BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine
Lecture 04 – Assembly and Mapping (16/09/2022)

Lecture Outline:
1. More about Dynamic Programming
2. Why and How to get Gene Expression Matrix
3. Sequence Assembly and Sequence Mapping

1. More About Dynamic Programming

1.1 Logic of Last Lecture

- Why do we care about sequence data?
- How do we get the sequence data?
- What kinds of processing do we do to the sequence data?
- How do we do sequence comparison?
- How do we do sequence comparison efficiently?
- How to invent dynamic programming by ourselves?
- What is the actual process of dynamic programming?

Prof Li reminded us that the thinking process and problem-solving skills are even more important than only remembering the actual implementation of dynamic programming.

1.2 How do we do dynamic programming?

1. Prepare a scoring matrix.
2. Fill in the dynamic programming table and draw arrows to indicate.
3. The value in the last cell is the best alignment score.
4. Tracing back could get the optimal alignment.

Scoring matrix:

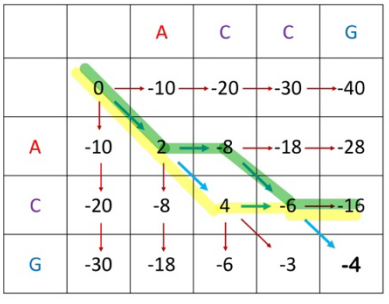
	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Gap penalty = -10

		A	C	C	G
	0	-10	-20	-30	-40
A	-10	2	-8	-18	-28
C	-20	-8	4	-6 (F(3, 2), F(4, 2))	-16
G	-30	-18	-6	-3 (F(3, 3))	?? (F(4, 3))

(-6)+2 =-4
(-16)+(-10) = -26
(-3)+(-10) =-13

For example, the optimal alignment of (ACCG, AC) could be obtained by tracing back in the table.



Noted that for global alignment, only the number in the last cell (-4 in this case) is considered as important.

For local alignment, only part of the sequence is checked.

1.3 More about Scoring Matrix

Mismatches happen due to mutations while gene duplications contribute to insertion or deletion in sequence alignment.

Finding an optimal alignment and the corresponding alignment score depends on the scoring matrices, which could be built from different databases.

Here lists one of the scoring matrices for protein alignment.

Noted that even if any two amino acids match, it does not necessarily mean the same value for different pairs of amino acids in calculating the alignment score.

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

Suppose we define the similarity between two sequences as the number of all matched pairs. Considering the alignment of (AAACCCTTT, ACT), the alignment score equals to three as there are three matched pairs (“A” and “A”, “C” and “C”, “T” and “T”) while mismatches and adding gaps result in no change in alignment score.

2. Why and how to get gene expression matrix?

2.1 Why do we study sequence data?

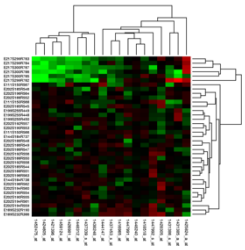
Phenotype, which means how one looks like, is determined by genotype and the environment. Genotype is believed to be determined by the sequences.

2.2 Problem of similar genomes between humans

There is little genetic variation, which only account for 0.001% in the genome of all people. Only 1% of human genome encodes protein. Hence, understanding the genome is not enough. By studying the gene expression of different individuals, understanding how different genes in different cells in different individuals varies in different expression levels, gene expression difference could explain the phenotype difference.

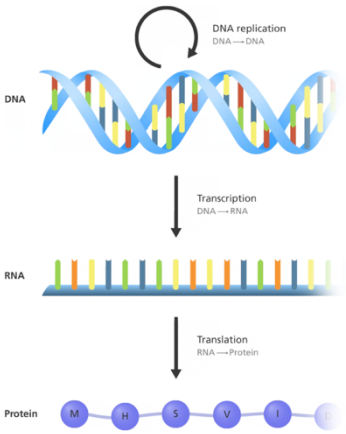
A gene expression matrix in the form of a data matrix could be generated by the sequence data to understand the phenotype difference of individuals.

2.3 Gene expression matrix



In the matrix, the columns represent different samples from different individuals, the rows represent genes and the colour intensity represent the gene expression level.

2.4 How to convert sequence data to gene expression matrix?



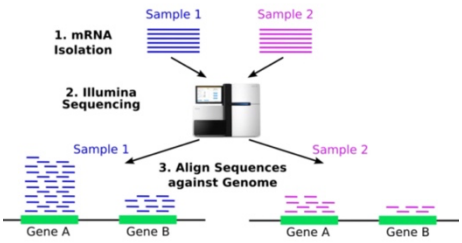
As DNA sequence is transcribed to RNA sequence and RNA sequence is translated to protein sequence, RNA sequence is required to understand gene expression and count the number of proteins that are generated by the genes.

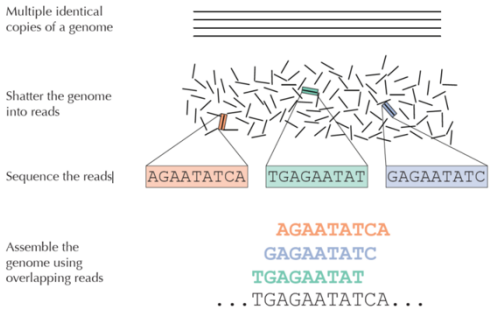
The goal is to map the short reads to the genome sequence, counting the number of reads to generate a gene expression matrix.

3. Sequence Assembly and Sequence Mapping

3.1 Sequence Assembly

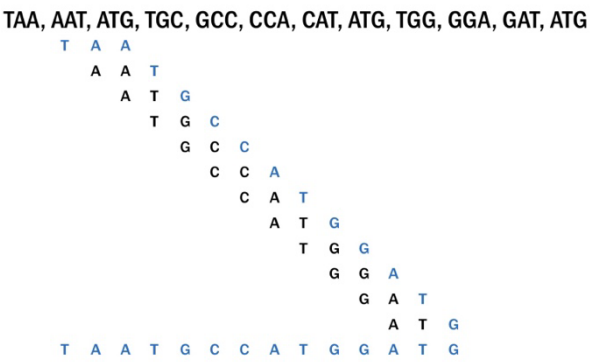
Illumina Sequencing gives two 200 base pairs of short reads, which always have overlap regions. The genome of around three billion base pairs long is assembled based on the shorted reads and overlap regions.





Suppose shorts reads of TAA, AAT, ATG, TGC, GCC, CCA, CAT, ATG, TGG, GGA, GAT, ATG are obtained. Below shows how sequence assembly is done to obtain the genome.

However, problems such as mutations, conflicts, repeated sequences and repeat genes increase the difficulty of sequence assembly.



3.2 Sequence Mapping

Suppose the genome TAATGCCATGGATG is assembled. By mapping the small reads TAA, CCA, GAT, GCC, CCA, ATG, the gene expression of gene GCCA is evaluated as shown below.

Slide each read along the genome to calculate the difference, either by counting or by dynamic programming.

This is the way to convert sequence data to a gene expression matrix.

Additional Resource:

1. A Survey of best practices for RNA-seq data analysis

2. Bioinformatics Algorithm: Chapter 6 & 9 (Textbook)