BMEG3105 Fall 2025

Lecture 4 – Assembly and Mapping

Lecturer: Yu Li(李煜) from CSE SID: 1155212306

Assembly and Mapping

Friday, 12 September 2025

1. Outline of Lecture 4

1.1. Dynamic programming
1.2. Why and how do we get gene expression matrix
1.3. Introduction of sequence assembly and mapping

2. Importance of sequencing

With the data analysis of DNA fragments inside the maternal blood of pregnant woman, a non-invasive DNA-based prenatal testing is able to be performed for Down syndrome test.

3. Recap of Last Lecture

3.1. Scoring matrix and Alignment score

By using a scoring matrix, we were able to find the optimal alignment score ***(sum of the score for each pair in the alignment)*** between two sequencing, determining whether the sequence has the highest similarity.

However, since there is too many possible alignment of sequence(Given the No. possible solution = (2n)!/(n!)^2) we use Dynamic Programming to determine which one is the best.

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

3.2. Dynamic Programming

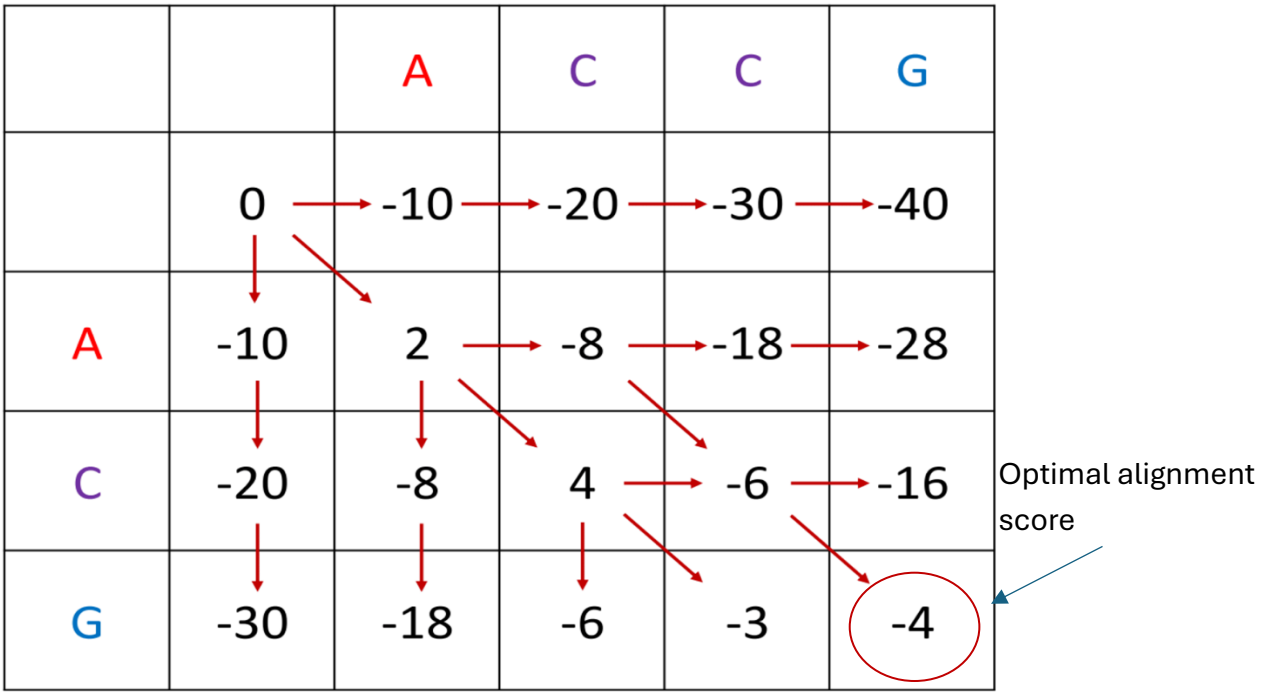
Dynamic Programming can help us simplify the reduction process of the sequence alignment and find the optimal alignment score to this problem.

Step 1: Fill in the table

Step 2: Input the score of different alignment

Step 3: Use arrows to trace the optimal alignment

DP Table:



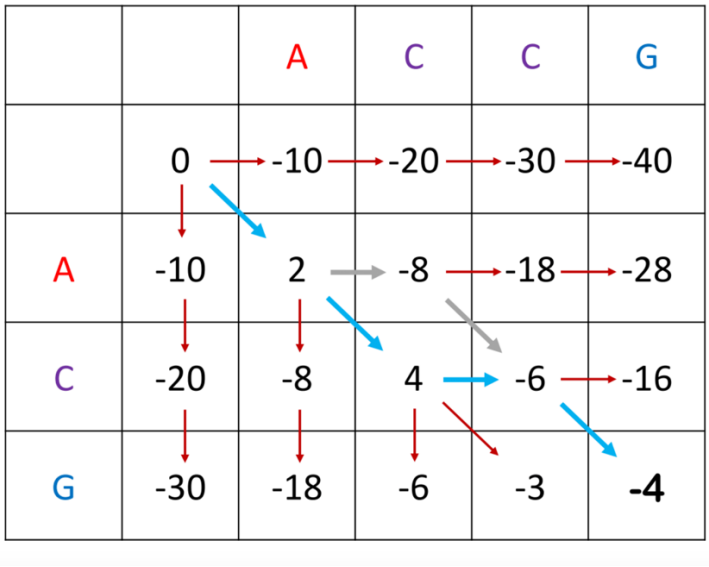
With this table, We can conclude that the best alignment score between ACCG and ACG is -4

Therefore, This method can significantly simplified the process of finding optimal alignment score.

4. More about dynamic programming

DP programming can also store the solution or path of alignment with different sequence***(May not be optimal)*** **.**

We can also trace back the alignment using arrows



(Two different path)

5. Importance of comparing different sequence and getting gene expression matrix

5.1. Why do we compare sequence?

To find the reason behind the difference between two sequence:

Mismatch -> Mutation

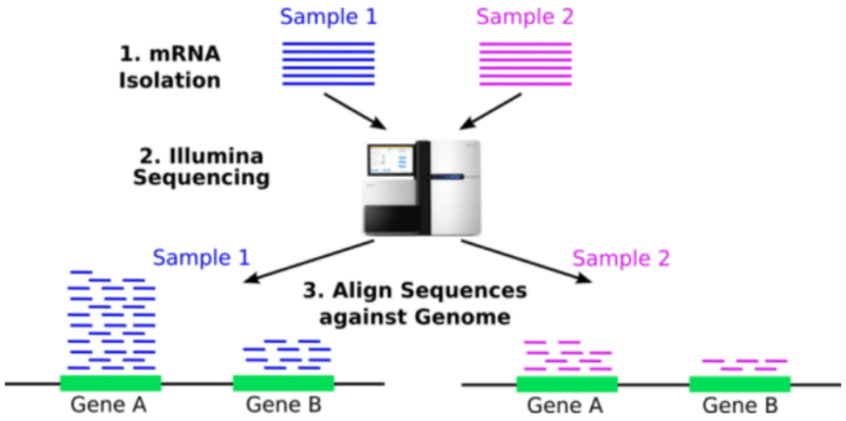
Gap -> Insertion/deletion, gene duplication

5.2. Why sequence data?

Human genome of all Humans are really similar ***(only around 0.0001% are varied)*** and only 1% of genome is used to encode protein. Therefore, everyone has different gene expression even with little variation.

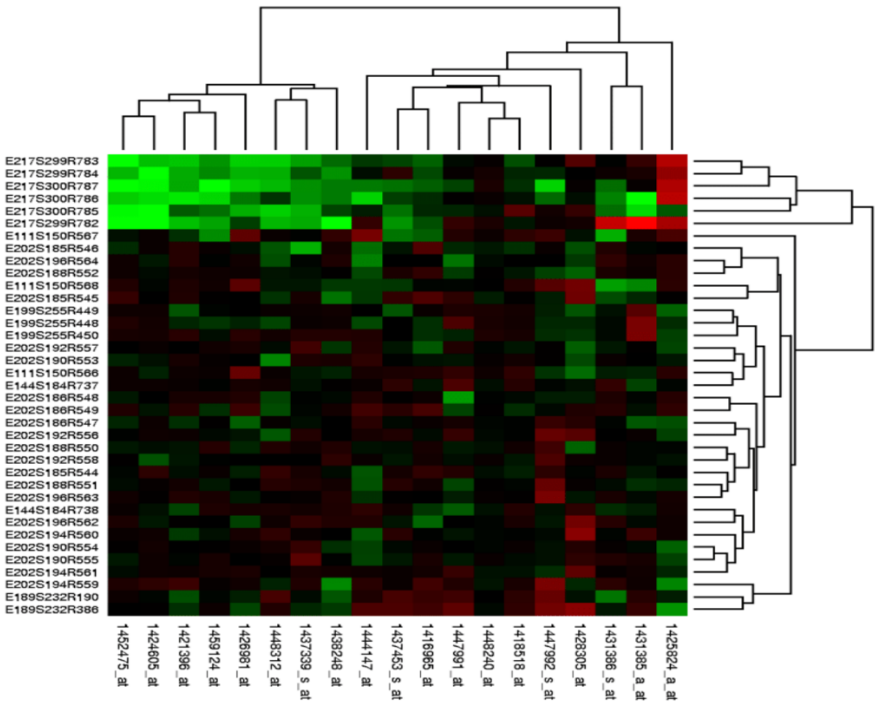
In conclusion, RNA sequencing is performed in order to obtain the data matrix of different individual to analysis the difference of genome using a data matrix. The steps are follows:

Step 1: RNA sequencing



Step 2: Compare the Number of Gene in different sample

Step 3: Compute a gene expression matrix using the result obtained



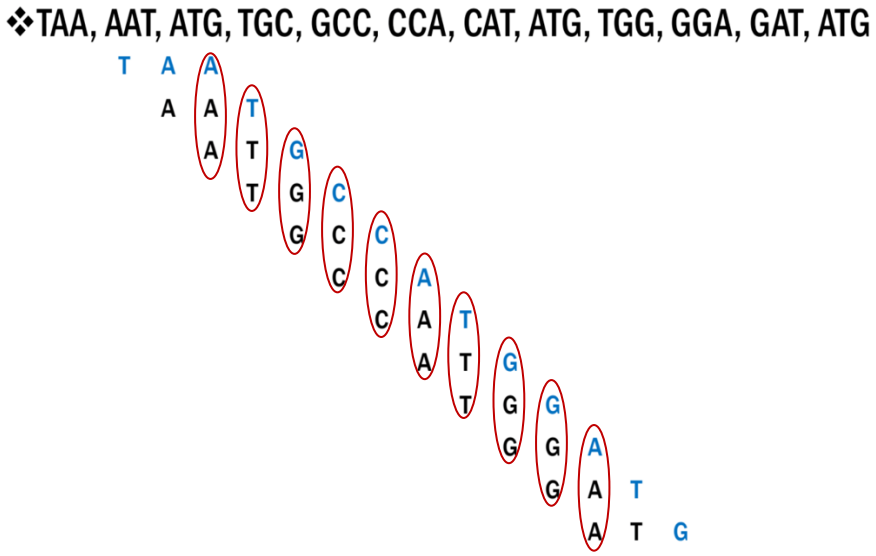
6. Genome Assembly and Mapping

6.1. Genome Assembly

The length of illumina sequencing is around 200bp and the human genome length is around 3 billion bp.

Then, we find the overlapped region between the two short read of genome(200bp) and assemble the genome using the sequenced read.

For example:



When we look at the above graph, we can see the overlapped region***(circled in red)*** when we align the overlapped region of the sequence. Therefore, we can assemble the sequence from the above data as TAATGCCATGGATG.

In conlusion, we can combine the short readings into a more simple and straight forward genome sequence. Enable us to study the structure and function of the genome more easily.

6.2. Genome Mapping

With each obtained reading, We slide it along the genome to calculate the number of difference between the reading and the aligned genome. After finding the best match***(0 difference)*** and it accumulate to 1 gene expression count.

We continue to do the mapping until we find the best match for all short readings.
Example:

Gene expression count: 3

Since we compare the obtained short readings with the normal genome. Mapping help us to analysis the structure of the genome which is useful to identify any genetic diseases.

7. Resource and uncovered topics

- Bioinformatics: Sequence and Genome Analysis---Chapter 2&3
- Time complexity and space complexity analysis
- Local alignment
- Multiple sequence alignment
- Affine gap penalty
- Sequence database search: BLAST
- Bioinformatics Algorithm: Chapter 6 & Chapter 9
- A survey of best practices for RNA-seq data analysis