

# BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine

## Lecture 03 – Sequence and Dynamic Programming (14/09/2022)

Lecture Outline:

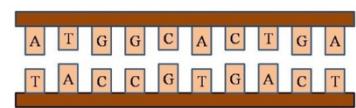
1. Sequence Data
2. Sequence Comparison and Alignment Score
3. Dynamic Programming

### 1. Sequence Data

#### 1.1 What are the sequence data?

##### ➤ DNA Sequence:

- ✓ Composed of A, T, C, G bases.
- ✓ Complementary double strand.
- ✓ Approximately 3 billion of base pairs.



Virus	RNA Sequence	MFE
BITV1	UCGACUACCCUCGGGUUCCUCUGCUCCCCAGUAACGGCGAACACGAUAGAUG	-13.6
CeRV1	CUUGGUAGAUACUUCUCCUGGUUCCUCUGCUCCCCAGUAUAAGCGAAACCAAGUUCAGA	-14.3
CmRV	GUUGCAAGGUGGUAGCUCUACUCGCGGCCGAACGUACUGGUACGGGUACAGUAVUG	-16.0
EIV1	GAUCUCCUCCAUAGCCUGAGCGUGUACCGUGUAGAAGCCCGUCCCCAGCAAGAUUG	-21.4
GsRV-L1	CCUCCCGCCAACCUUCGCCGAGCGCGAGCGUACGUAGCGGCCUCAAGAUUG	-16.5
HmTV-17	CAUGAGGCUGAGGCCALGCGAAGCAAGCGGCCCAAGCGUCAAGAUAGG	-20.2
Hv19MS	ACUUCACUACAGCCACCCCCGAGCUACCGAGUAGUAGCGAGGAGCAAGAUUG	-17.0
MoV1	GAUCAGGGCGAACCCCCGAGCCACGCGAGCGUAGUAGCGAGGAGCAAGAUUG	-13.8
MoV2	CCCGAAAGGCAAACACAGGAGCGGCCGAGCGUACCGAGUAGUAGCGAGGAGCAAGAUUG	-23.4
SrRV1	ACCCUGCCCCCGUCGAGACCGGCGGCCAACGUACGUAGCGGCCAACAGAUAGG	-16.2
SrRV2	CCCCUGCCUCGAAAGCGCCAGCGGAAAGUACCCUCGGCGCCGCGAACAGAUAGG	-21.8

##### ➤ RNA Sequence:

- ✓ Composed of A, U, C, G bases.

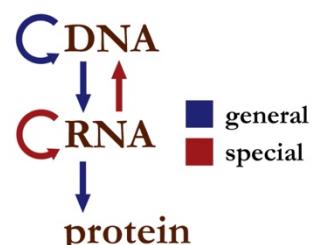
##### ➤ Protein Sequence:

- ✓ Usually composed of 20 amino acids.
- ✓ Allows multiple sequence alignment.

CAA37898.1	-----HPTLGRGFTTE--EQEALVVRWSKAMMWNAGEILGLIEFLKFLIFEIAPSQ	47
PF6871.2	-----PPTFPEESR-----VTPV-----VTPV-----VTPV-----VTPV-----	40
CAA77743.1	MHSSTVILATVLFVIAKTRKELCWSLSEANVQ-----QDQDQDQDQDQDQDQDQD	59
AAA29796.1	MHSSTVILATVLFVIAKTRKELCWSLSEANVQ-----TSKEAKQDQDLYHMEHEYPMH	59
CAA37898.1	-----HPTLGRGFTTE--EQEALVVRWSKAMMWNAGEILGLIEFLKFLIFEIAPSQ	47
PF6871.2	-----PPTFPEESR-----VTPV-----VTPV-----VTPV-----VTPV-----	40
CAA77743.1	KYPERHENY-YTFADQVDPDFFTIKQQNILL-ACHVLCATY-DDR-----ETFDAYGEMLA	112
AAA29796.1	KYPERHENY-YTFADQVDPDFFTIKQQNILL-ACHVLCATY-DDR-----ETFDAYGEMLA	112
CAA37898.1	KHGVADE-----ENPVTVKALLETTKRAVEETVPEENANGKAYDKLUVAAIKLEKQ	158
PF6871.2	IISLLMCIDLHVDPENRLLGIVLVICLAHIFHGEFTTIPVQAQYQVVAVGANALAH-----	145
CAA77743.1	RHE---EDHVKIPNDVWHFWERFTIRFLG---SRTTLDPEPKHAWQEIIGREFSHEISHRRH	168
AAA29796.1	RHE---EDHVKIPNDVWHFWERFTIRFLG---SRTTLDPEPKHAWQEIIGREFSHEISHRRH	168

#### 1.2 Why do we study sequence data?

Sequence data is central dogma. Genetic information is hidden in DNA sequences.

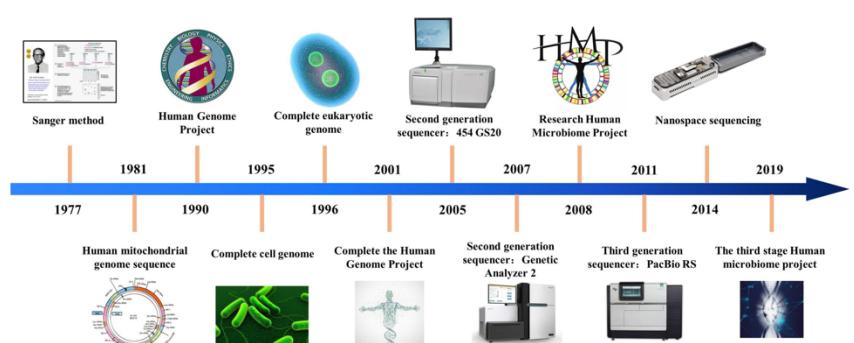


We can understand phenotype with genotype and the environment, in which genotype is believed to be determined by the sequences.

#### 1.3 How do we get the sequences?

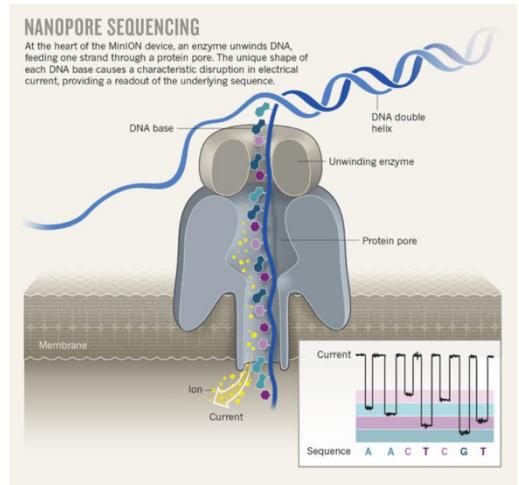
DNA and RNA sequencing are still under active development.

Scientists strive to obtain long reads.



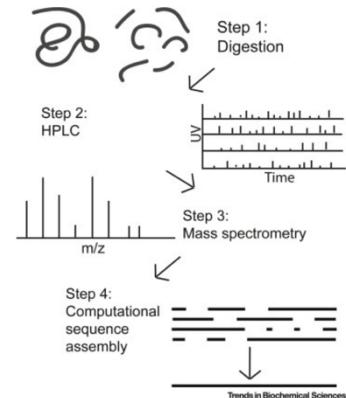
## ➤ Nanopore Sequencing:

- ✓ One of the most advanced methods.
- ✓ A, T, C, G bases have different electrical current.
- ✓ Sequencing by detecting the change in current.
- ✓ Due to noisy signals, error rate is relatively high.
- ✓ Able to obtain very long sequences.



## ➤ Protein Sequencing:

- ✓ Based on mass spectrometry (MS).
- ✓ Break the long sequence into short pieces.
- ✓ Determine the weight of each piece by MS.
- ✓ Assemble the short pieces into the raw sequence.



## 1.4 What can we do with sequence data?

### DNA Sequences:

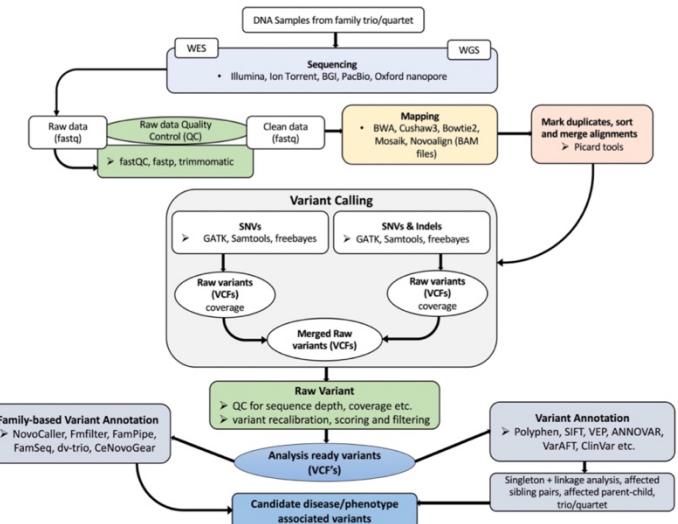
Step 1: Read raw sequence.

Step 2: Perform quality control to delete noises.

Step 3: Map this sequence to reference genome.

Step 4: Variant Calling to check for mutations.

Step 5: Check if the genotypes are related to phenotype associated variants.



### Protein Sequences:

- ✓ Compare two or more sequences by sequence alignment.
- ✓ Similar sequences imply similar structure, which implies similar function.
- ✓ Comparing similar sequences may find out the common ancestor.

CAA37898.1	-----MSTLEGRGFTE--EQEALVVKWSAMKPNAGELGLKFFLKIFEIAPSQ	47
P68871.2	MVHLTPEERSA-----VITALWG-KV-NVDEVGGEEALGRILLVVYFWQT	40
CAA77743.1	MHSSIVLATVLFVAIASASKTRELQMSLAEHKVG-TSKEAKDGEIGDLYKHMFEHYPAWK	59
AAA29796.1	MHSSIVLATVLFVAIASASKTRELQMSLAEHKVG-TSKEAKDGEIGDLYKHMFEHYPAWK	59
CAA37898.1	-----KLFNSFLKDSNVPL--ERNPPLKLKSHAMSVFIMTCESAVOLRKAGKVTVRESSLKQLGASHF	105
P68871.2	REFESFGDLSTPDAVMGNPKVKAHGKKVLG-AFS-----DGL-----AHLDNLKGTFAF	88
CAA77743.1	KYFKHRENVY-TPADVQKDPFFIKQQQNILL-ACHVLCATY-DDR-----ETFDAYVGEIMA	112
AAA29796.1	KYFKHRENVY-TPADVQKDPFFIKQQQNILL-ACHVLCATY-DDR-----ETFDAYVGEIMA	112
CAA37898.1	KHGVAD-----EHFEVTKFALLETIKEAVPETWSPEMKNAWGEAYDKLVAIAKLEMKP	158
P68871.2	LSELHCDKLHVDFENFRLLGNVILCVLAHHFGKEFTPVQAYQKVAGVANALAHK---	145
CAA77743.1	RHE--RDHVVKIPNDVWNHFWEHIFIEFLG--SKTTLDEPTKHAWQKEIGKEFSHEISHHGRH	168
AAA29796.1	RHE--RDHVVKIPNDVWNHFWEHIFIEFLG--SKTTLDEPTKHAWQKEIGKEFSHEISHHGRH	168

## 2. Sequence Comparison and Alignment Score

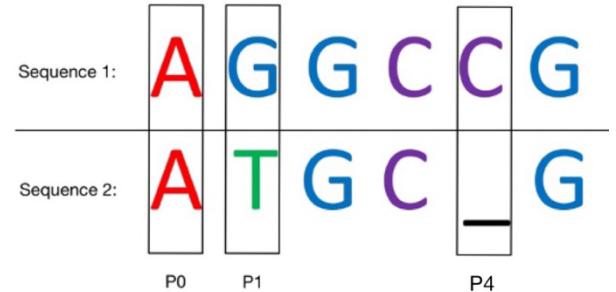
### 2.1 What is sequence alignment?

Sequence alignment is to determine the similarity between two or more sequences. Through pairwise or multiple sequence alignment, we aim to maximize the similarity between them.

### 2.2 What is sequence alignment score?

Consider two sequences:

- ✓ In Position 0 of two sequences, the two “A”s **match**.
- ✓ In Position 1 of two sequences, the “G” in Sequence 1 and the “T” in Sequence 2 **mismatch**.
- ✓ In Position 4 of two sequences, by **insertion or deletion**, a **gap** results.



There are many ways to align two or more sequences. Alignment score is calculated according to the information in scoring matrix. To maximize the similarity between them and find the optimal alignment, the alignment with a relatively higher alignment score will be chosen.

Here lists two possible alignments. The first alignment is chosen as it has a higher alignment score.

Scoring matrix:

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Gap penalty = -10

$$\text{Alignment score 1} = 2 + (-7) + 2 + 2 + (-10) + 2 \\ = -9$$

A G G C C G  
A T G C \_ G

$$\text{Alignment score 2} = 2 + (-7) + 2 + 2 + (-7) + (-10) \\ = -18$$

A G G C C G  
A T G C G \_

### 2.3 How to perform sequence alignment?

Enumerating all the possible alignments is straightforward. However, dynamic programming is used instead as there are too many possible alignments.

### 3. Dynamic Programming

#### 3.1 What is dynamic programming?

- ✓ Break the problem into smaller sub-problems.
- ✓ Solve these sub-problems optimally and recursively.
- ✓ Use these optimal solutions to construct the optimal solution for the original problem.

#### 3.2 How dynamic programming is used in sequence alignment?

There is finite choice for each base, either aligning to another base or aligning to a gap. The alignment score is the sum of the scores for each pair in the alignment.

Consider two sequences:

Goal:

Find the optimal alignment score  $F(ACCG, ACG)$  and the optimal alignment.

Scoring matrix:

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

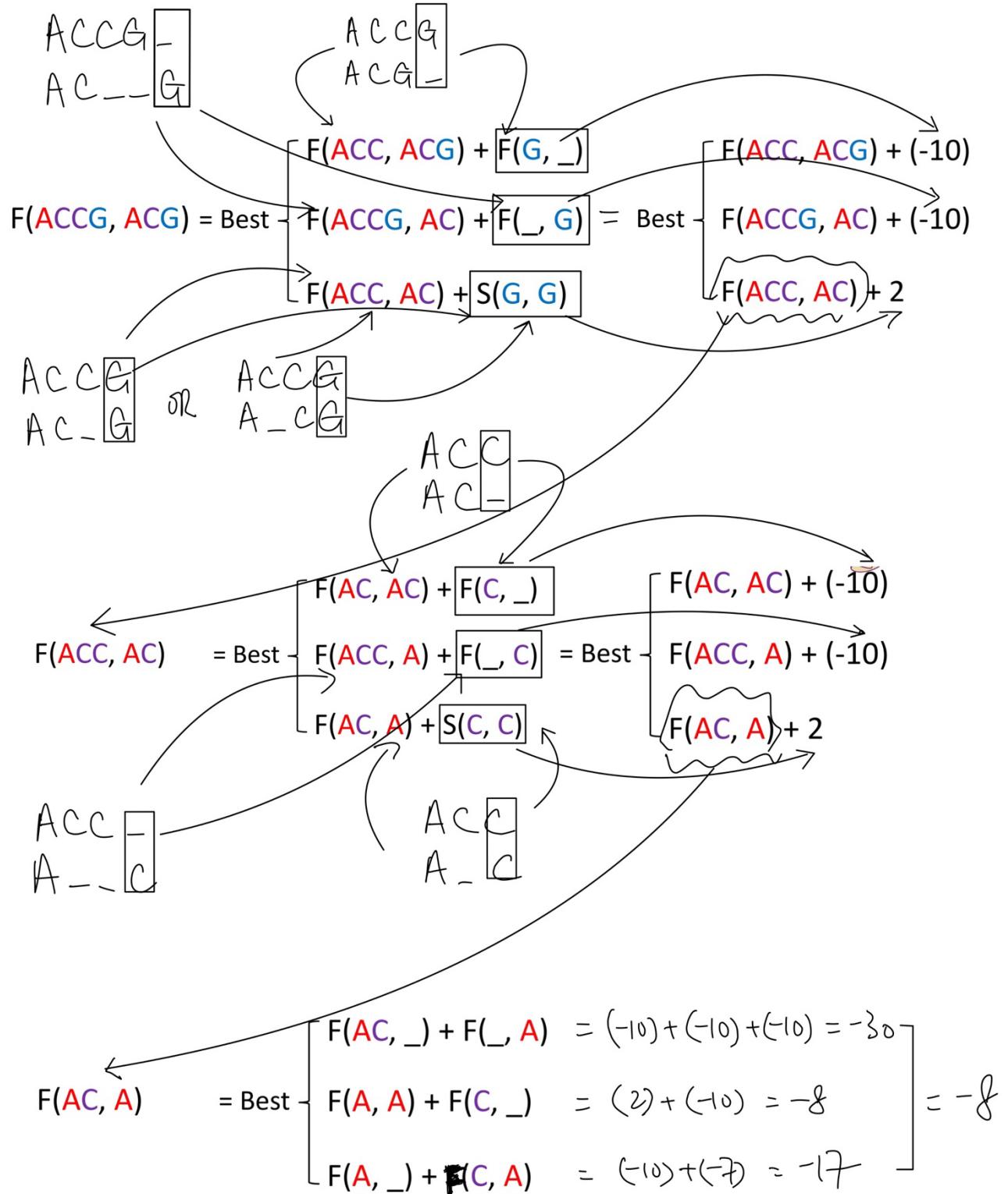
Input sequences:  
ACCG  
ACG

Gap penalty = -10

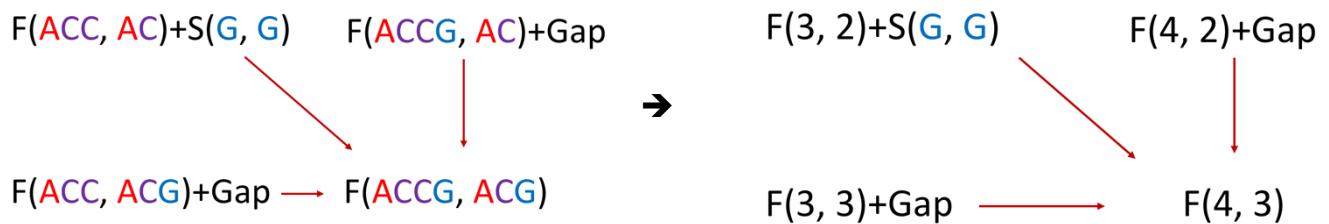
There are three possible ways to align the last pair of the alignment.



Below shows how dynamic programming break the original problem into sub-problems. Noted that the problem size is reduced by one to two bases each time.  $F(XXX, XXX)$  will finally be reduced to  $F(X, X)$  or  $F(X, \_)$  in the scoring matrix, which are the boundary cases.



The original problem can be simplified as follows.



## Table representation

		ACCG				X
		A	C	C	G	
		0	-10	-20	-30	-40
A		-10	2	-8	-18	-28
C		-20	-8	4	F(3, 2) (-6)+2 = -4	F(4, 2) -16
G		-30	-18	-6	F(3, 3) -3	F(4, 3) ??

Scoring matrix:

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

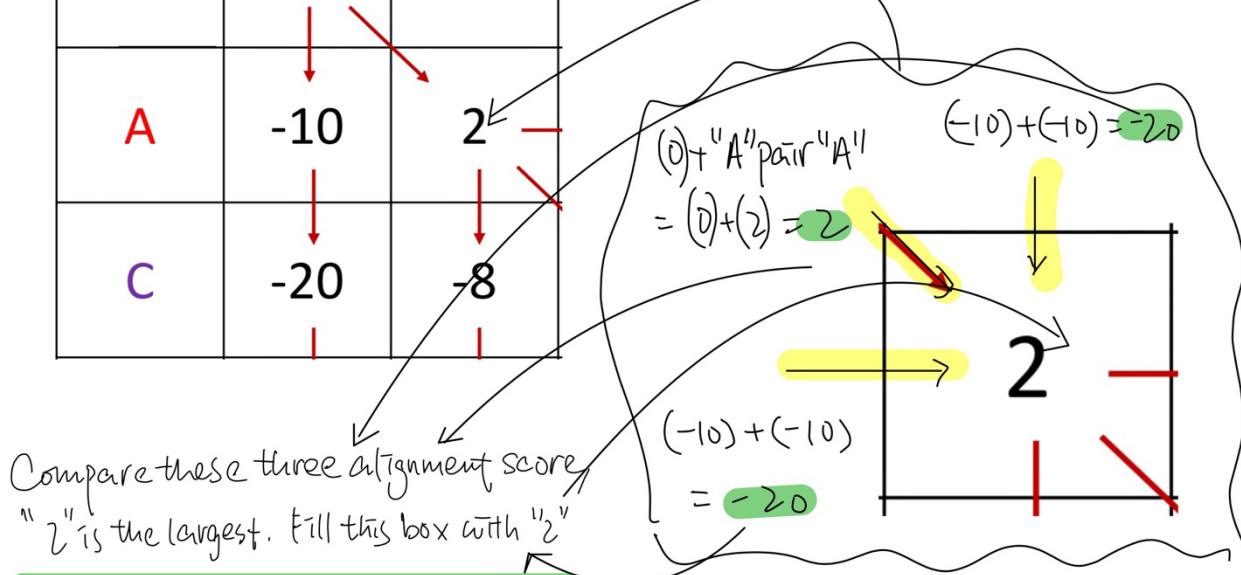
$$(-16) + (-10) = -26$$

$$\text{Gap penalty} = -10$$

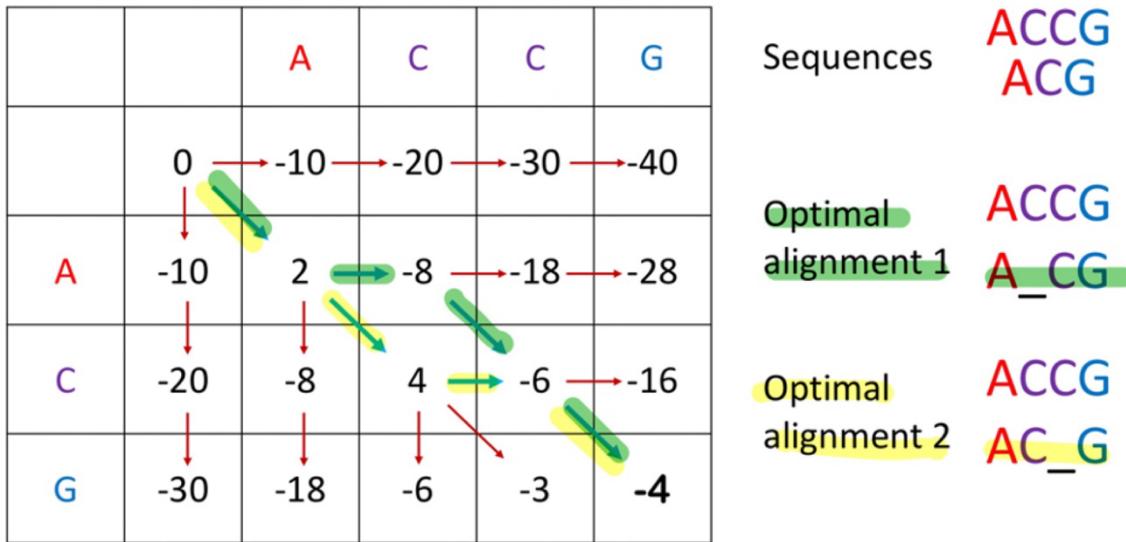
$$(-3) + (-10) = -13$$

		A	
		0	-10
		-10	2
A		-10	2
C		-20	-8

move vertically OR horizontally:  
"adding or gap"  $\Rightarrow -10$



From the table, optimal alignment(s) could be obtained by tracing back. For the input sequences ACCG and ACG, there are two optimal alignments, both with alignment score -4.



### 3.3 What controls the final alignment?

The score matrix.

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Gap penalty = -10

BLOCKS SUbstitution Matrix (BLOSUM)																							
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V				
7	-3	-3	-3	-1	-2	-2	0	-3	-3	-3	-1	-2	-4	-1	2	0	-5	-4	-1	-5	-4	-4	
R	-3	9	-1	-3	1	-1	-4	0	-5	-4	3	-3	-5	-3	-2	-2	-5	-4	-4	-5	-4	-5	
N	-3	-1	9	2	-5	0	-1	-1	1	-6	-6	0	-4	-6	1	0	-7	-4	-5	-7	-4	-5	
D	-3	-3	2	10	-7	-1	2	-3	-2	-7	-7	-2	-6	-6	-3	-1	-2	-8	-6	-6	-5	-2	
C	-1	-6	-5	-7	13	-5	-7	-6	-7	-2	-3	-6	-3	-4	-6	-2	-2	-2	-2	-2	-5	-2	
Q	-2	1	0	-1	-5	9	3	-4	1	-5	-4	2	-1	-5	-3	-1	-1	-4	-3	-4	-5	-2	
E	-2	-1	-1	2	-7	3	8	-4	0	-6	-6	1	-4	-6	-2	-1	-2	-6	-5	-4	-5	-4	
G	0	-4	-1	-3	-6	-4	-4	9	-4	-7	-7	-3	-5	-6	-5	-1	-3	-6	-6	-6	-6	-6	
H	-3	0	1	-2	-7	1	0	-4	12	-6	-5	-1	-4	-2	-4	-2	-3	-4	3	-5	-4	-5	
I	-3	-5	-6	-7	-2	-5	-6	-7	7	2	-5	2	-1	-5	-4	-2	-5	-3	4	-2	-5	-3	
L	-3	-4	-6	-7	-3	-4	-6	-7	-5	2	6	-4	3	0	-5	-4	-3	-4	-2	1	-3	-4	
K	-1	3	0	-2	-6	2	1	-3	-1	-5	-4	8	-3	-5	-2	-1	-1	-6	-4	-4	-4	-4	
M	-2	-3	-4	-6	-3	-1	-4	-5	-4	2	3	-3	9	0	-4	-3	-1	-3	-3	1	-3	-3	
F	-4	-5	-6	-6	-4	-5	-6	-6	-2	-1	0	-5	0	10	-6	-4	-4	0	4	-2	-4	-2	
P	-1	-3	-4	-3	-6	-3	-2	-5	-4	-5	-5	-2	-4	-6	12	-2	-3	-7	-6	-4	-5	-4	
S	2	-2	1	-1	-2	-1	-1	-2	-4	-4	-1	-3	-4	-2	7	2	-6	-3	-3	-3	-3	-3	
T	0	-2	0	-2	-2	-1	-2	-3	-3	-2	-3	-1	-1	-4	-3	2	8	3	0	-5	-3	0	
W	-5	-5	-7	-8	-5	-4	-6	-6	-4	-5	-4	-6	-3	0	-7	-6	-5	16	3	3	-5	-5	
Y	-4	-4	-4	-6	-5	-3	-5	-6	3	-3	-2	-4	-3	4	-6	-3	-3	3	11	-3	-3	-3	
V	-1	-4	-5	-6	-2	-4	-4	-6	-5	4	1	-4	1	-2	-4	-3	0	-5	-3	7	-3	-3	

### Additional Resource:

1. Webserver for Sequence Alignment: [https://www.ebi.ac.uk/Tools/psa/emboss\\_needle/](https://www.ebi.ac.uk/Tools/psa/emboss_needle/)
2. Biopython: <https://biopython.org>
3. Bioinformatics: Sequence and Genome Analysis Chapter 2 & 3 (Textbook)