BMEG3105: Data Ana for Personalized Geno

2025-2026 Semester 1

Lecture--4 Assembly & Mapping

Lecturer: Professor Yu Li

1. **Review of DP (Dynamic Programming)**

   A. main goal: find the DNA base pair with the minimum value of alignment score.

   E g: F (ACCG, ACG)

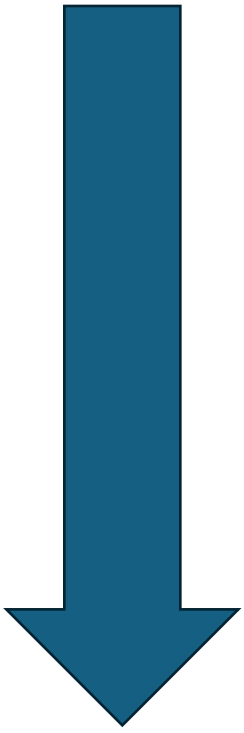
   Best= 1. (ACC, ACG) +F (G, _)

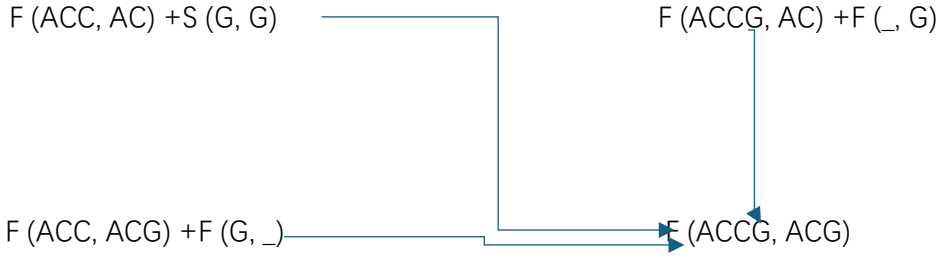
   2.F (ACCG, AC) +F (_, G)

   3.F (ACC, AC) +S (G, G)

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

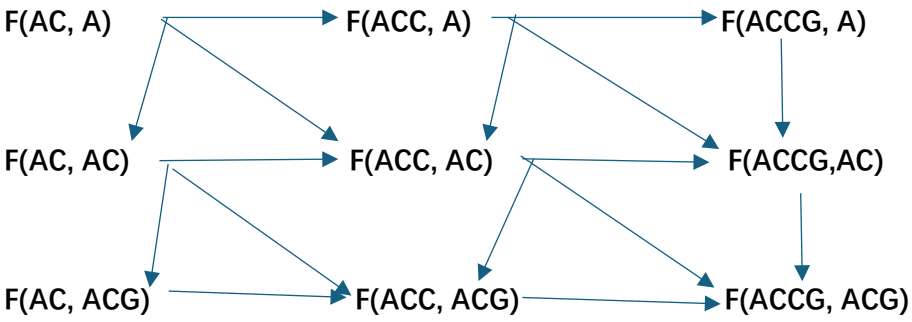
Scoring matrix (gap -10)





B. Construct a DP Table.

1. Break it into smaller sub-base.



2. Construct the table according to the sub-base and scoring matrix.

		A	C	C	G
	0	-10 (A _) =-10(GAP)	-20	-30	-40
A	-10	2	-8	-18	-28
C	-20	-8 (A_ AC) =2-10 =-8	4	-6 (ACC AC_) =2+2-10 =-6	-16
G	-30	-18	-6	-3	-4

The one we care about

= (ACCG
AC_G)
OR
(ACCG
A_CG)

OPTIMUM SCORE AND ALIGNMENT

=2+2+2-10=-6

C. Summary.

Good: --Can identify sequence similarity

--can find the optimum alignment

--easy way to solve the sequence recursively

--DP table shows a clear sub-sequence and construction path for understanding

Limitation: --The calculation becomes complicated when the length of sequence is large

→ No. of possible alignment=(2n)!/(n!)*2

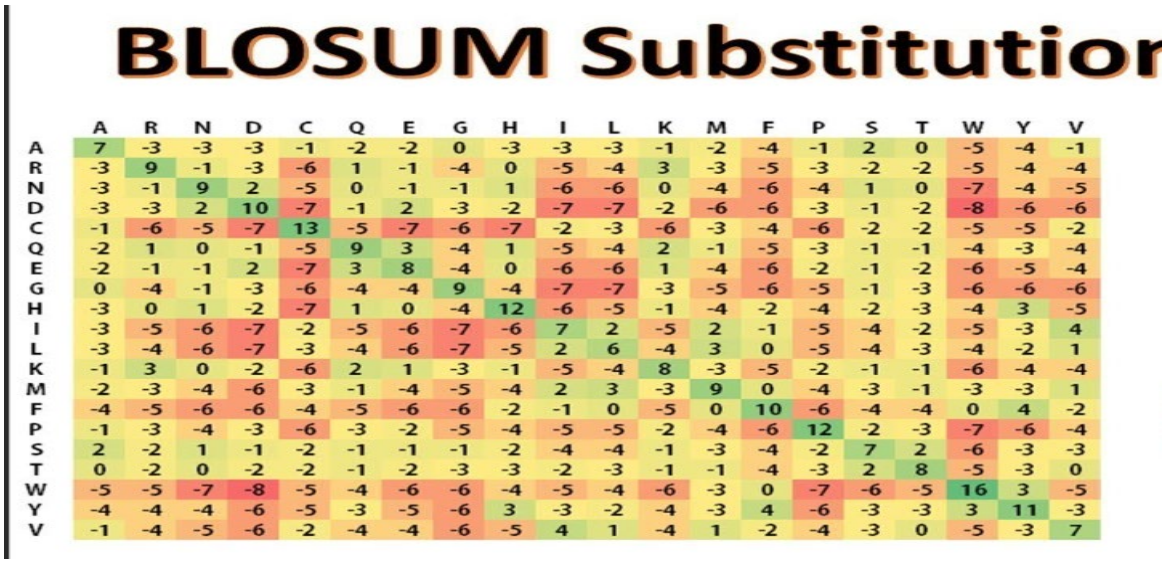
→ Difficult to construct a DP table

--For global alignment, we only care about the bottom right corner one.

--Sometimes, there is mismatch which is because of mutation (sequence change)

--Gap occur from insertion,deletion,gene duplication.(Gene mutation)

--The scoring matrix may be different in different sources. (depends on how we define the similarity between two sequence)



BLOSUM Substitutior

	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

D. Appendix

--Online Website for sequence alignment

https://www.ebi.ac.uk/Tools/psa/emboss_needle/

https://biopython.org

3. Assembly and mapping.

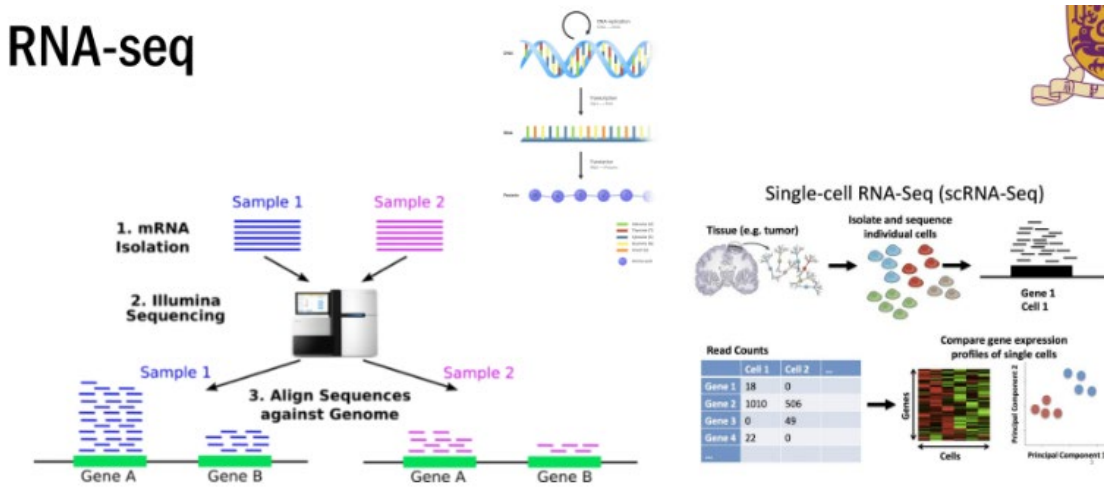
A. Reasons for sequence data

1. Central dogma (DNA—RNA—Protein)
2. DNA sequence contains genetic information
3. Phenotype is related to Genotype and environment, and genotype is determined by the sequence

--genetic variation between all people only accounts for 0.01%

--1% of the gene is responsible for protein coding

--gene expression different make the phenotype different



Sequence data → data matrix

1. Map the short read to the genome

2. Count the no. of reads to construct a gene expression matrix

B. Genome Assembly

Why need it?

→Human genome is very long (~3*10^9 bp), and we need to get the genome from the long reading.

Principle: The 200bp short reads have some overlap regions, we can assembly them by assembly the overlap regions and get the interested genome.

Eg. TAA,AAT,ATG,TGC,GCC,CCA,CAT,ATG,TGG,GGA,GAT,ATG

→TAA

AAT

ATG

TGC……

→TAATG……

C. Mapping

Main goal: find out the number of differences between different base in the genome during sliding each other. And find out the least differences match.

Eg.TAATGCCATGGATG TAA,CCA,GAT,GCC,CCA,ATG

TAATGCCATGGATG ⎫ **first time**

CCA ⎪

2 ⎭

TAATGCCATGGATG ⎫

CCA ⎪ **starts from the second neuclotide**

3 ⎭ **and see the difference**

……

→ **TAATGCCATGGATG**

CCA

233320233233

--Finally, we find that the least difference is in the 6th place which has 0 different.

3. **The Upcoming Topic**

   -- Data exploration and data cleaning