

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

20th September, 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:

<p>Positive Feedback:</p> <ul style="list-style-type: none">- Good.- Nice, learn something new.- Very interesting and interactive lecture with good examples.- It was taught well, and the algorithm of the last part was interesting to learn.- Clear PowerPoint presentation and speed.	<p>Constructive Feedback:</p> <ul style="list-style-type: none">- The cheapest flight problem was overexplained.- Sequence Alignment with DP was explained confusingly, and many did not understand it.- Could explain how to get the subproblems part more clearly.- Consider providing steps on how to 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.
<p>Specific Question:</p> <ul style="list-style-type: none">- Why do we add an extra gap in AGGC when comparing it with AGC in sequence alignment with DP? Why do we substitute the last pair C in AGGC with "-" in the dynamic programming algorithm?	<p>Technical Issues:</p> <ul style="list-style-type: none">- The volume is a bit small, and some words are unclear at the back.- It took a long time before the lecture started teaching the topic.- Question regarding the survey: Uncertainty about what is counted as topic 1, 2, 3.

◆ **Last Lecture Recap**

- Flight Metaphor
 - KAUST → 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)
 - $F(\text{ACCG}, \text{ACG})$ can come from three sources
 - ✧ $F(\text{ACC}, \text{AC}) + S(\text{G}, \text{G})$
 - ✧ $F(\text{ACCG}, \text{AC}) + S(_, \text{G})$
 - ✧ $F(\text{ACC}, \text{ACG}) + F(\text{G}, _)$

◆ **DP Table (Simplify Reduction Process)**

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

		A	C	C	G
	Start Point	→	→	→	→
A	↓	↘	→	↘	→
C	↓	↓	↘	→	↘
G	↓	↓	↓	↘	↘
					Final Result

- 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

ACCG_ _ _

		A	C	C	G
	Start Point	→	→	→	→
A					↓
C					↓
G					Final Result

ACCG

		A	C	C	G
	Start Point				
A		→			
C			→	→	→
G					Final Result

◆ Scoring Matrix

- 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
- Genetic Variation = 0.001%
- Genome encodes protein = 1%

◆ Gene Expression Matrix

- RNA-seq
 - Map Short read to genome
 - Count number of reads → Gene Expression Matrix

◆ Genome Assembly

- Find out Overlap Regions (Use short reads)
- Limitation
 - Mutation / Conflict / Repeat Sequence / Repeat Gene
- Mapping Example

T	A	A	T	G	C	C	A	T	G	G	A	T	G
T	A	A			C	C	A			G	A	T	
				G	C	C					A	T	G
					C	C	A						

個T係GCCA出面
 C A T <= 有啲方法會計埋, 有啲方法唔會計 (睇你一開始的假設)

Gene expression count: 3

- 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