



Pairwise Sequence Alignment



Today's Goal

> DNA Sequence 1

```
ACTGCGATTGACGTACGATCATCGTACGATCATGCTGAGCTATCATCATCGTACTGA  
TCGTAGACTACGTAGCTAGCATGCAGTCTGATGACGTCATGCTGACGTAGCATGC
```

> DNA Sequence 2

```
GACTAGCAGCGAGAGATCTCTCGAGTATGCGAGAGCTGATGCATCTACGTATGCAGTCGT  
GCTAATGCGAGCGTATACGCGGCATGTAGAGACTTCTTAGTAC
```

How related are two sequences?

> Protein Sequence 1

```
KGLAHDGHNADFLKAMGGPIAFPIDADPFIDFKLHMNI
```

> Protein Sequence 2

```
LHASDGFKHSADFHNAIFDPAFLKADFPIMADSFN
```




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

CGTTACA--TG
| | | |
T-GT-CACGT- C-GTT-ACATG -TGTCACGT-
| | | |
-TG-TCACGT- CG-TTACATG
| |
TGTC-A-CGT

CGTTACATG-
| | | |
TGT--CACGT

CGTTACATG
TGTCACGT

-CGTTAC-ATG C-----GTTACATG CGTTACATG
| | | | | | | | | | | | | |
TGTCACGT----- CGTT-ACATG- TGTCACGT-
| | | | | | | | | | | | | |
TG-TCAC--GT CGTTACATG-
CGT-TACATG- CGTTACATG
| | | | | | | |
T-G-T-CACGT T-GTCACGT --TGTCACGT



Pairwise Sequence Alignment

Pairwise Alignment Problem:

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



How many different alignments?

Diagram illustrating various pairwise sequence alignments between the sequences **CGTTACATG** and **TGTCACGT**. The sequences are shown in a grid-like format with vertical lines indicating matches between corresponding nucleotides. A central box highlights the sequences **CGTTACATG** and **TGTCACGT**.

```

      CGTTACA--TG          CGTT-ACATG
      | | | |           | | | |
C-G-T-TACATG          C-GTT-ACATG          -TGTCACGT-
      | | | |           | | | |
TG-T-C-AC-GT          -TG-TCACGT-          CG-TTACATG
      | | | |           | | | |
      CGTTACATG-          -CGTTACA-TG
      | | | |           | | | |
      TGT--CACGT          T-G-T-CACGT

      -CGTTAC-ATG          C-----GTTACATG          CGTTACATG
      | | | |           | | | |           | | | |
      T-GTCA-C-GT          TGTCACGT-----          TGTCACGT-

      CGT-TACATG-          CGTTACATG          CGTTACATG-
      | | | |           | | | |           | | | |
      T-G-T-CACGT          T-GTCACGT          --TGTCACGT
  
```



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

AGCGTTA
ACGTGA

AGCGTT + A
ACGTGA -



The Elegance of Alignment

AGCGTTA
ACGTGA



AGCGTT	+	A	AGCGTTA	+	-
ACGTGA	-		ACGTG	-	A



The Elegance of Alignment

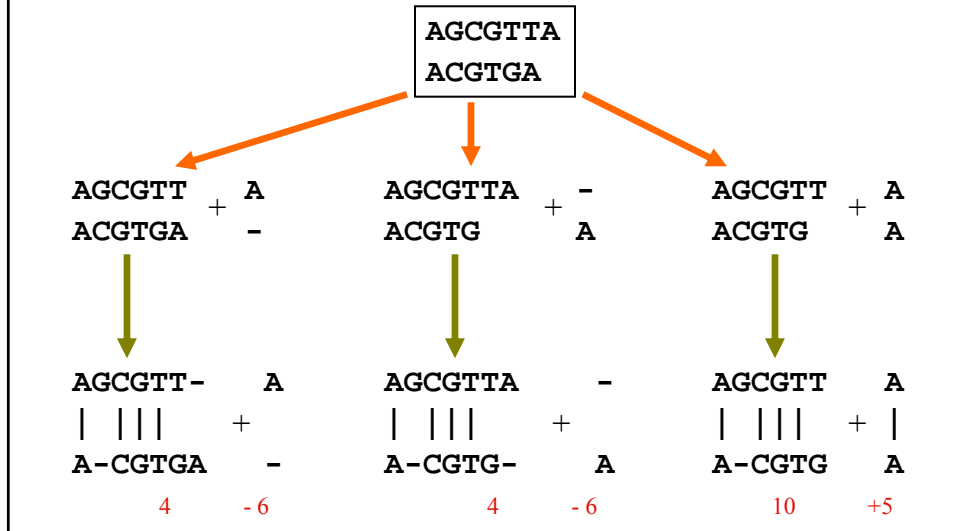
AGCGTTA
ACGTGA



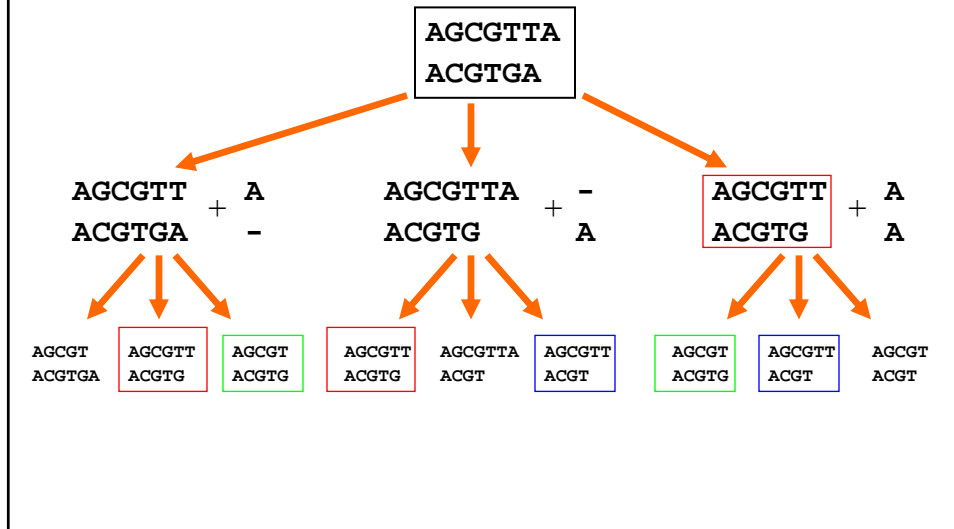
AGCGTT	+	A	AGCGTTA	+	-	AGCGTT	+	A
ACGTGA	-		ACGTG	-	A	ACGTG	-	A



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

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



Score Table

AGCGTTA
ACGTGA

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

AGCGT
ACG



Score Table

AGCGTTA
ACGTGA

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

A
ACGTG



How is each block 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)

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

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



The Elegance of Alignment

AGCGTTA
ACGTGA

AGCGTT + A
ACGTGA -

AGCGTTA + -
ACGTG A

AGCGTT + A
ACGTG A

AGCGTT- A
| | | | +
A-CGTGA -

AGCGTTA -
| | | | +
A-CGTG- A

AGCGTT A
| | | | + |
A-CGTG A



Alignment Score Table

AGCGTTA
ACGTGA

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



Alignment Score Table

AGCGTTA
ACGTGA

	A	G	C	G	T	T	A
0	-6	-12	-18	-24	-30	-36	-42
A	-6	5					
C	-12						
G	-18						
T	-24						
G	-30						
A	-36						



Alignment Score Table

AGCGTTA
ACGTGA

	A	G	C	G	T	T	A
0	-6	-12	-18	-24	-30	-36	-42
A	-6	5	-1				
C	-12						
G	-18						
T	-24						
G	-30						
A	-36						



How do we re-create the alignment?

AGCGTTA
ACGTGA

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



How do we re-create the alignment?

AGCGTTA
ACGTGA

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

A
|
A



How do we re-create the alignment?

AGCGTTA
ACGTGA

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

TA
|
GA



How do we re-create the alignment?

AGCGTTA
ACGTGA

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

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!



Local Alignment

AGATCAC
CGACAG

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



Local Alignment

AGATCAC
CGACAG

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



Local Alignment

AGATCAC
CGACAG

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



Local Alignment

AGATCAC
CGACAG

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



Local Alignment

AGATCAC
CGACAG

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

A
|
A



Local Alignment

AGATCAC
CGACAG

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

CA
||
CA



Local Alignment

AGATCAC
CGACAG

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

GATCA
|| ||
GA-CA



Linear Gap Penalty

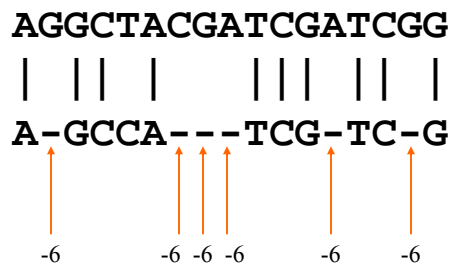
With linear gap scoring, every gap has the same score

AGGCTACGATCGATCGG
| | | | | | | |
A-GCCA---TCG-TC-G
↑ ↑ ↑ ↑ ↑ ↑
c c c c c c



Linear Gap Penalty

With linear gap scoring, every gap has the same score



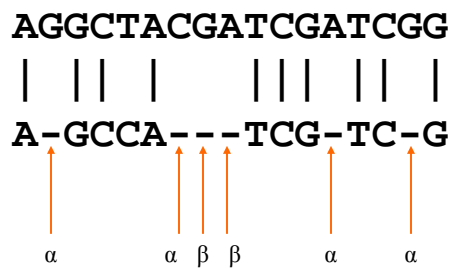
If the match score is +5, the mismatch score is -4, and the linear gap score is -6, then the alignment score is 10.

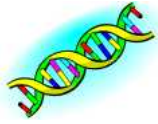


Affine Gap Penalty

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

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

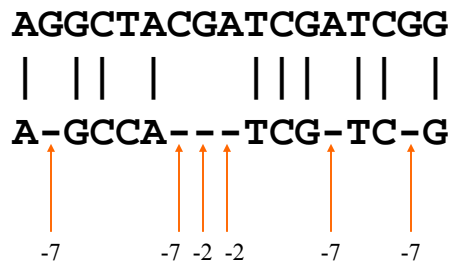




Affine Gap Penalty

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 $\alpha = -7$ and $\beta = -2$, then the alignment score is 14.



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



20 Amino Acids

Alanine	A	Leucine	L
Arginine	R	Lysine	K
Asparagine	N	Methionine	M
Aspartic acid	D	Phenylalanine	F
Cysteine	C	Proline	P
Glutamine	Q	Serine	S
Glutamic acid	E	Threonine	T
Glycine	G	Tryptophan	W
Histidine	H	Tyrosine	Y
Isoleucine	I	Valine	V



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