Pairwise Sequence Alignment

Today's Goal

```
> DNA Sequence 1
ACTGCGATTGACGTACGATCATCGTACGATCATCATGCTGAGCTATCATCATCGTACTGA
TCGTAGACTACGTAGCTAGCATGCAGTCTGATGACGTCATGCTGACGTAGCATGC

> DNA Sequence 2
GACTAGCAGCGAGAGATCTCTCGAGTATGCGAGAGCTGATGCATCTACGTATGCAGTCGT
GCTAATGCGAGCGTATACGCGGGCATGTAGAGACTTCCTAGTAC
```

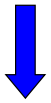
How similar are two sequences?

```
> Protein Sequence 1
KGLAHDGHNADFLKAMGGPIAFPIDADPFIDFKLHMNI

> Protein Sequence 2
LHASDGFKHSADFHNAIFDPAFLKADFPIMADSFN
```


Alignment

```
CGTAGCAGC
TGTAGTTCAGC
```

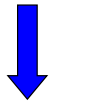


```
CGTAG--CAGC
 ||||  ||||
TGTAGTTCAGC
```


Scoring Alignments

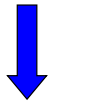
Match: +5 Mismatch: -4 Gap: -6

```
CGCGTTA
CGGGTCA
```



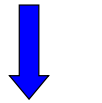
```
CGCGTTA
|| || |
CGGGTCA
```

```
ACTCGATCG
ACTTCG
```

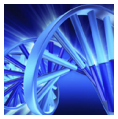


```
ACTCGATCG
|||   |||
ACT---TCG
```

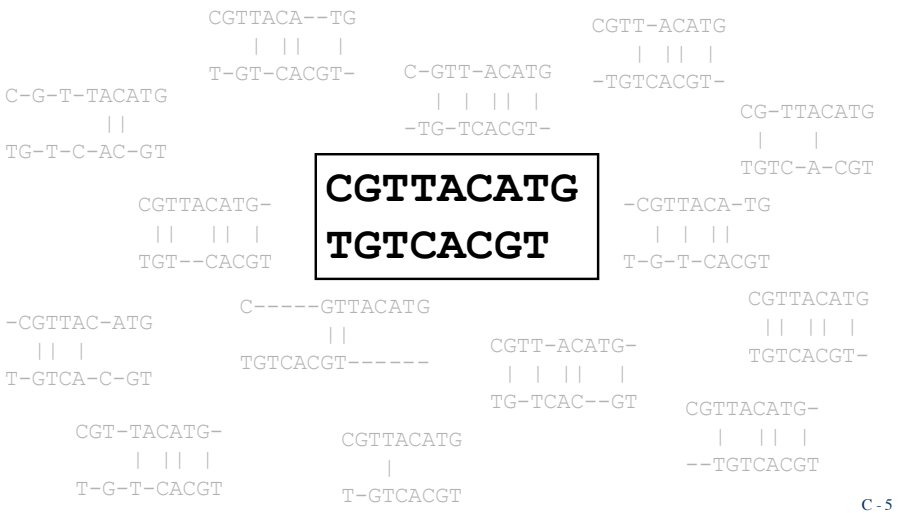
```
CGTAGCAGCT
CATACAGGACT
```

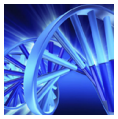


```
CGTAGCAG--CT
| ||  |||  ||
CATA-CAGGACT
```



Use the Optimal (best scoring) Alignment



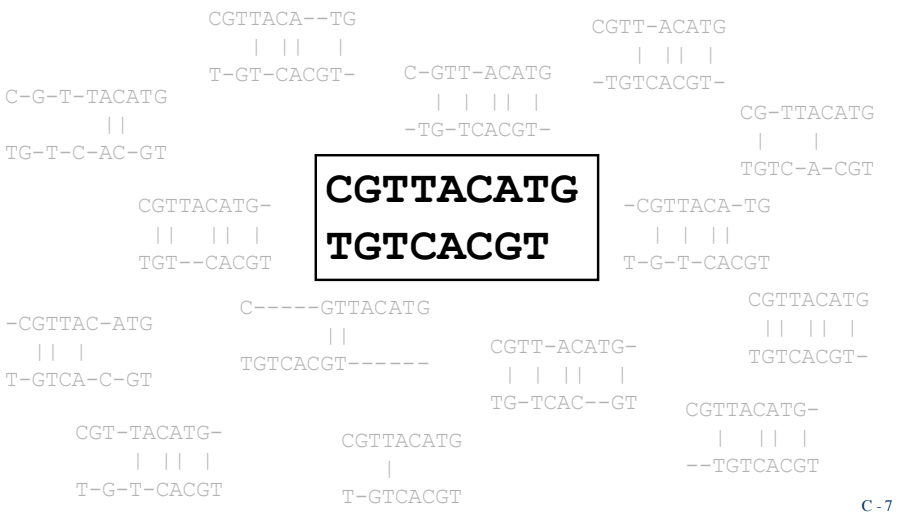


Pairwise Sequence Alignment

Pairwise Alignment Problem:

Given two sequences, determine their optimal (i.e., best scoring) alignment.

How Many Different Alignments?



The Elegance of Alignment

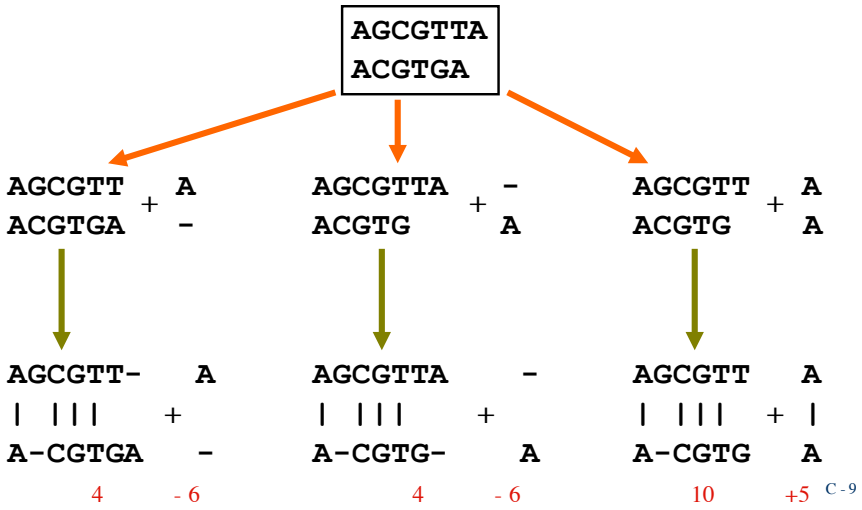
The problem of finding the best alignment of two sequences has two important properties:

(1) The solution can be found by looking at the solutions to subproblems

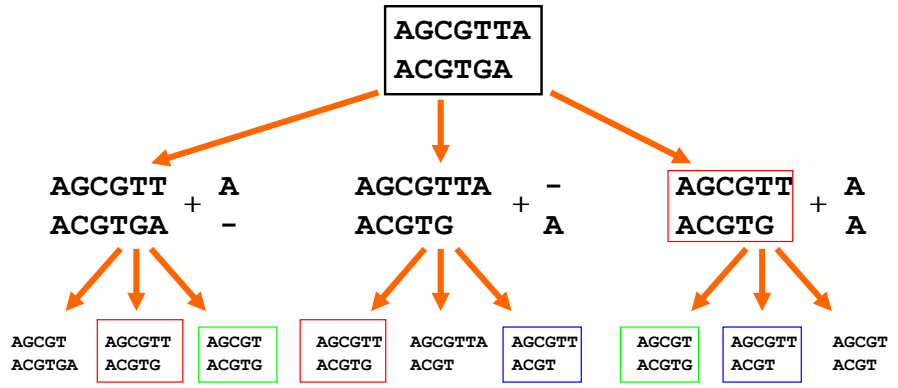
(2) Subproblems often overlap

Indeed, to find the best alignment of two sequences, we need only look at 3 slightly smaller alignments (i.e., remove one or two characters from the sequences).

The Elegance of Alignment



The Elegance of Alignment



The Elegance of Alignment

The problem of finding the best alignment of two sequences has two important properties:

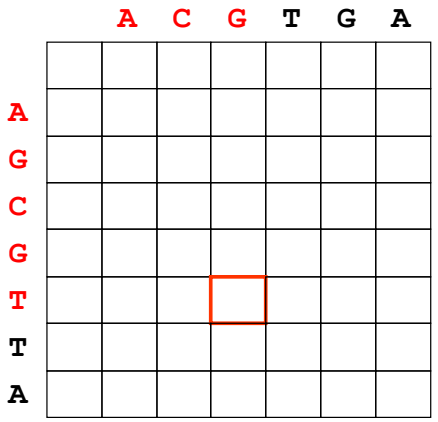
(1) The solution can be found by looking at the solutions to subproblems

(2) Subproblems often overlap

The method for determining the best alignment is known as a *dynamic programming algorithm*.

Score Table

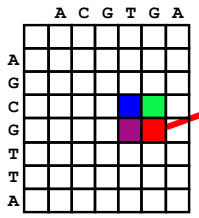
AGCGTTA
ACGTGA



AGCGT
ACG

How Is Each Entry in the Table Determined?

- Each entry depends on 3 previous entries (because of problem's "elegance")
- Each entry also depends on scores used (match, mismatch, gap)



max of 3

- Score in block to the left minus *gap penalty*
- Score in block above minus *gap penalty*
- Score in block diagonally left/above plus *match/mismatch score*

Alignment Score Table

AGCGTTA
ACGTGA

		A	C	G	T	G	A
	0	-6	-12	-18	-24	-30	-36
A	-6						
G	-12						
C	-18						
G	-24						
T	-30						
T	-36						
A	-42						

Alignment Score Table

AGCGTTA
ACGTGA

		A	C	G	T	G	A
	0	-6	-12	-18	-24	-30	-36
A	-6	5	-1	-7	-13	-19	-25
G	-12	-1	1	4	-2	-8	-14
C	-18	-7	4	-2	0	-6	-12
G	-24	-13	-2	9	3	5	-1
T	-30	-19	-8	3	14	8	2
T	-36	-25	-14	-3	8	10	4
A	-42	-31	-20	-9	2	4	15

How Do We Re-Create the Alignment?

AGCGTTA
ACGTGA

		A	C	G	T	G	A
	0	-6	-12	-18	-24	-30	-36
A	-6	5	-1	-7	-13	-19	-25
G	-12	-1	1	4	-2	-8	-14
C	-18	-7	4	-2	0	-6	-12
G	-24	-13	-2	9	3	5	-1
T	-30	-19	-8	3	14	8	2
T	-36	-25	-14	-3	8	10	4
A	-42	-31	-20	-9	2	4	15

```
AGCGTTA
| |||  |
A-CGTGA
```

Let's Recap, Shall We?

- The problem of finding the best alignment for two sequences has a couple of interesting properties:
  - (1) The best alignment can be determined using the best alignments of subproblems
  - (2) Subproblems often overlap
- Because of these properties, we can fill in a table of solutions to subproblems
- Each table entry is determined from 3 of the preceding entries
- The filled-in table tells us the best alignment!

Global Alignment

AGCGTTA
ACGTGA

		A	C	G	T	G	A
	0	-6	-12	-18	-24	-30	-36
A	-6	5	-1	-7	-13	-19	-25
G	-12	-1	1	4	-2	-8	-14
C	-18	-7	4	-2	0	-6	-12
G	-24	-13	-2	9	3	5	-1
T	-30	-19	-8	3	14	8	2
T	-36	-25	-14	-3	8	10	4
A	-42	-31	-20	-9	2	4	15

Global vs. Local

```
TGGTAGATTCCCACGAGATCTACCGAGTATGAGTAGGGGGACGTTCGCTCGG
GCCTCTAACACACTGCACGAGATCAACCGAGATATGAGTAATACAGCGGTACGGG
```

Global Alignment Score: 60

```
---TGGTAGATTC-C--CACGAGATCTACCGAG-TATGAGTAGGGGGAC-GTTCGCT-C-GG
   |  || |  | |  ||||||||| ||||||| ||||||||     || |  || | | ||
GCCT-CTA-ACACACTGCACGAGATCAACCGAGATATGAGTA---ATACAG--CGGTACGGG
```

Local Alignment Score: 105

```
CACGAGATCTACCGAG-TATGAGTA
||||||||| ||||||| ||||||||
CACGAGATCAACCGAGATATGAGTA
```


Local Alignment

AGATCAC
CGACAG

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

Local Alignment

AGATCAC
CGACAG

		C	G	A	C	A	G
	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0
G	0	0	5	0	1	0	10
A	0	0	0	10	4	6	4
T	0	0	0	4	6	0	2
C	0	5	0	0	9	3	0
A	0	0	0	5	3	14	8
C	0	5	0	0	10	8	10

Local Alignment

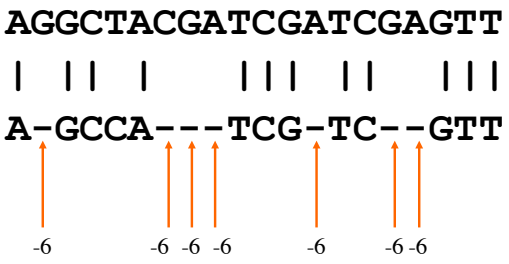
AGATCAC
CGACAG

		C	G	A	C	A	G
	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0
G	0	0	5	0	1	0	10
A	0	0	0	10	4	6	4
T	0	0	0	4	6	0	2
C	0	5	0	0	9	3	0
A	0	0	0	5	3	14	8
C	0	5	0	0	10	8	10

GATCA
|| ||
GA-CA

Linear Gap Penalty

With linear gap scoring, every gap has the same score

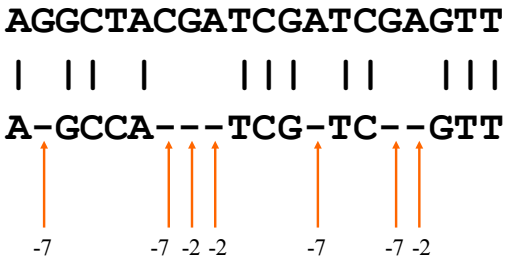


If the match score is +5, the mismatch score is -4, and the linear gap score is -6, then the alignment score is 14.

Affine Gap Penalty

With affine gaps, gap scores are determined from two scores:

- alpha, α, is the gap opening score
- beta, β, is the gap extension score



If the match score is +5, the mismatch score is -4, and the affine gap scores are α = -7 and β = -2, then the alignment score is 22.

Not All Nucleotides Are Created Equal!

Match score: 5

Mismatch score: -4

	A	C	G	T
A	5	-4	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4
T	-4	-4	-4	5

	A	C	G	T
A	5	-4	-1	-4
C	-4	5	-4	-1
G	-1	-4	5	-4
T	-4	-1	-4	5

Amino Acids Work Too!!!

MLVIGSL
MHWNLV

		M	H	W	N	L	V
M							
L							
V							
I							
G							
S							
L							

BLOSUM62 Matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Protein vs. Nucleotide

- Protein searches tend to find more distant similarities
- Why?
  - 4 vs. 20 letter alphabet
  - Different nucleotide sequences can code for the exact same sequence of amino acids
  - Better protein substitution matrices
  - Protein databanks are smaller