BMEG3105 **Fall2022**

Data analytics for personalized genomics and precision medicine
Lecture3 scribing

Lecturer:YuLI
Wednesday, 14 September 2022
Scriber: Kam Hei Man 1155144587

Sequences in biology:

- DNA sequence, RNA sequence, protein sequence

Nanopore sequencing:

- DNA goes through a chemical pore and read the sequence by detecting different signal given by different base.
- +ve: no need to break sequence but able to read very long sequence.

Protein sequencing:

- Step 1. Break long sequence into short pieces.
- Step 2. Each short pieces can be determined by their weight (mass spectrometry).
- Step 3. Assemble the short pieces back to the raw sequence.

Raw data (DNA sequence) handling:

- Quality control
- Map reads to reference genome
- Variant calling
- Phenotype associated variants

Sequence alignment:

- To compare more than 1 sequence by finding the similarity between sequences.
- Example: Pairwise sequence alignment.

Pairwise sequence alignment:

- Step 1. Arrange 2 sequences.
- Step 2. Maximize the similarity by inserting gap.
- Step 3. Calculate the maximum sequence alignment score (i.e. score that represent the alignment with the highest similarity) by the defined scoring matrix.

Scoring method:

- Method 1: Enumeration
  - Problem: Too many possibility !!!
- Method 2: Dynamic programming

Dynamic programming:

- Reduce big problem to smaller sub-problems
- Aim: reduce all problem to boundary case so that values are known and scoring matrix can be used
- Example:

Method 1

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

$$F(ACCG, ACG) = \text{Best} \begin{cases} F(ACC, ACG) + F(G, \_) \\ F(ACCG, AC) + F(\_, G) \\ F(ACC, AC) + S(G, G) \end{cases}$$

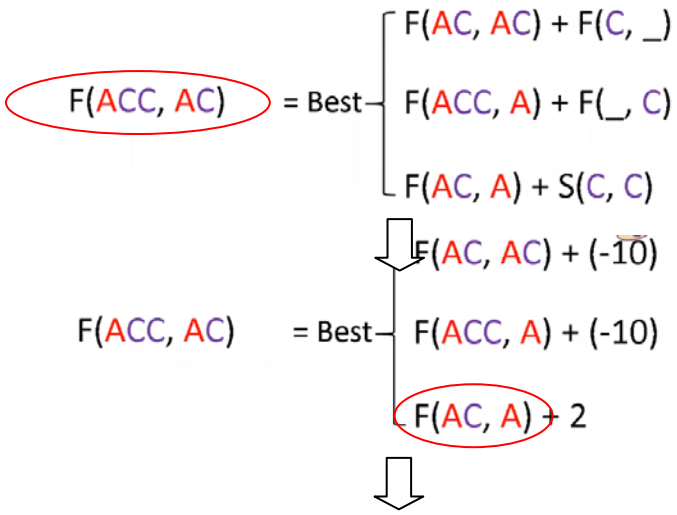
F(G,_) = F(_,G) = gap penalty

S(G,G) = match

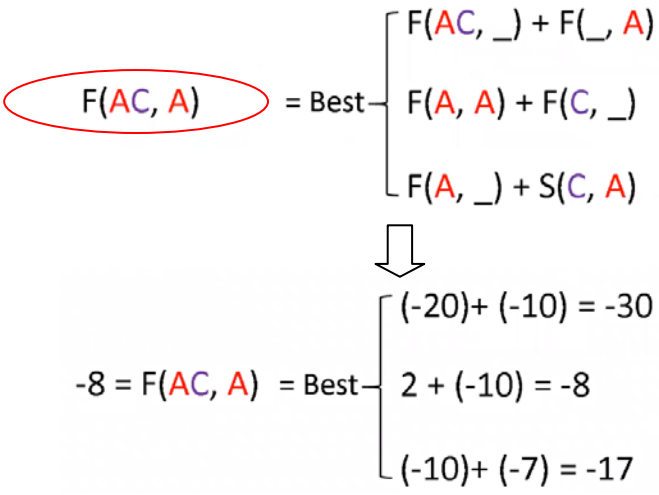
According to scoring matrix:

$$F(ACCG, ACG) = \text{Best} \begin{cases} F(ACC, ACG) + (-10) \\ F(ACCG, AC) + (-10) \\ F(ACC, AC) + 2 \end{cases}$$

Solve sub-problem:



All are boundary cases !!!



Best= Highest score= -8 (-30<-17<-8)

…….

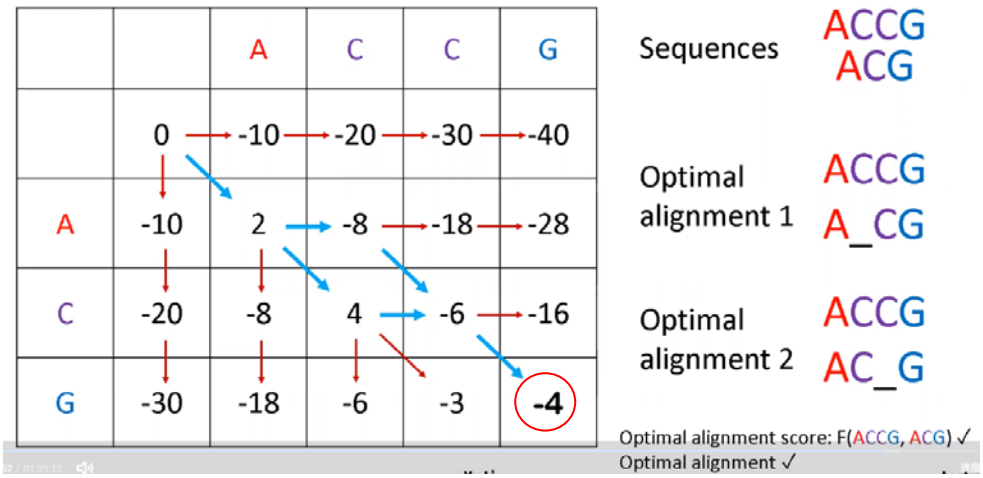
Method 2

Table representation:

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



- Step 1. Fill in the table according to the scoring matrix.
- Step 2. Draw arrows that have the smallest value difference between the previous one and the one on ↓ → ↘ direction.
- We can find:
  - Arrow = the alignment arrangement of 2 sequences
  - ◯ = sequence alignment score