BMEG3105 Fall 2023

Data Analytics for Personalized Genomics and Precision Medicine

Lecture 4: Dynamic Programming

Lecturer: Yu LI (李煜) from CSE

Liyu95.com, liyu@cse.cuhk.edu.hk

15 September 2023

Lecture outline:

- Recap from last lecture
- Dynamic Programming

Part 1. Recap from last lecture

Sequence Data:

- DNA sequence
  - Composed of A, T, C, G bases
  - Consists of complementary double strands
- RNA sequence:
  - Composed of A, U, C, G bases
- Protein sequence:
  - Composed of 20 amino acids
  - Multiple sequence alignment
- To find the best pairwise alignment
  - With two sequences and scoring matrix known
  - Enumeration as a solution
    - Enumerate all the possible alignments between two sequences
    - Calculate scores for all alignments
    - Choose the alignment with the highest score
    - Problem with this solution is that there are too many possible alignments

Part 2. Dynamic Programming (DP):

- Break down a problem into smaller sub-problems
- Solve these sub-problems optimally and recursively
- Use these optimal solutions to construct the optimal solution for the original problem

Flight problem example:

- Direct flight not always cheap
- Finite pre-destination
- Break the flight into several flight trips
- Compare the sum of flight trips for each option and choose the cheapest option

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

Similarly with Optimal alignment score:

- Each base either aligns to a gap or another base
- Alignment score equals to the sum of score for each alignment pair

Sequence alignment with DP:

- Consider the possibilities of the last pair of the alignment
- Calculate alignment score and choose the best option
- Break down the best option and repeat the last step

ACCG and ACG alignment example:

1. Consider the last base pair. (_ represent blank)

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

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

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

➔ Choose the option with highest last base pair score, i.e. F(ACC, AC) + 2

2. Break down unknown part the option chosen

$$\text{F(ACC, AC)} = \text{Best}\begin{cases}\text{F(AC, AC)+ F(C, \_)}\\ \text{F(ACC, A) + F(\_, C)}\\ \text{F(AC, A) + S(C, C)}\end{cases} = \text{Best}\begin{cases}\text{F(AC, AC)} - 10\\ \text{F(ACC, A)} - 10\\ \text{F(AC, A)} +2\end{cases}$$

The Formula is reduced to boundary case

$$F(AC, A) = \text{Best}\begin{cases} F(AC, \_) + F(\_, A) \\ F(A, \_) + S(C, A) \\ F(C, \_) + S(A, A) \end{cases}$$

$$= \text{Best}\begin{cases} -20 - 10 = -30 \\ -10 - 7 = -17 \\ 2 - 10 = -8 \end{cases}$$

The optimal alignment is ACCG AC_G or ACCG A_CG

Table representation of the method:

	Gap	A	C	C	G
Gap	0 → ↓ ↘	-10 →	-20 →	-30 →	-40
A	-10 ↓	2 F(A, A) → ↓ ↘	-8 → ↘	-18 →	-28
C	-20 ↓	-8 ↓	4 F(A, A)+F(C, C) → ↓ ↘	-6 F(AC, AC)+F(C, _) / F(AC, A)+F(C, C) → ↘	-16
G	-30	-18	-6	-3	-4 F(ACC, AC)+F(G,G)

Optimal alignment 1:

ACCG
A_CG

Optimal alignment 2:

ACCG
AC_G

- The arrows preserve the path information
- Arrows in blue highlight the direction that gives the best alignment score
- Trace back the arrows to get the optimal alignment