Rebecca Millum - 1155210329
The Chinese University of Hong Kong

BMEG 3105

Assembly & Mapping

Lec-5, September 20, 2023

Lecture's agenda

- More about DP (Dynamic Programming)
- Why and how do we get gene expression matrix
- Introduction to sequence assembly and sequence mapping

DP (Dynamic Programming)

First, when we want to make an alignment, we have two options: Align to another base or align to a gap. From the alignment we can obtain the alignment score, buy using a scoring matrix.

Motivation for Dynamic programming to make alignment:

Enumeration takes too long, results into many alignments, to align two sequences of length n see equation 1

$$\begin{pmatrix} 2n \\ n \end{pmatrix} = \frac{(2n)!}{(n!)^2} \tag{1}$$

Whereas for dynamical programming, filling in the DP table will be described by using the time complexity n^2.

Scoring matrix

The scoring matrix is what defines the best alignment. It is built on how we define the similarity and gap penalty.

Example: if we decide that the gap penalty = 10 and the match pair = 2, that will result in the "best" alignment being the one with the most gaps.

We have an example of a scoring matrix for DNA in figure: 1,

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

Figure 1: Scoring matrix for DNA

We also have Blocks Substitution Matrix (BLOSUM) which is used for protein. See figure 2

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

Figure 2: Soring matrix for protein

How to DP

- First, decide on a scoring matrix.
- Second, fill in in DP table example of an empty DP table in figure 3, ***KEEP THE ARROWS*** to show the path.
- Third, for global alignment find best alignment score in the last cell see figure 3, the aria coloured in in red.
- Four, back track the arrows to obtain the alignment.

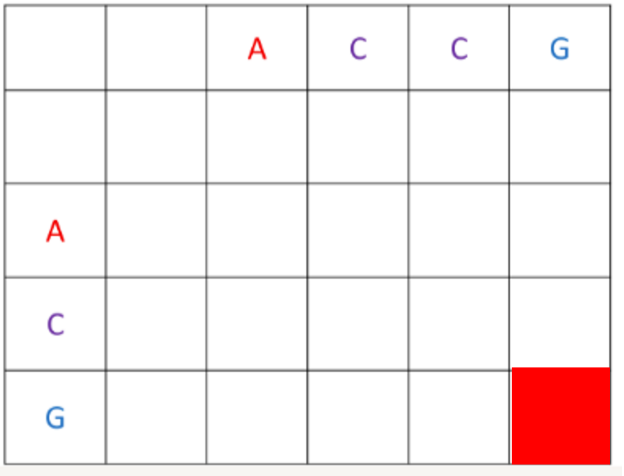
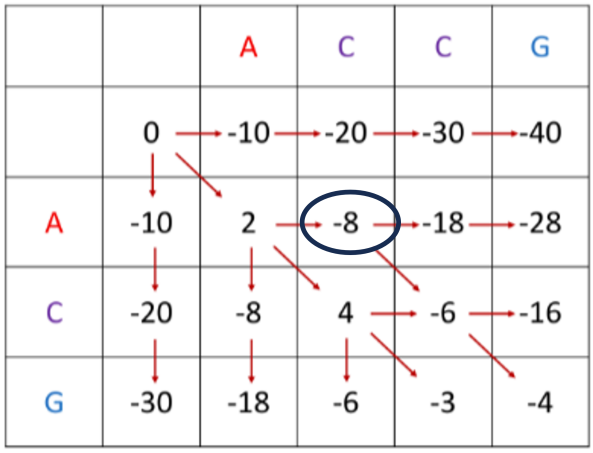


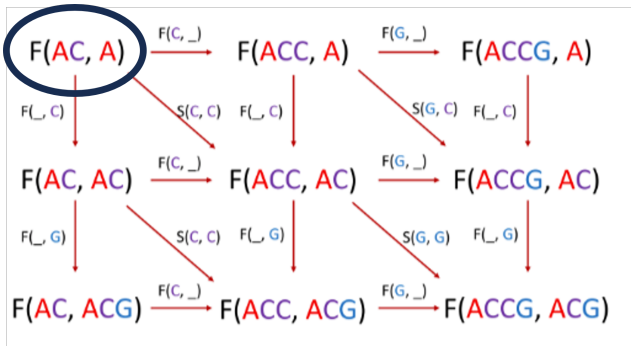
Figure 3: DP table ready to be filled out for sequence: ACCG and ACG

To solve the final problem using DP

Example, given seq1: ACCG and seq2:ACG. In figure 4, on the left we have the DP matrix filled out, and on the right we have a figure illustrating the path. Black circles indicate where on the DP matrix the path starts. The arrows on the right is with a note that has $F()$, within the brakes, it is specified what is the step that is needed to accrue, for the sequence is going to align, commencing from the previous position.



(a) DP table filled out for sequence: ACCG and ACG



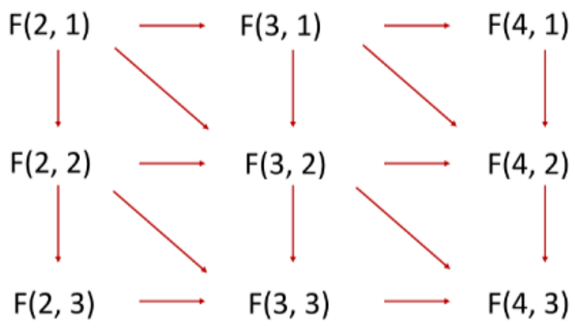
(b) Table to resolve the final problem

Figure 4: DP table and table to resolve the final problem

The function of the DP table is to simplify the reduction process. We can then turn figure 4 into figure 5 by substituting the letters out with numbers corresponding to the index.

		A	C	C	G
	F(0,0)	F(1,0)	F(2,0)	F(3,0)	F(4,0)
A	F(0,1)	F(1,1)	F(2,1)	F(3,1)	F(4,1)
C	F(0,2)	F(1,2)	F(2,2)	F(3,2)	F(4,2)
G	F(0,3)	F(1,3)	F(2,3)	F(3,3)	F(4,3)

(a) DP table using numbers



(b) To resolve the final problem table for the path

Figure 5: To resolve the final problem using the index

The DP table for seq1: ACCG and seq2: ACG will appear as in figure 6.

		A	C	C	G
	0	-10	-20	-30	-40
A	-10	2	-8	-18	-28
C	-20	-8	4	-6	-16
G	-30	-18	-6	-3	**-4**

Figure 6: DP table to Trace back the optimal alignments

We know now that the best aliment score is -4. To find the optimal alignment we can now back track the arrows marked in blue and see that there are two equally good alignments $\begin{matrix}\text{ACCG}\\ \text{A\_CG}\end{matrix}$ and $\begin{matrix}\text{ACCG}\\ \text{AC\_G}\end{matrix}$.

The alignment between seq1: ACCG and seq2: ACG looking like this $\begin{matrix}\text{ACCG\_\_\_}\\ \text{\_\_\_\_ACG}\end{matrix}$, would be found in the DP table going in the path indicated by yellow and black arrow, (note: the numbers down is not corresponding to the alignment score of this alignment) see figure 7

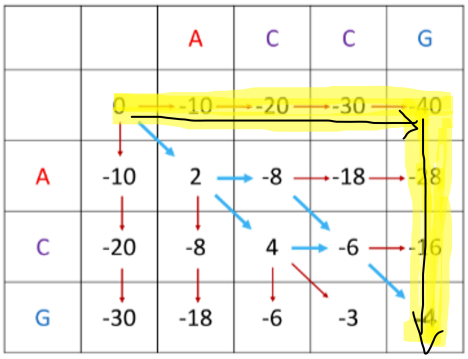


Figure 7: Example of how to find alignments in the DP table

Gene expression matrix

Reasons why the gene expression matrix is relevant

- There is little variation. (0.001% of the genome varies).
- Only 1% of the genome encodes proteins.
- The sequence of the genome does not give the full picture, we also need the expression of the genes.

How do we get the gene expression matrix?

We know that RNA will be present when protein is generated, therefore we know that the active (transcript) genome are the ones we can find the RNA from. When we ***map*** the smaller reads of RNA against the genome we get the expressed part of the genome, see figure 8.

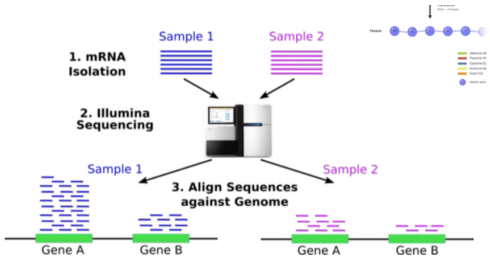


Figure 8: How the active part of the genome is found

We can count the reads which is the indication of expression, and the expression can be compared between the cells and across different people; see figure 9 .

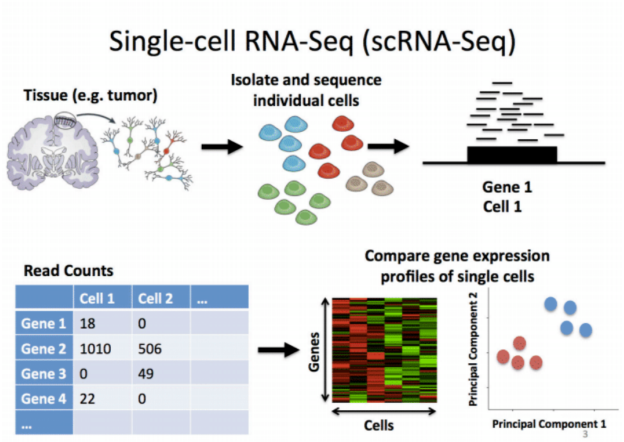


Figure 9: Process of gene expression

Introduction to sequence assembly and sequence mapping

Sequence assembly

The genome is assembled by overlapping regions, think of it as a puzzle, see figure 10.

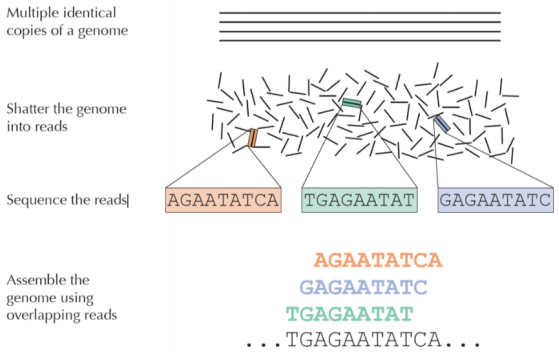


Figure 10: sequence assembly

In figure 11, you find an example of sequence assembly where each read is 3 base pairs long.

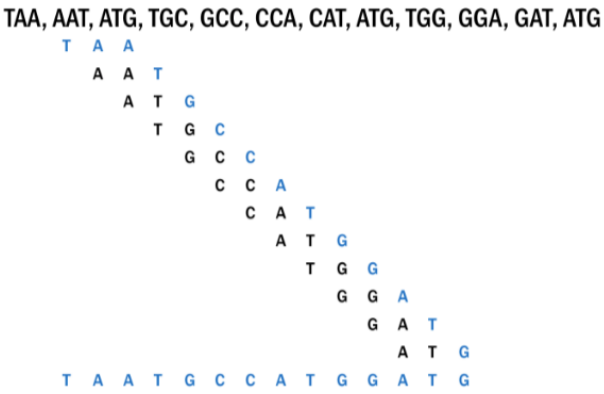


Figure 11: Sequence assembly, example using 3 base pairs

List of things that will complete the genome assembly

- Mutations.
- Conflict FX: AAT and AAA.
- Repeat sequences FX (AAAAAAAAAAAA...AAAA) → hard to determine the length of the repeat.
- Repeat genes (if two sequences are present in more than one place) →problem determining the copy number.

Sequence mapping

Sequence mapping is what we do to get the gene expression count which is used in the expression matrix.
To find where in a genome a read belongs, the read will slide along the genome, and in the process the differences between the read and the genome will be calculated. In figure 12 we have the genome *TAATGCCATGGATG* and we are trying to map the sequences *CCA* to it, where you can see that the read is mapped to the part of the genome where the difference is equal to 0.

Figure 12: Example of mapping a read consisting of 3 base pairs to a sequence

This process will happen with multiple reads, now we have all the reads (TAA, CCA, GAT, GCC, CCA, ATG) and we are interested in the gene expression count of the GCCA part of the genome.

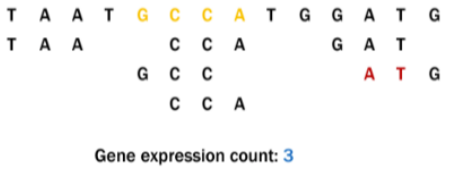


Figure 13: Example of mapping multiple reads consisting of 3 base pairs to a sequence to get the gene expression count

In this case the expression count is 3, because there are 3 reads in the part of the gene we are interested in. Note: when we state the expression count we declare whether or not a read is both overlapping with the region of interest but also covers some of the genome if it counts towards the genome count.

wep

- https://www.ebi.ac.uk/Tools/psa/emboss_needle/ ***(Webserver for sequence alignment)***
- https://biopython.org/ ***(Python tools for computational molecular biology)***
- https://github.com/VGP/vgp-assembly/tree/master/pipeline ***(error-free genome assemblies)***