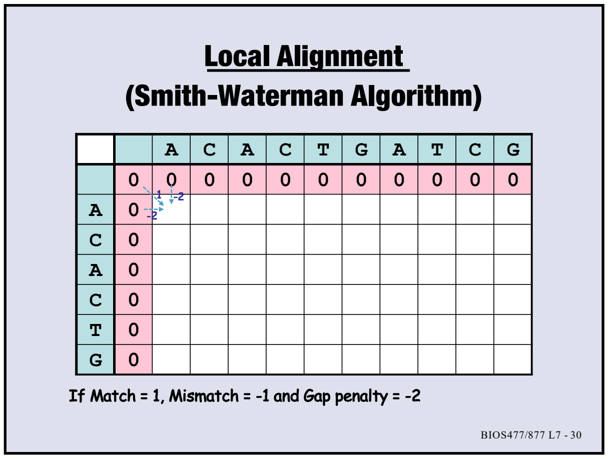


Local Alignment
(Smith-Waterman Algorithm)

		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

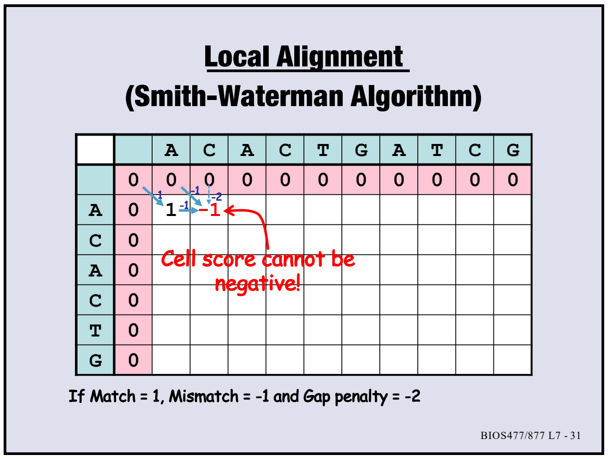
If Match = 1, Mismatch = -1 and Gap penalty = -2



Local Alignment
(Smith-Waterman Algorithm)

		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										

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



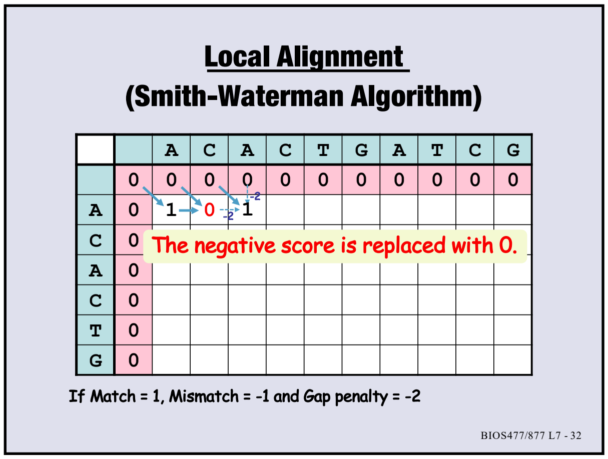
Local Alignment
(Smith-Waterman Algorithm)

		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!

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

31



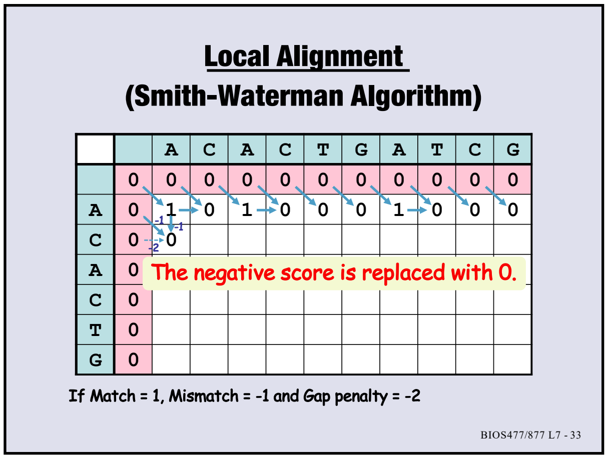
Local Alignment
(Smith-Waterman Algorithm)

		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	0										
A	0										
C	0										
T	0										
G	0										

The negative score is replaced with 0.

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

32



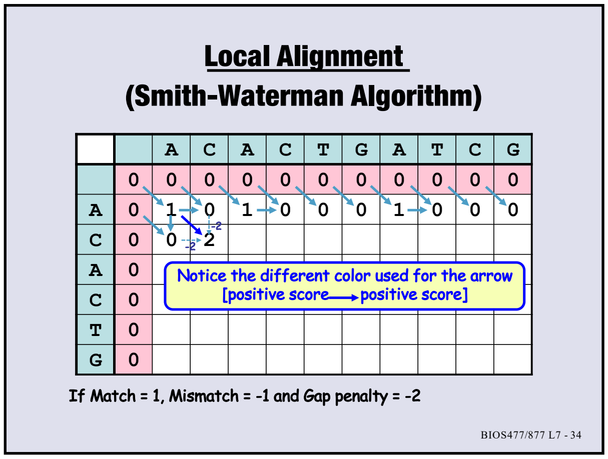
Local Alignment
(Smith-Waterman Algorithm)

		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									
A	0										
C	0										
T	0										
G	0										

The negative score is replaced with 0.

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

33



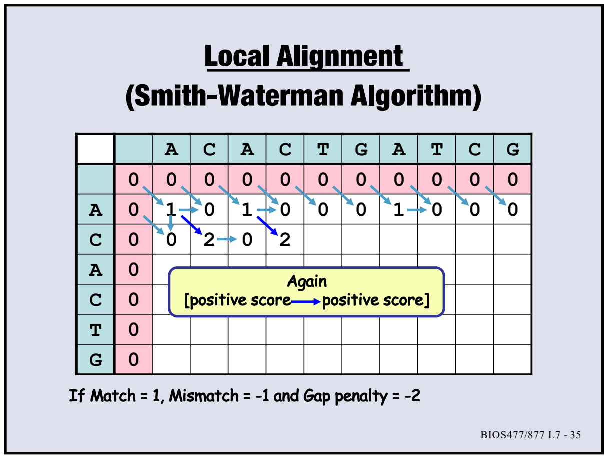
Local Alignment
(Smith-Waterman Algorithm)

		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								
A	0										
C	0										
T	0										
G	0										

Notice the different color used for the arrow
[positive score ⟶ positive score]

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

34



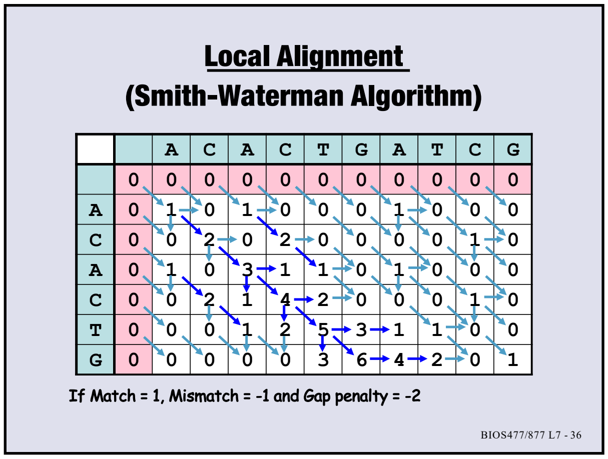
Local Alignment
(Smith-Waterman Algorithm)

		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										

Again
[positive score ⟶ positive score]

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

35

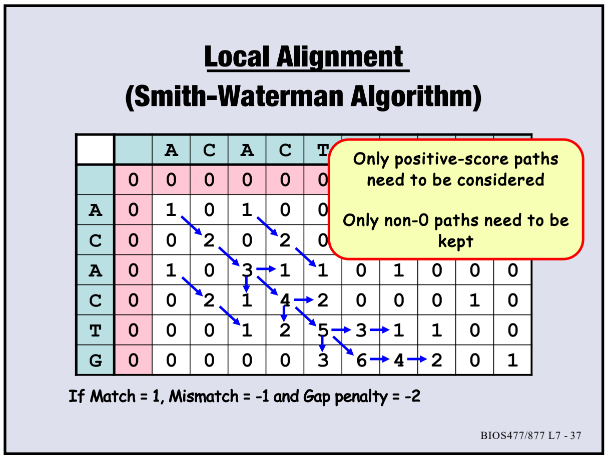


Local Alignment
(Smith-Waterman Algorithm)

		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

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

36



Local Alignment
(Smith-Waterman Algorithm)

		A	C	A	C	T					
	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

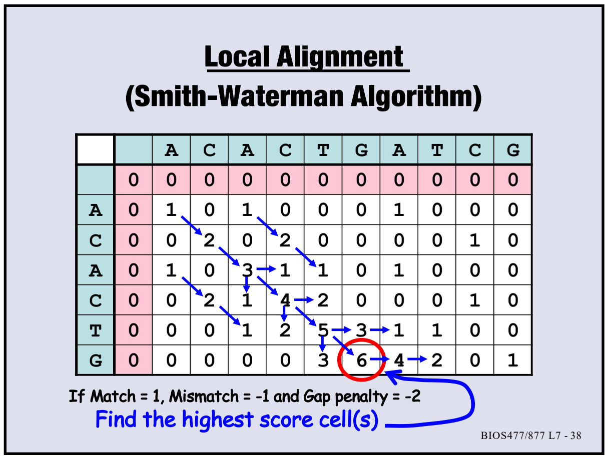
Only positive-score paths need to be considered

Only non-0 paths need to be kept

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

BIOS477/877 L7 - 37

37



Local Alignment
(Smith-Waterman Algorithm)

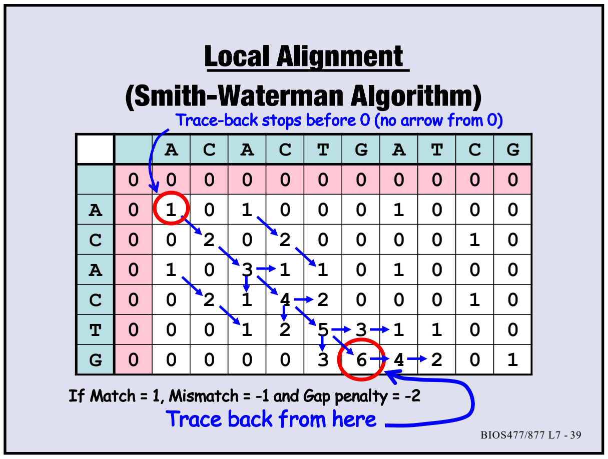
		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0	0	2	0	2	0	0	0	0	1	0
A	0	1	0	3	1	1	0	1	0	0	0
C	0	0	2	1	4	2	0	0	0	1	0
T	0	0	0	1	2	5	3	1	1	0	0
G	0	0	0	0	0	3	6	4	2	0	1

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

Find the highest score cell(s)

BIOS477/877 L7 - 38

38



Local Alignment
(Smith-Waterman Algorithm)

Trace-back stops before 0 (no arrow from 0)

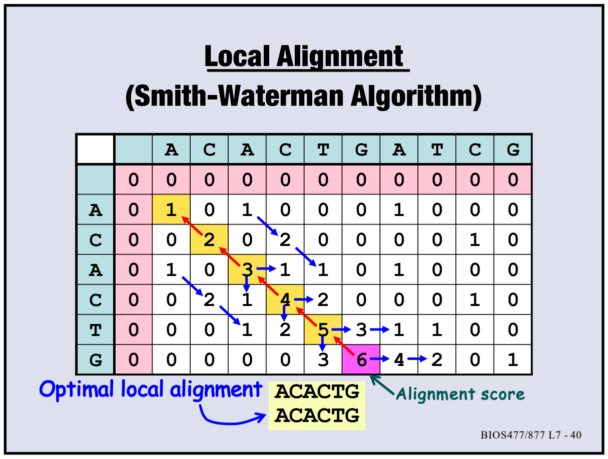
		A	C	A	C	T	G	A	T	C	G
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	1	0	0	0
C	0	0	2	0	2	0	0	0	0	1	0
A	0	1	0	3	1	1	0	1	0	0	0
C	0	0	2	1	4	2	0	0	0	1	0
T	0	0	0	1	2	5	3	1	1	0	0
G	0	0	0	0	0	3	6	4	2	0	1

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

Trace back from here

BIOS477/877 L7 - 39

39



Local Alignment
(Smith-Waterman Algorithm)

		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

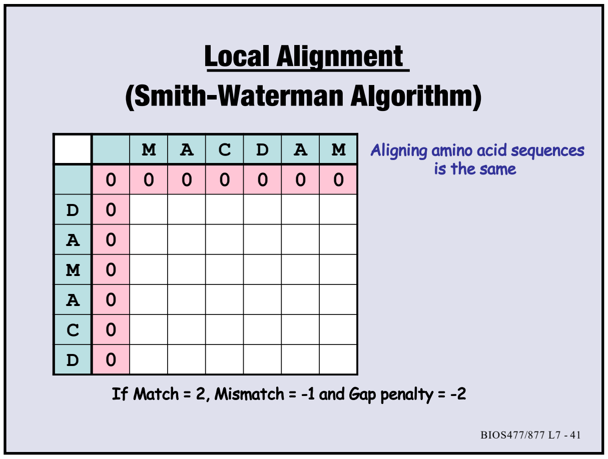
Optimal local alignment

ACACTG
ACACTG

Alignment score

BIOS477/877 L7 - 40

40



Local Alignment
(Smith-Waterman Algorithm)

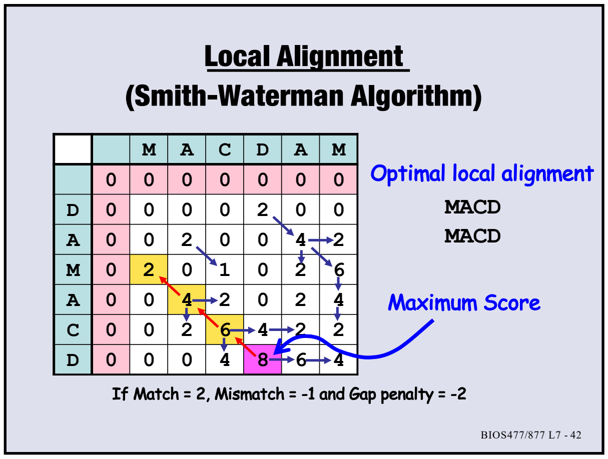
		M	A	C	D	A	M
	0	0	0	0	0	0	0
D	0						
A	0						
M	0						
A	0						
C	0						
D	0						

Aligning amino acid sequences is the same

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

BIOS477/877 L7 - 41

41



Local Alignment
(Smith-Waterman Algorithm)

		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

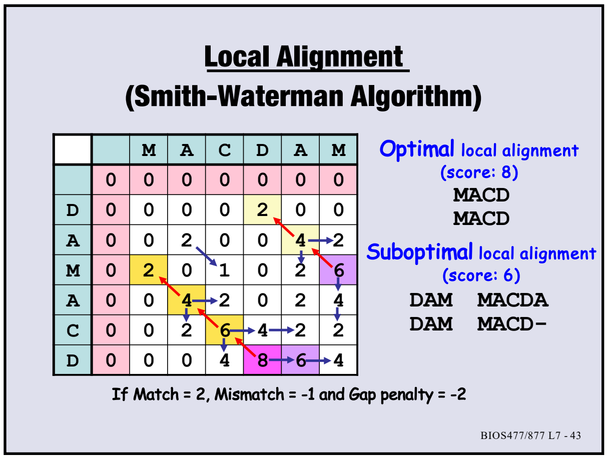
MACD
MACD

Maximum Score

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

BIOS477/877 L7 - 42

42



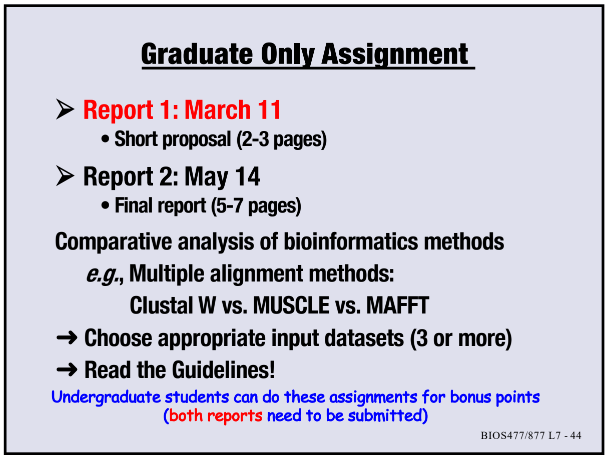
Local Alignment
(Smith-Waterman Algorithm)

		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

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

Optimal local alignment
(score: 8)
MACD
MACD

Suboptimal local alignment
(score: 6)
DAM MACDA
DAM MACD-



Graduate Only Assignment

- ➢ **Report 1: March 11**
  - Short proposal (2-3 pages)
- ➢ Report 2: May 14
  - Final report (5-7 pages)

Comparative analysis of bioinformatics methods
e.g., Multiple alignment methods:
Clustal W vs. MUSCLE vs. MAFFT

➜ 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)