BMEG3105: Data analytics for personalized genomics and precision medicine

Lecture 3 Scribing [Sequence data and Dynamic programming], Sep 10 (Wed)

Outline

- Questions and comments
- Recap
- Sequencing methods
- Sequence Processing
- Sequence alignment and similarity
- Introduction to DP
- Summary and resources

A. Review of Comments

Student Feedback:

- Mostly positive: "good," "interesting," "clear."
- Suggestions for more detailed explanations, especially for Python.

B. Recap of Types of Data and Sequence Types

- Biomedical data that we will encounter: genes expression, proteome, molecular & cellular network hospital test result, travel history, lifestyle, social data, etc.
- Types of data:
    - Sequential data
    - Data matrix
    - Spatial data
    - Temporal data
    - Graph/Network data
    - Text data
    - Multimodality data
- The purpose of learning programming language: program is like a communication software connects between humans and computers. We need to learn the language to let the translator (Python) translate the programming codes to the codes that machines know.

C. Sequence Data: Foundation of Genomics

Why Sequence Data?

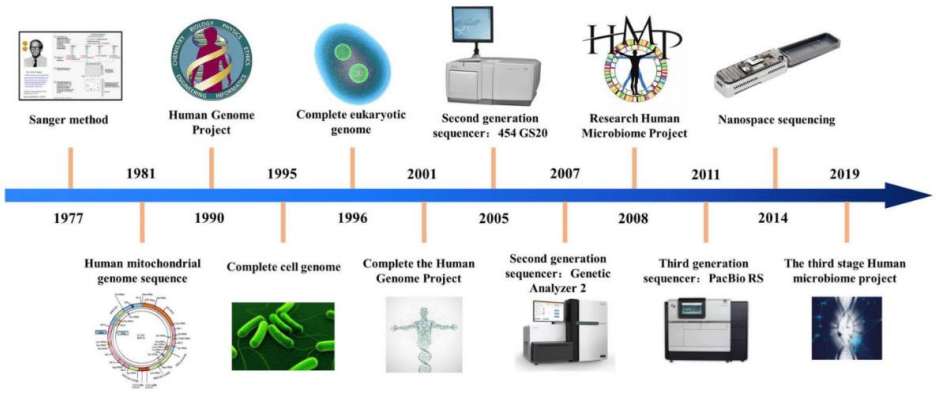
- Follows the central dogma (DNA → RNA → Protein).
- We have to decode the genetic sequence to know the genetic information (some genes may represent a higher possibility of having a disease).
- Phenotype arises from both genotype and environment. And genotypes are determined by the DNA sequence.

What Constitutes Sequence Data?

- **DNA sequence:** A, T, C, G (double-stranded, ~3 billion bp)
- **RNA sequence:** A, U, C, G
- **Protein sequence:** 20 amino acids; analyzed with multiple sequence alignment

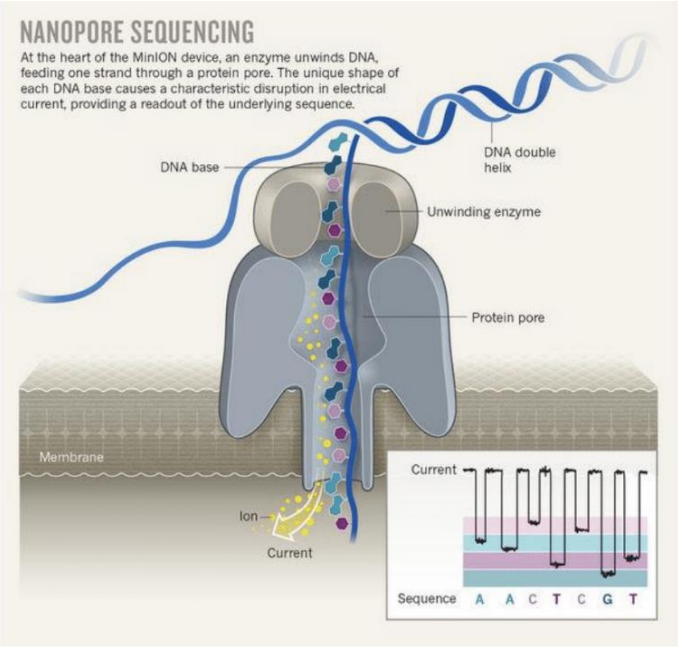
GETTING Sequences

- **DNA/RNA sequencing:** Rapidly developing, moving toward long reads.



*** the first generation can only sequence ~200bp and it uses radioactive markers (low efficiency). ***the Human Genome Project sequences the whole genome of a person.

- **Nanopore sequencing:**
    - While each DNA base in the genome passes through the chemical pore, it gives an unique current change.



    - It supports long reads (up to 4Mb) and the error rate is about 1% now, while PacBio is about 0.001%.

- **Protein sequencing:** Uses mass spectrometry—fragments are weighed and assembled like a jigsaw puzzle.

D. Processing and Using Sequence Data

General Flow to Deal with DNA Raw Data:

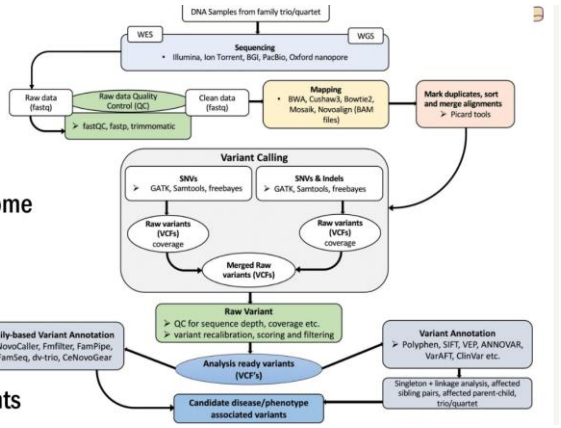
❖DNA sequences

❖Quality control

❖Map reads to reference genome

❖Variant calling

❖Phenotype associated variants



For Protein:

- Sequence comparison
- Multiple sequence alignment
- Sequence similarity helps predict structure and function, or infer evolutionary homology

E. Sequence Comparison & Alignment

Definition: Sequence Similarity= The alignment score of the best alignment.

***Why Compare Sequences?

- Determine similarity, infer functions, or discover evolutionary mechanisms by identifying the conservative/non-conservative regions, etc.

Pairwise Alignment

- Align two sequences to maximize similarity.
- Score alignment via:
    - Match
    - Mismatch (substitution) [may be caused by mutations]

- Gap (insertion/deletion)

Sample Scoring Matrix & Example

	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

- Matches score +2, mismatches -5 or -7, gap penalty -10
- Example:

  AGGCCG and ATGC_G

  Alignment score: $ 2 + (-7) + 2 + 2 + (-10) + 2 = -9

F. Introduction to Dynamic Programming (DP)

Enumeration Problem

- The number of possible alignments grows exponentially with sequence length (e.g. for a 300bp sequence, the combination is 7 *10^ {88}; more than atoms in the universe).

DP Solution Overview

Analogy of DP: finding the cheapest way from KAUST to CUHK

independently

$$\text{Cheapest}(\text{KAUST, CUHK}) = \text{Cheapest}\begin{cases}\text{Cheapest}(\text{KAUST, Qatar}) + \text{Cheapest}(\text{Qatar, CUHK}) \\ \text{Cheapest}(\text{KAUST, Dubai}) + \text{Cheapest}(\text{Dubai, CUHK}) \\ \text{Cheapest}(\text{KAUST, GZ}) + \text{Cheapest}(\text{GZ, CUHK})\end{cases}$$

- Adopt the recursion method, reduce the combination of the above example to 300*300=90000 (efficiency improvement).
- **Steps in DP:**
  - Break the alignment problem into smaller subproblems.
  - Solve these subproblems recursively and optimally.
  - Build up to the global solution.

Sequence Alignment with DP: Example in class

- Input: Aligning "ACCG" and "ACG" using the given scoring matrix and gap penalty

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))	-16 (F(4, 2))
G	-30	-18	-6	-3 (F(3, 3))	?? (F(4, 3))

(-6)+2 =-4

(-16)+(-10) = -26

(-3)+(-10) =-13

Y

***Notes

- Keep the arrows so that we can trace back the original path, also we can identify if there is more than one way to achieve the result.
- To simply put, both the horizontal and vertical moves just add the gap penalty. e.g. if penalty gap is -10, moving rightward a frame just needs to add (-10).

H. Summary & Links

- Sequence alignment with scoring matrices quantifies similarity.
- Dynamic programming offers a computationally feasible approach which the calculation to n^2.
- Some other score matrix can be used e.g. BLOSUM

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

- Website for sequence alignment: https://www.ebi.ac.uk/Tools/psa/emboss_needle/
- We can also use Biopython to perform alignment in Python

  https://www.ebi.ac.uk/Tools/psa/emboss_needle/