Outline of Lecture

- Sequence data
- Sequence comparison and alignment score
- Dynamic programming
- Uncovered resources

1 Sequence data

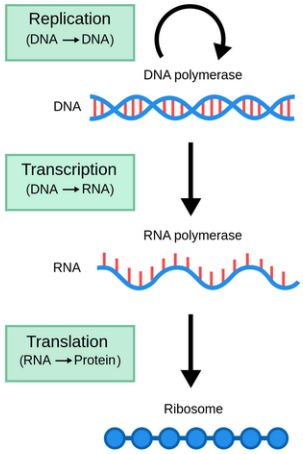


Fig. 1 central dogma

What is sequence data
- DNA (A, T, C, G)
- RNA (A, U, C, G)
- Protein (20 symbols of amino acids)

Why sequence data?
- Central dogma
  "DNA makes RNA, and RNA makes protein", the genetic information and genotype is hidden in DNA sequences

How to acquire sequence data
- DNA sequencing
  - Sanger method (1977), involves electrophoresis and make use of DNA replication
  - Next generation sequencing NGS (1996), based on the measurement of luminescence generated because of pyrophosphate synthesis
    - Principal of NGS
      - Sample/ library preparation
      - Amplification and sequencing
      - Data output and analysis
  - Third generation sequencing (2010), using single molecule real-time (SMRT) sequencing technologies
    - Nanopore sequencing (electrical current change)
- Protein sequencing

- Mass spectrometry MS
  - Break long sequence into short pieces
  - Measure MS of pieces and combine them

How to make use of sequence data

- Compare newly discovered data with already known one (reference genome) to see if there is candidate disease or phenotype associated variants, method such as
  - Multiple sequence alignment

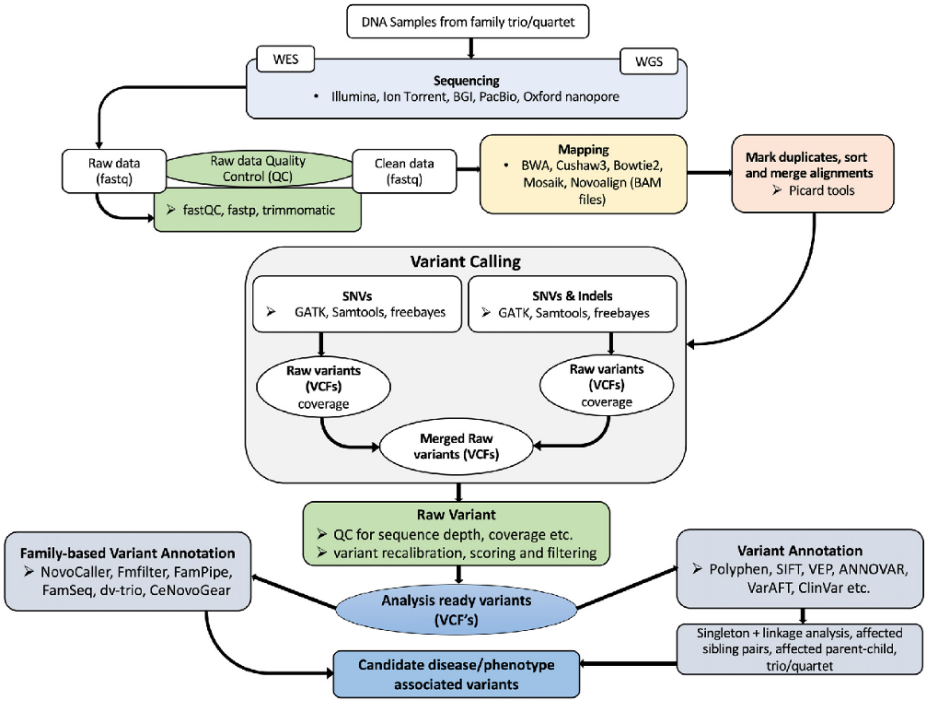


Fig. 2 Ways and methods to use sequence data

2 Sequence comparison and alignment score

Idea: To determine the similar regions between sequences, where similarity might give

- Biomolecular function
- Property prediction
- Evolution of a kind: identifying conservation region and investigating mechanism

How to do

- Maximize the similarity between sequences
- Define similarity

- Cases can be combined to give Scoring matrix (value in the score matrix can be customized)
  - DNA

	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 *Fig. 3 Example of DNA scoring matrix*

    - Match (A <-> A gives +2)
    - Mismatch (substitution, A <->C gives -7)
    - Gap (Insertion or deletion, A <-> _ gives -10)
  - Protein
    - 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

Fig. 4 Example of protein scoring matrix

- To find the best pairwise alignment (the highest score alignment), dynamic programming is used to reduce computation complexity and expenses of enumeration

3 Dynamic programming

Idea: Reduce the problem into smaller problem and reuse the result of smaller problem

Example of DNA pairwise alignment:

- Given the scoring matrix **(SM)** same as above, align ACCG and ACG, (Note SM (A, A) as A<->A = 2)
- Steps:
  - Create a DP table

		A	C	C	G
	(0,0)	(0,1)	(0,2)	(0,3)	(0,4)
A	(1,0)	(1,1)	(1,2)	(1,3)	(1,4)
C	(2,0)	(2,1)	(2,2)	(2,3)	(2,4)
G	(3,0)	(3,1)	(3,2)	(3,3)	(3,4)

- Start from (0, 0) as 0, propagate towards (3,0) and (0,4), as all those are type ‘gap’, -10 score each time.

		A	C	C	G
	(0,0) 0	(0,1) -10	(0,2) -20	(0,3) -30	(0,4) -40
A	(1,0) -10	(1,1)	(1,2)	(1,3)	(1,4)
C	(2,0) -20	(2,1)	(2,2)	(2,3)	(2,4)
G	(3,0) -30	(3,1)	(3,2)	(3,3)	(3,4)

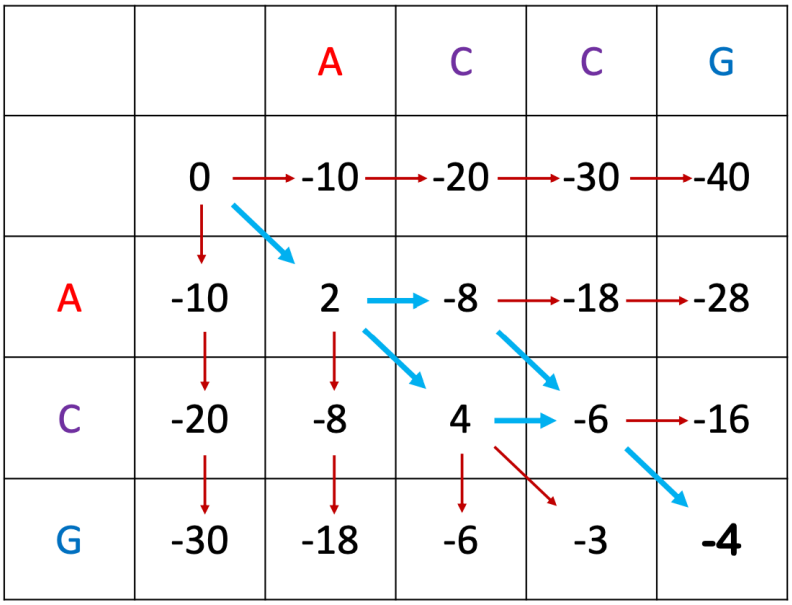
- Then all other cell (x, y) is determined by
  - Min (SM (A, A) + (x-1, y-1),
    SM(GAP) + (x-1, y),
    SM(GAP) + x, y-1))
  - Remember to keep the pointer which gives the smaller value

		A	C	C	G
	(0,0) 0	(0,1) -10	(0,2) -20	(0,3) -30	(0,4) -40
A	(1,0) -10	(1,1) 2	(1,2)	(1,3)	(1,4)
C	(2,0) -20	(2,1)	(2,2)	(2,3)	(2,4)
G	(3,0) -30	(3,1)	(3,2)	(3,3)	(3,4)

- Gives

		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

- Trace back the optimal alignment by the red arrows and highlight it in blue



Gives two optimal alignments with optimal score "-4"

- ACCG, A_CG
- ACCG, AC_G

Existing tool for pairwise sequence alignment

- EMBOSS Needle (online)
  - https://www.ebi.ac.uk/Tools/psa/emboss_needle/
- Biopython (library package of python)
  - https://biopython.org/

4 Uncovered resources

- Time and space complexity analysis
  - O(m*n) to align two sequence of lengths m and n by DP, which is still time consuming
  - O(l*m*n) time and O(m*n) space to find sequence in database with l length-n sequences that compare with query sequence of length m
  - Heuristic method like BLAST and FASTA is used

- Local alignment
  - A local alignment aligns a substring of the query sequence to a substring of the target sequence.
- Multiple sequence alignment
  - Multiple sequence alignment is the alignment of three or more sequences of similar length
- Affine gap penalty

- Gap penalty is higher when there is consecutive gap in alignment, e.g.
  - -10 value every time
  - -10 value for first gap, -25 value for two consecutive gaps

Reference

1. Central Dogma https://theory.labster.com/central_dogma_molecular_biology_pre/
2. Dynamic programming https://www.geeksforgeeks.org/dynamic-programming/#:~:text=Dynamic%20Programming%20is%20mainly%20an,compute%20them%20when%20needed%20later.
3. History of DNA sequencing https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
4. Huristic of BLAST https://www.youtube.com/watch?v=jzSIC2UzxZ4
5. Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance https://www.semanticscholar.org/paper/Next-Generation-Sequencing-and-Bioinformatics-of-Kanzi-San/eee8049718a926efe01e331a2e3eb5c8b2c23457/figure/1