Scribing: Lecture 4: Assembly & Mapping

Created	@September 17, 2025 2:21 PM
Class	BMEG3105
Instructor	Ng Sui Ip 1155213651

Dynamic Programming:

Using the optimal solutions to construct the optimal solution for the original problem, ie, finding the highest similarity score (the best score) for comparing two genes' alignment.

- Dynamic Programming Matrix

Example:

Assume that we now have two Genes: ACCG and ACG. By using dynamic programming:

Set up a scoring matrix for the pairing of nucleotide base:

	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	-10	-10	-10	-10

Here, the Gap penalty is -10.

Fill in the DP table with ACCG and ACG (The first row and column is filled by cascading the gap penalty), Follow the rules of highest score, where every time the pairing either adding gap with the first gene, adding gap with the second gene, and pairing up the nucleotide base directly, select the highest score in every cell, we then get:

		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

From the table, here are some rules:

1. If the direction is horizontal, it indicates that the sequence add a gap from the column gene, ie F(base, _).
2. If the direction is vertical, it indicates that the sequence add a gap from the row gene, ie F(_, base).
3. If the direction is diagonal, it indicates that the sequence is pairing, ie F(base, base).

Focus on the last cell (G,G)=-4 , trace back the selection of the origin of the cell value (which cell combines it and what decision have it made), we then get the best alignment score:

ACCG and AC_G = -4 where _ stands for gap.

More information:

Gap: due to insertion/deletion, gene duplications.

Mismatch: due to mutation.

The scoring matrix is defined by the similarity between two sequences.

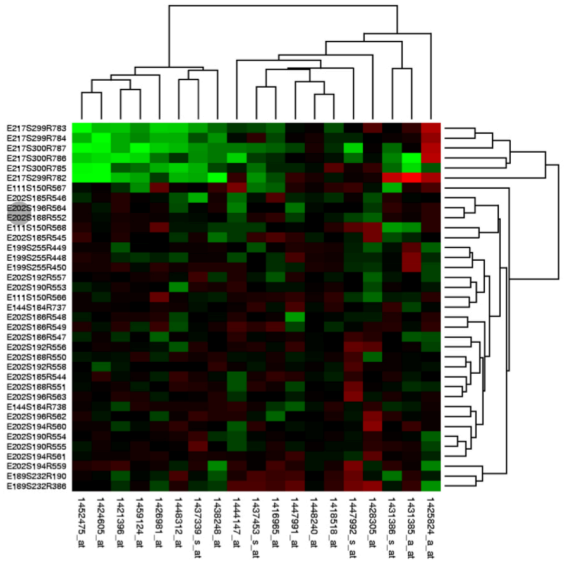
To compare the size difference between enumeration and DP: for n =300, enumeration = 300: $7 * 10^{88}$. DP = 300: 90000

Gene Expression Matrix:

The gene expression matrix represents the expression levels of genes across different conditions or samples. Each row typically corresponds to a gene, while each column represents a different experimental condition or sample.

Why do we Gene Expression Matrix?

For some cases (specifically, in homo sapiens), the genome is mostly the same (the genetic variation only account for 0.001% in the genome). Also, only very small portion of human genome encodes protein (~1% of the entire genome). That is why, knowing the sequence of the genome is not enough. Gene expression difference may account for the phenotype difference and we find them using Gene Expression Matrix.



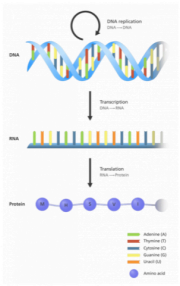
How do we transit data matrix from sequence data?

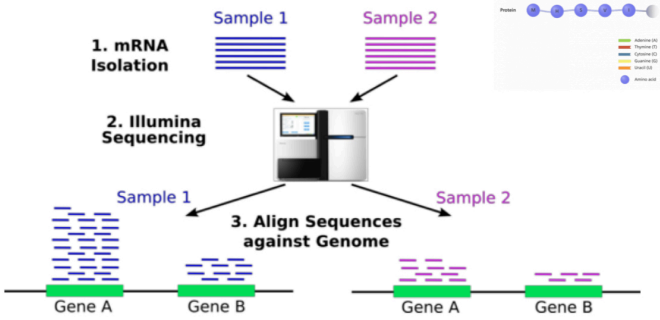
RNA-Seq detects and quantify messenger RNA (mRNA) molecules in a biological sample at a given time, enabling us to measure gene expression levels.

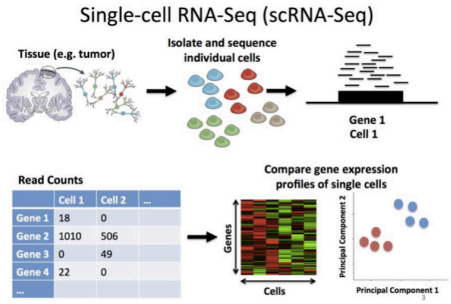
From Sequencing to Gene Expression Matrix:

- **RNA extraction and library preparation:** RNA is extracted from samples and prepared for sequencing.
- **Sequencing:** Sequencing produces millions of short reads.
- **Alignment/Mapping:** Reads are aligned to a reference genome to identify their origin.
- **Counting:** The number of reads mapping to each gene is counted to generate raw count data.
- **Matrix construction:** Raw counts are normalised and arranged in a matrix where rows represent genes and columns represent samples or conditions.

RNA-seq







Genome assembly and mapping during sequencing:

Assembly

As the Illumina sequencing length is around 200bp, where the short reads can have overlap regions. We can utilise the property the seeable base on the short reads and overlap regions (like a jigsaw puzzle).

The idea would be like this:

❖TAA, AAT, ATG, TGC, GCC, CCA, CAT, ATG, TGG, GGA, GAT, ATG

```
T A A
  A A T
    A T G
      T G C
        G C C
          C C A
            C A T
              A T G
                T G G
                  G G A
                    G A T
                      A T G
T A A T G C C A T G G A T G
```

Mapping

Mapping can be done by sliding each read along the genome, then we calculate the difference using dynamic programming (or other methods). If there is no difference between read and genome, gene expression is then counted.

Mapping example

❖TAATGCCATGGATG TAA, CCA, GAT, GCC, CCA, ATG

❖Slide each read along the genome, calculate the difference

- ➢Each time, we may use dynamic programming to calculate the difference
- ➢For simplicity, we would not use it for now

```
T A A T G C C A T G G A T G
        C C A
  2 3 3 3 2 0 2 3 3 2 3 3
```


Mapping example

❖TAATGCCATGGATG TAA, CCA, GAT, GCC, CCA, ATG

❖Slide each read along the genome, calculate the difference

- ➢Each time, we may use dynamic programming to calculate the difference
- ➢For simplicity, we would not use it for now

```
T A A T G C C A T G G A T G
        C C A
  2 3 3 3 2 0 2 3 3 2 3 3
```