BMEG 3105 Fall 2023 Data analytics for personalized genomics and precision medicine Topic: Assembly and mapping Lecture: Lecturer: Yu LI (李煜) from CSE Liyu95.com, liyu@cse.cuhk.edu.hk Student: CHANG Hing Lam SID: 1155143887 20<sup>th</sup> Spetember,2023

Expected Outcome:

- 1. Understand what gene expression matrix is and how to use it.
- 2. Introductions to sequence assembly and sequence mapping.

Pre-course survey result:

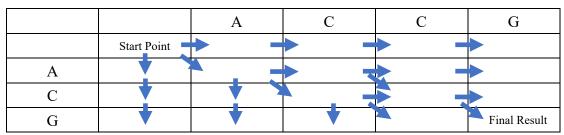
Positive Feedback:	Constructive Feedback:		
- Good.	- The cheapest flight problem was		
- Nice, learn something new.	overexplained.		
- Very interesting and interactive	- Sequence Alignment with DP was		
lecture with good examples.	explained confusingly, and many did		
- It was taught well, and the algorithm	not understand it.		
of the last pair was interesting to	- Could explain how to get the		
learn.	subproblems part more clearly.		
- Clear PowerPoint presentation and	- Consider providing steps on how to		
speed.	get each alignment score in the		
	matrix and showing steps for tracing		
	back.		
	- Request for a code snippet or		
	supplementary material to		
	implement the algorithm.		
Specific Question:	Technical Issues:		
- Why do we add an extra gap in	- The volume is a bit small, and some		
AGGC when comparing it with	words are unclear at the back.		
AGC in sequence alignment with	- It took a long time before the lecture		
DP? Why do we substitute the last	started teaching the topic.		
pair C in AGGC with "-" in the	- Question regarding the survey:		
dynamic programming algorithm?	Uncertainty about what is counted as		
	topic 1, 2, 3.		

## Last Lecture Recap

- Flight Metaphor
  - KAUST  $\rightarrow$  Hong Kong
  - Finite pre-destination (GZ? Dubai? Qatar?)
  - Figure out the cheapest price. (Price = Sum of each flight price)
- Finite Choice of each base (Flight Metaphor)
  - Align to another base / align to another gap.
  - Alignment Score = Sum of each pair in alignment. (Flight Metaphor)
  - Use dynamic programming (Split the problem to sub-problems)
  - F(ACCG, ACG) can come from three sources
    - $\Rightarrow F(ACC, AC) + S(G,G)$
    - $\Rightarrow$  F(ACCG, AC) + S(\_,G)
    - $\Rightarrow F(ACC, ACG) + F(G, \_)$

DP Table (Simplify Reduction Process)

		А	С	С	G
	0	(Gap) -10	(Gap) -20	(Gap) -30	(Gap) -40
		(0-10)	(0-10-10)	(0-10-10-10)	(0-10-10-10-10)
А	(Gap) -10	2	-8	-18	-28
	(0-10)	(0+2)	(0+2-10)	(0+2-10-10)	(0+2-10-10-10)
С	(Gap) -20	-8	4	-6	-16
	(0-10-10)	(0+2-10)	(0+2+2)	(0+2+2-10)	(0+2+2-10-10)
G	(Gap) -30	-18	-6	-3	-4
	(0-10-10-10)	(0+2-10-10)	(0+2+2-10)	(0+2+2-7)	(0+2+2-10+2)



- > Two paths to the result = two optimal alignments.
- Store answers of sub-problems and construction path.
- Procedure of DP
  - Find the suitable scoring matrix.
  - Fill in the DP Table.
  - Best Alignment Score = Last Cell
  - Trace Back to find alignment.

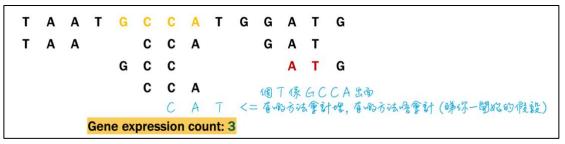
## More Example – Local Alignment

		А	С	С	G
	Start Point 💻	-			
А					Ĭ
С					Ĭ
G					Final Result

## ACCG

		А	С	С	G
	Start Point				
А					
С				-	
G					Final Result

- Scoring Matrix
  - $\blacktriangleright Mismatch = Mutations$
  - Gap = Insertion / Deletion / Gene Duplication
  - > Definition depends on how you define the similarity.
- Sequence Data
  - Central Dogma
  - Hidden Genetic Information
  - Phenotype = Genotype + Environment
  - $\blacktriangleright$  Genetic Variation = 0.001%
  - Senome encodes protein = 1%
- Gene Expression Matrix
  - ► RNA-seq
    - Map Short read to genome
    - Count number of reads  $\rightarrow$  Gene Expression Matrix
- Genome Assembly
  - Find out Overlap Regions (Use short reads)
  - Limitation
    - Mutation / Conflict / Repeat Sequence / Repeat Gene
  - Mapping Example



- Further Improvement
  - Speed / In One Pass / Mutation or Errors
- Resource and Uncover Part
  - Bioinformatics: Sequence and Genome Analysis---Chapter 2&3
  - > Time complexity and space complexity analysis
  - Local alignment
  - Multiple sequence alignment
  - Affine gap penalty
  - Sequence database search: BLAST