

BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine

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

Lecture Outline:

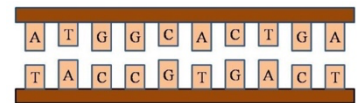
1. Sequence Data
2. Sequence Comparison and Alignment Score
3. Dynamic Programming

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.



➤ RNA Sequence:

- ✓ Composed of A, U, C, G bases.

Virus	RNA Sequence	MFE
BTV1	UCGACUACCCUCCCGUCCUCCUCCUCCUUAACCGGCGGACACGAUAGUGUG	-13.6
CeRV1	CUGGUGAGUUAUCUUCUUCUCCUCCUCCUUAUAAGGCGGAAACCGUAGUGA	-14.3
CmRV	GUGGAGGUGUGAGUACUCCUCCUCCUCCUUAUACCGGUGCCACAGUAGUG	-16.0
ERV1	GAUCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCU	-21.4
GaRV-L1	CCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCU	-16.5
HmTV-17	CAUGAGGUGAAGGACGACAAAGUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCU	-20.2
Hv198S	ACUGCAGCCCGACCCCGCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCU	-17.0
MoV1	GAUCGAGCCCGACCCCGCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCU	-13.8
MoV2	CCGGAAGGCAUAACAACGAGCAAGUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCU	-23.4
SaRV1	ACCUGCCCGCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCU	-16.2
SaRV2	CCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCUCCU	-21.8

➤ Protein Sequence:

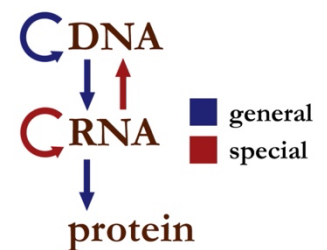
- ✓ Usually composed of 20 amino acids.
- ✓ Allows multiple sequence alignment.

CAA37898.1	-----MSTLEGGGTE--EQEALVVKSMSPNAGELGLFFFLIETAPSAQ	47
P68871.2	-----MHLTPEEKA-----YIALNG-EV-NDEVGGEALGLLVVFTQ	40
CAA7743.1	MISISVLAIVLVAIASANTRELCMSLEHANG-TSEANQDGLDYRHFHFYFAM	59
AAAS9796.1	MISISVLAIVLVAIASANTRELCMSLEHANG-TSEANQDGLDYRHFHFYFAM	59
CAA37898.1	HLFSLADSNVPL--SRNPKLSHMSVFMTCESAVQLRAGKIVTRESLKLGSHP	105
P68871.2	HFESFDLSTPDAMGNPKVFAHGGVGLG-AFS-----DGL----AHLNLAGTFAT	88
CAA7743.1	HYFIRHENY-TPADVQIDFFFIKQGNILL-ACHVLCATY-DR-----ETFDVIGELAM	112
AAAS9796.1	HYFIRHENY-TPADVQIDFFFIKQGNILL-ACHVLCATY-DR-----ETFDVIGELAM	112
CAA37898.1	HRQVAD-----EHFVFKFALLETLEAVPTKSPENANAGRAYDLVLAALTEMLSP	158
P68871.2	LSLILCDLHVDNFENFLLGNVLVLAHVFGEFTFPQQAAYQVAVANALAH---	145
CAA7743.1	FHE--HRIWVLPIDVNHNRHETFLG--SRITLDEPTTHAMQETGFESHEISRHGRH	168
AAAS9796.1	FHE--HRIWVLPIDVNHNRHETFLG--SRITLDEPTTHAMQETGFESHEISRHGRH	168

1.2 Why do we study sequence data?

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

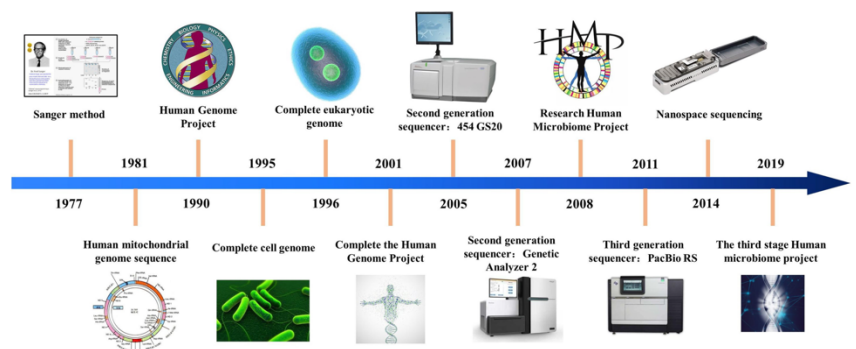
We can understand phenotype with genotype and the environment, in which genotype is believed to be determined by the sequences.



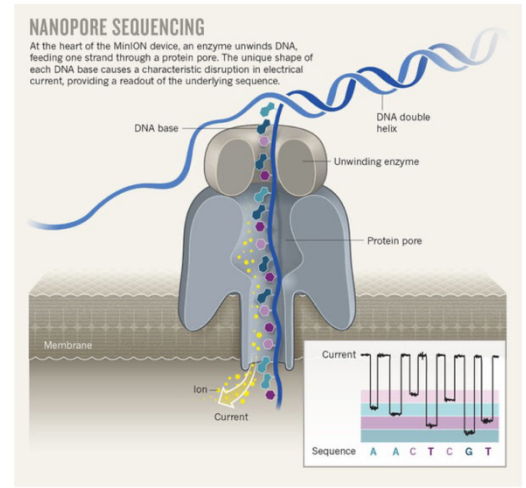
1.3 How do we get the sequences?

DNA and RNA sequencing are still under active development.

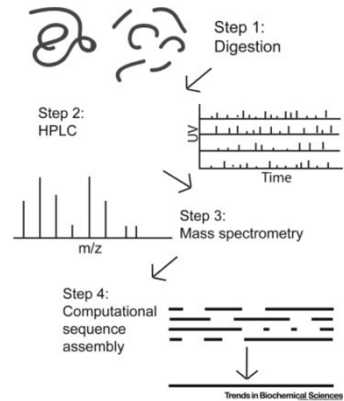
Scientists strive to obtain long reads.



- 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.



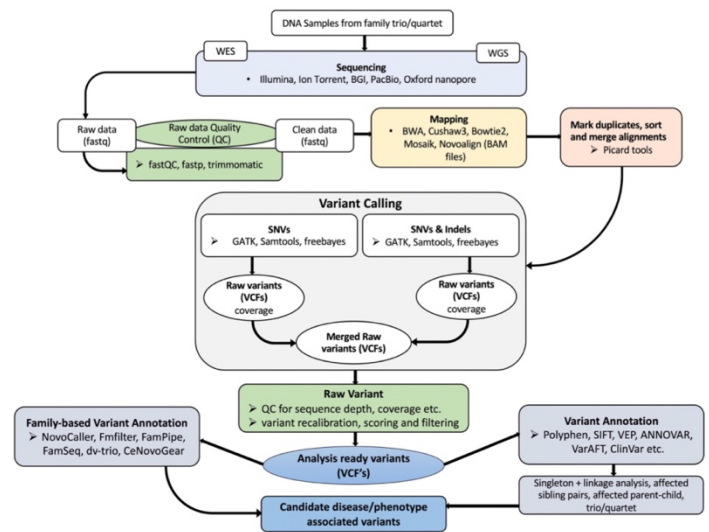
- Protein Sequencing:
 - ✓ Based on mass spectrometry (MS).
 - ✓ Break the long sequence into short pieces.
 - ✓ Determine the weight of each piece by MS.
 - ✓ Assemble the short pieces into the raw sequence.



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 -----MSTLEGRGFTF--EQEALVVKSWSAMKPNAGELGLKFLKIFFIAPSAQ 47
P68871.2 -----MVHLTPEEKSA-----VTALMG-KV-NVDEVGGEALGRLLVVPWQT 40
CAA77743.1 MHSIIVLATVLFVAIASASKTRELCKMKSLEHAKVGT-SRKAQDGDLDLYKHFHYHPAMK 59
AAA29796.1 MHSIIVLATVLFVAIASASKTRELCKMKSLEHAKVGT-SRKAQDGDLDLYKHFHYHPAMK 59
          : : : : : : : : : : : : : : : : : : : : : : : : : : : :
          : * : * : * : * : * : * : * : * : * : * : * : * : * : * : * :

CAA37898.1 KLFSLKDSNVPL--ERNPKLKHSHMSVFLMTCESAVQLRKAGKVTVRESLKKLGASHF 105
P68871.2 RFESFGDLSTPDVVMGNPKVKAHGKVLG-AFS-----DGL---AHLNLRKGFAT 88
CAA77743.1 KYFKHRENY-TPADVQKDPFFIKQGQNIL-ACHVLCATY-DDR---ETFDAYVGLMA 112
AAA29796.1 KYFKHRENY-TPADVQKDPFFIKQGQNIL-ACHVLCATY-DDR---ETFDAYVGLMA 112
          : * : * : * : * : * : * : * : * : * : * : * : * : * : * : * :

CAA37898.1 KHGVAD-----EHFVTKFALLETIKEAVFETWSPMKNAWGEAYDKLVAAIKLEMKP 158
P68871.2 LSELHCDKLVDPENFRLLGNVLCVLAHFGKEFTPPVQAAYQKVVAGVANALAHK--- 145
CAA77743.1 RHE--RDHVKIFNDVWNHFWHFIEFLG--SKTTLDEPTKHAWQEIIGKFSHEISHHGRH 168
AAA29796.1 RHE--RDHVKIFNDVWNHFWHFIEFLG--SKTTLDEPTKHAWQEIIGKFSHEISHHGRH 168
          : : : : : : : : : : : : : : : : : : : : : : : : : : : :
  
```

2. Sequence Comparison and Alignment Score

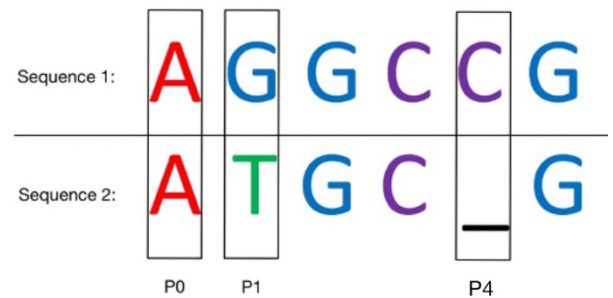
2.1 What is sequence alignment?

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 $F(\text{ACCG}, \text{ACG})$ and the optimal alignment.

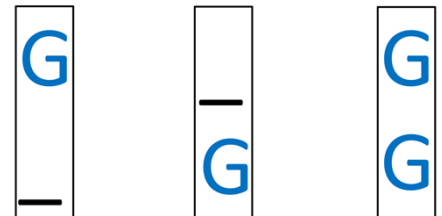
Scoring matrix:

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

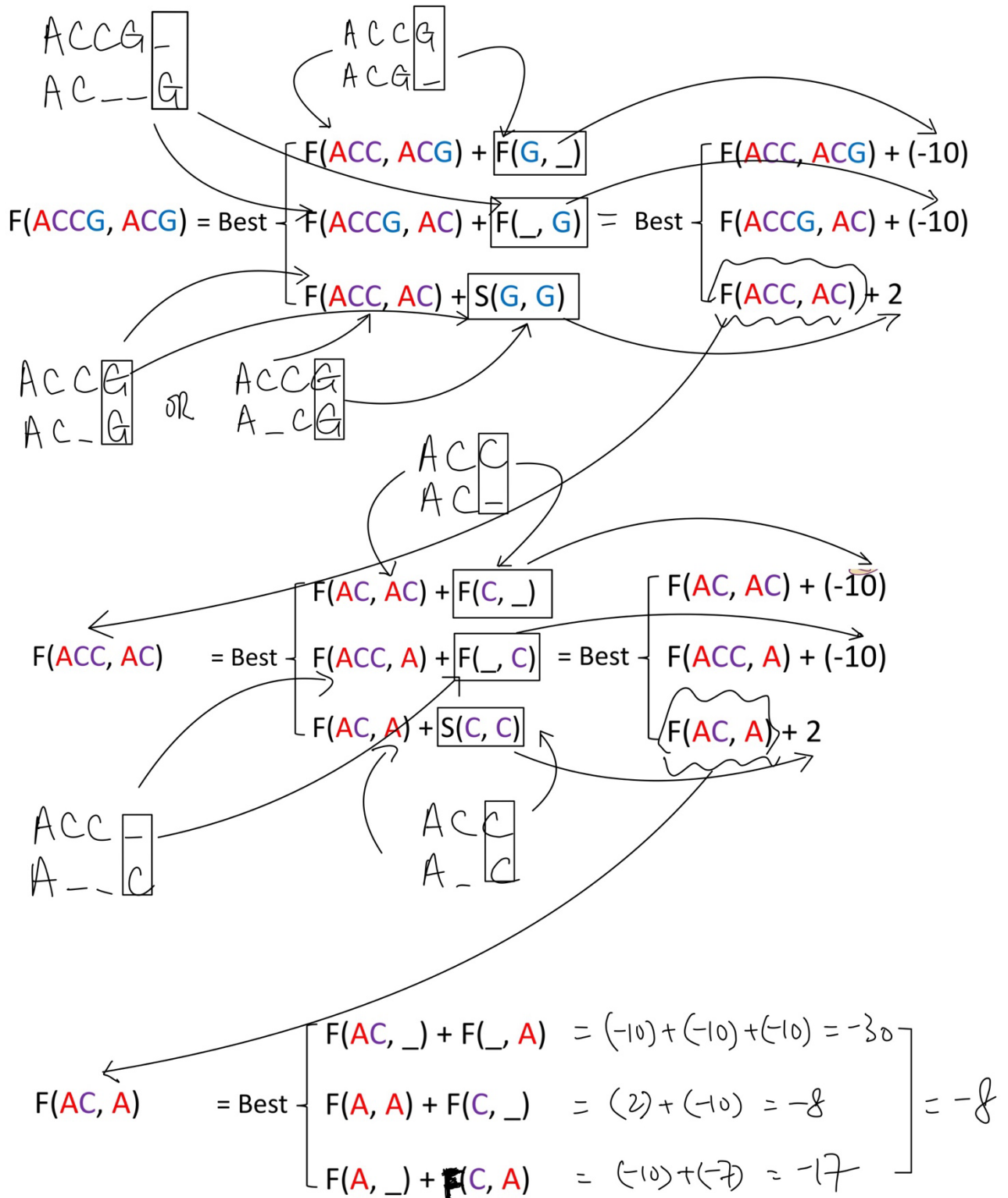
Input sequences: **ACCG**
ACG

Gap penalty = -10

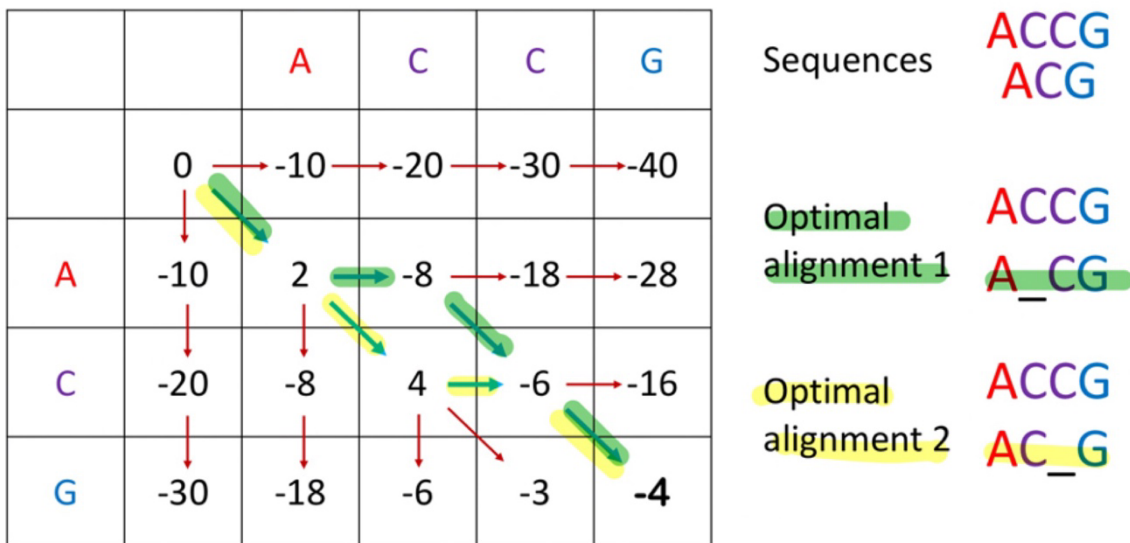
There are three possible ways to align the last pair of the alignment.



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. $F(\text{XXX}, \text{XXX})$ will finally be reduced to $F(\text{X}, \text{X})$ or $F(\text{X}, _)$ in the scoring matrix, which are the boundary cases.



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

Additional Resource:

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