BMEG3105 Lecture 5

From Sequence to Gene Expression matrix: Assembly and Mapping

Friday, 20 September 2024

LIPeiran 1155174020

Lecture's agenda

- Specification about Dynamic Programming
- Process of getting gene expression matrix

Dynamic Programming:

Motivation for Dynamic programming to make alignment:

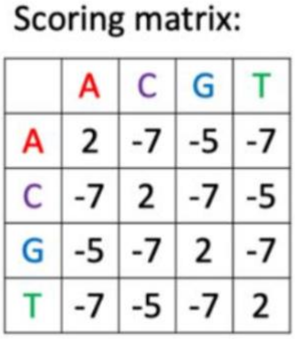
The regular process takes a large computational cost, result into too many alignments. If we align two sequences of the gene sequence with length n:

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2}$$

However, if we use the dynamic programing, filling the table will be done by the time complexity n^2

Scoring Matrix:

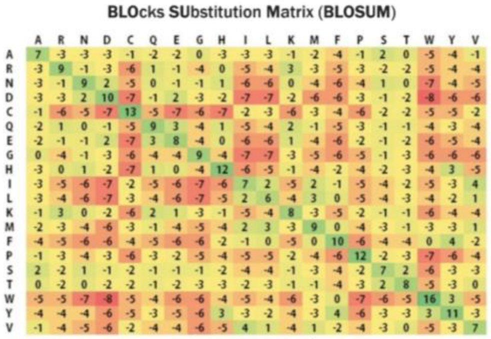
Scoring matrix for DNA: The scoring matrix is defined by the similarity score and gap penalty. Here is an example of scoring matrix with the match pair score = 2, the gap penalty = 10.



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

Scoring matrix for protein:



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

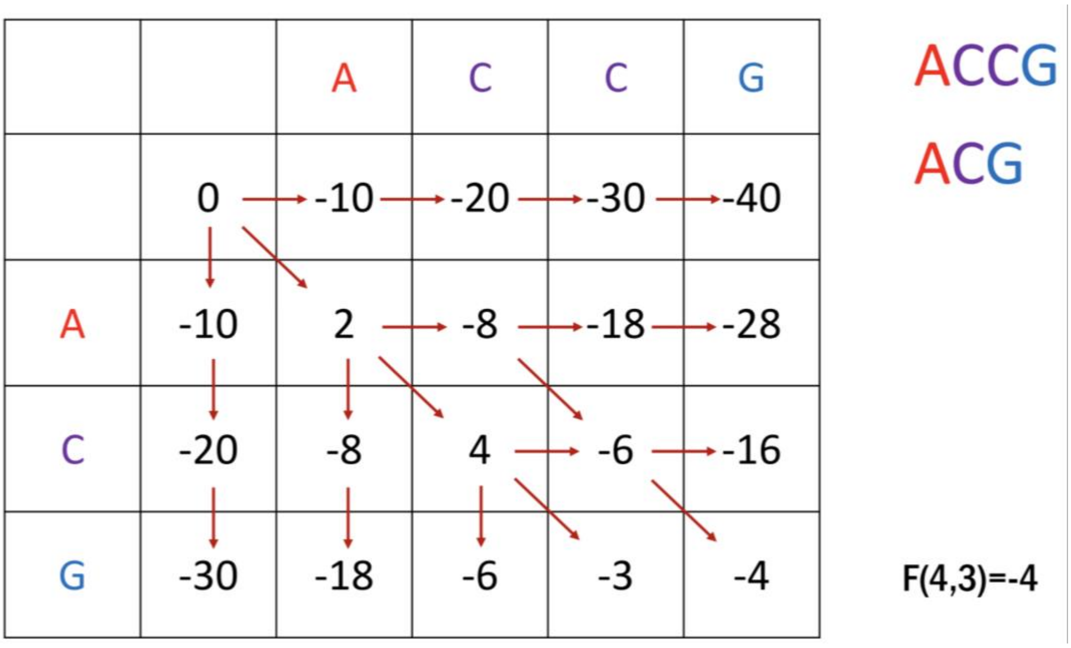
Dynamic Table & Sequencing

Key Idea: Using the dynamic programming (DP) to resolve the final problem. DP refers to divide the original problem into subproblem and solve those problem recursively.

Process:

1. Create a scoring matrix with clearly penalty and similarity score
2. Fill in DP table example of an empty DP table and mark the path with directed arrow
3. Trace back the arrows to obtain the alignment

Example Demonstration:



		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

- Convert the possible combination into table representation
- Horizontal and vertical direction means the gap since the total number of the gene is different
- Diagonal direction has two possible outcomes that is match or mismatch
- Trace back from the last cell and go along the arrows to get the best alignment
- Notice there could be two possible optimal alignments

Other Information:

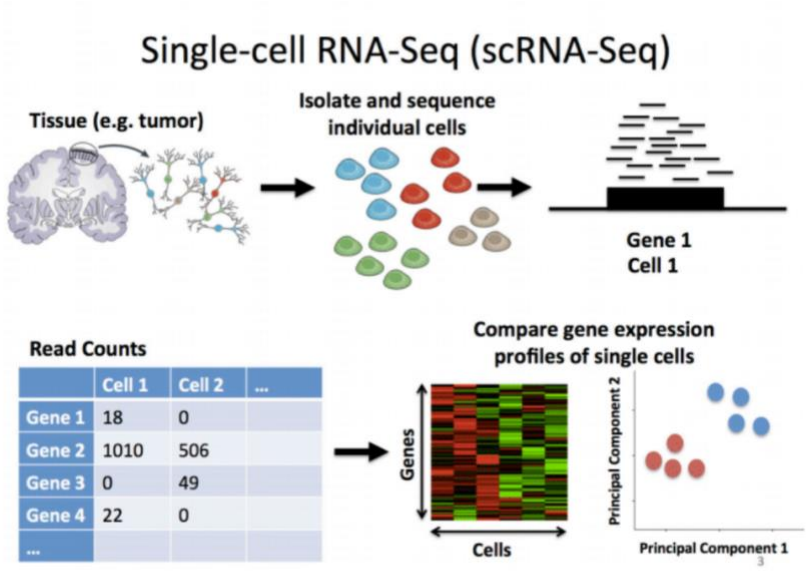
- Local alignment has similar components, motifs, and domains, in dissimilar sequences. Global alignment only care about the number in this cell
- Scoring matrix define the similarity between two sequences, like “the number of all the matched pairs”

Gene Expression Matrix

Motivation:

- Biologists believe genotype is determined by the sequences
- There is only 0.001% of the genome varies and only 1% of the genome encodes proteins
- Knowing the sequence of the genome is not enough since the gene expression difference may account for the phenotype difference

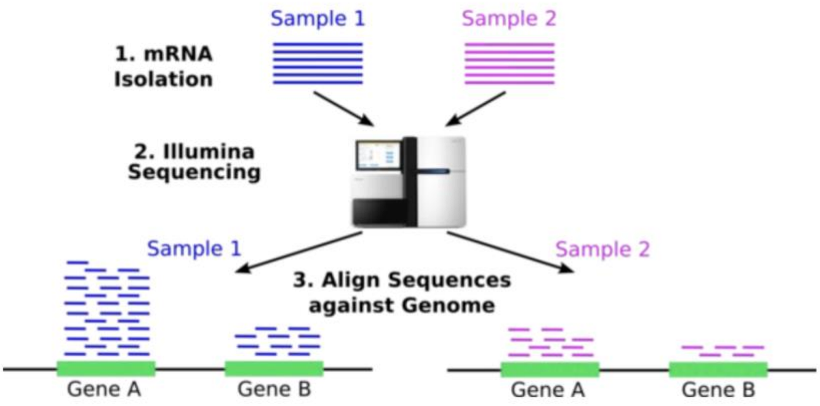
Basic Process of gene expression:


Single-cell RNA-Seq (scRNA-Seq)
Tissue (e.g. tumor)
Isolate and sequence individual cells
Gene 1
Cell 1
Read Counts
Compare gene expression profiles of single cells
Genes
Cells
Principal Component 2
Principal Component 1

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

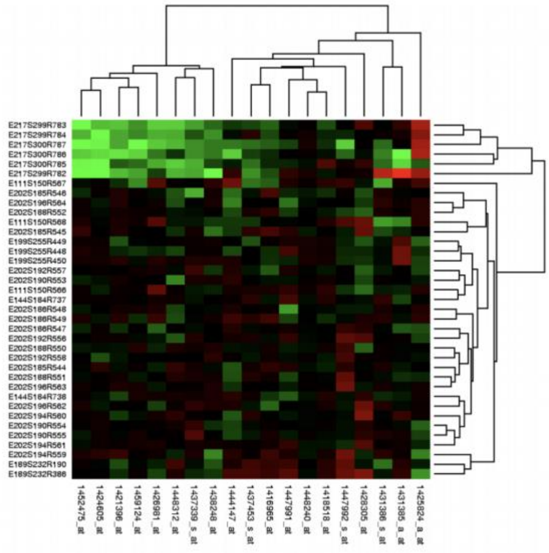
The image illustrates the process of single-cell RNA sequencing (scRNA-Seq). It begins with tissue sampling, such as from a tumor, where individual cells are isolated and sequenced. Each cell's gene expression is measured, producing read counts for various genes. The data is visualized in a heatmap, showing gene expression profiles across different cells. Additionally, a principal component analysis (PCA) plot compares the gene expression profiles of individual cells, highlighting differences and similarities among them.

Process of finding the activate part of the genome:



1. mRNA isolation: isolate the mRNA in sample1 and sample2
2. illumine sequencing: sequence the mRNA from the sample, generating a large number of short sequences reads
3. Sequence Alignment: short sequence reads are aligned to the genome to determine the expression level of each gene

Matrix Demonstration:



This matrix represents the expression level of each gene. The vertical directions mean the different gene sequence and the horizontal line means the different cells or condition.

Normally,
Red cell indicates a large expression value
Green cell indicates a small expression value