BMEG3105: Data analytics for personalized genomics and precision medicine

Fall 2025

Lecture 4: From sequence to gene expression matrix: Assembly and mapping

Lecturer: Professor Yu Li Scriber: Tam Chun Hin

Lecture 4 outcome:

1) Recap on last lecture
2) Using DP table to solve sequence alignment questions
3) Understanding gene expression matrix
4) Brief introduction on sequence assembly and mapping

Recap on previous lecture:

- DNA sequencing is essential.

  DNA holds genetic information

  (phenotype = genotype + environment)

- But human genome is highly similar.

- Data is required to measure gene expression (RNA-seq)

e.g. Professor Lo discover cell-free fetal DNA, with the aid of DNA sequencing => non-invasive prenatal testing for down syndrome.

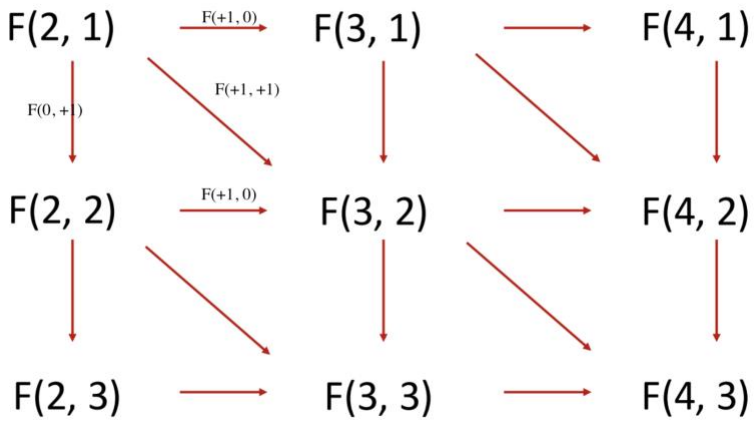
- The use of DP table

  The pre-destination is not unlimited, the base choice as well

  ■ For example: a 7 bases sequence will not have a best result of more than 6 bases.

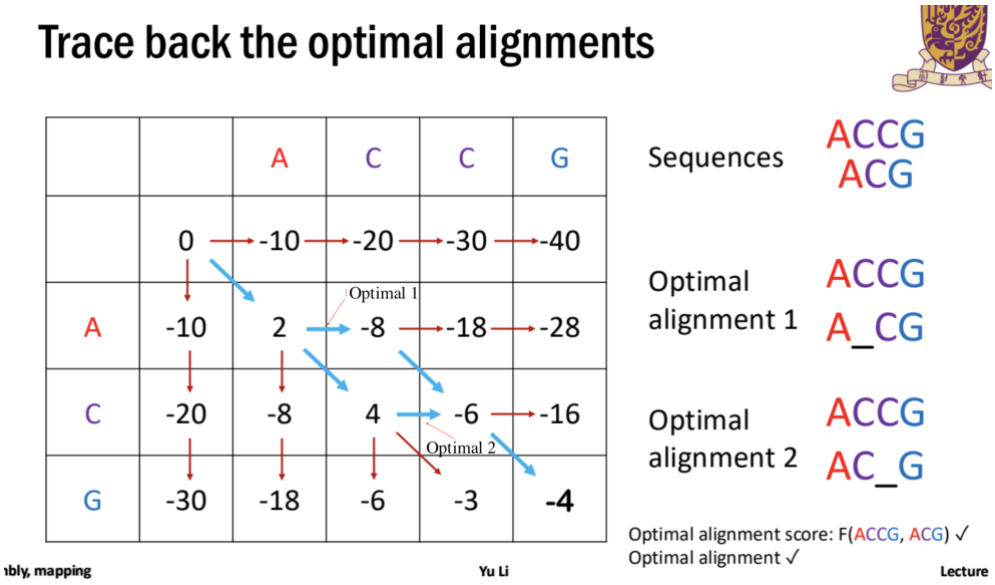
<u>More on DP table:</u>

- Each slot on the DP table can be either adding or reducing at least one base on the adjacent slot.



- Difficult by just remembering, DP table can simplify this process

- Also help if we want to find out the optimal alignment for a sequence (cheapest pathway).



- DP table show which pathway is valid and hence easily determine the pathway to get the optimal alignments.
- Now this sequence alignment can be used:
  1. finding sequence similarity
  2. calculate alignment score (with a scoring matrix)
  3. answering the sub-problems and the optimal path

- Obtained 2 sequences, scoring matrix=> obtain solution
- However, the n can be extremely large.
- Dynamic programming needed

Webserver for sequence alignment:

https://www.ebi.ac.uk/Tools/psa/emboss_needle/

Biopython:

Tool for finding out the alignments

https://biopython.org

```
fromBioimportpairwise2
alignments = pairwise2.align.globalxx("ACCGT", "ACG")
fromBio.pairwise2importformat_alignment
print(format_alignment(*alignments[0]))
```

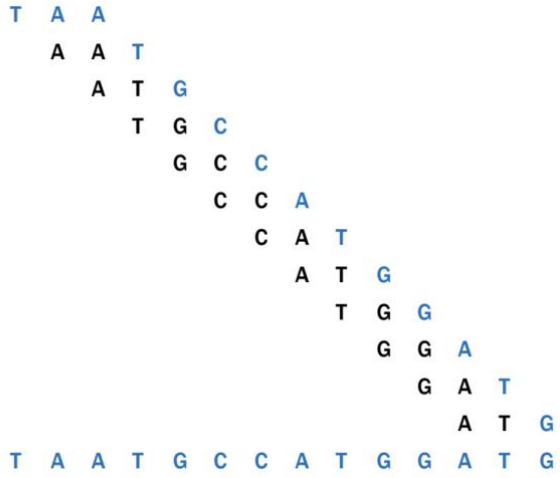
Make a DP table:

- Scoring matrix is needed
- Use the best alignment score as the last cell value
- Put on the arrows

Genome Assembly:

- Illumina sequencing ~200bp
- Human genome ~ 3 billion bp

- Too long for sequencing
- 2 reads with overlapping regions => assembled on the overlapping regions (TAA, AAT => TAAT) like a jigsaw.
- However, problems will be found
    1. Mutation occurred
    2. Conflict on overlap
    3. Repeating units (AAA,AAA,AAA)
    4. Repeating genes (AAT,AAT,AAT)
    5. The process is not efficient



After getting the reads:

- Slide along each read and calculate difference (DP table, but very slow)
- After mapping, count the population of specific gene, obtain

expression level => making data matrix

Readings and uncovered topics:

- 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: BLA
- RNA sequence analysis

  https://www.dropbox.com/s/j6gu44xszr9e8jk/A%20survey%20of%20best%20practices%20for%20RNA-seq%20data%20analysis.pdf?dl=0