BMEG 3105 Fall 2022

**Data analytics for personalized genomics and precision medicine**

**Lecture 5 Scribing**

Lecturer: Yu LI (李煜) from CSE

Friday 20 September 2024

**Lecture agenda**:

- Recap of last lecture
- Discussion of further detail in dynamic programming (DP)
- Understand the function and application of gene expression matrix
- Introduction to sequence assembly and sequence mapping (Notes are included in slides. but due to limited time, it did not mention in this lecture)

**Expected outcomes**:

- Understand the principles of DP and able to apply it to sequence alignment
- Use DP to trace back the optimal alignment
- Able to transition from sequence data to data matrix

**Feedback and comments from last lecture**:

- Positive feedback:
  - Clear and easy to follow
  - Liked the calculation part
- Request and suggestion:
  - Slower explanation of calculations.
  - More examples needed
  - Use of animations for complex graphs

**Recap**:

***Fundamental understanding of Dynamic Programming***:

- Purpose:
  - it solves alignment problems by breaking them into smaller problems
    *[e.g. when we calculate F(4,3), we can break it into F(4,2) + F(3,3) + F(3,2). ]*
  - Find the optimal alignment score to determine the optimal alignment
- Matches of each base:
  - Finite choice for each base
    a. Align to another base
    b. Align to a gap

**Lecture**:

***Dynamic Programming (DP) in Sequence Alignment***:

- Further analysis/application of DP
  - Merge the result of small questions to fix the final problem
  - Arrows show the alignment pathway/arrangements
    a. Trace back from the right bottom conner to left upper conner
    b. Determine the optimal alignment by reversing the alignment pathway (the optimal alignments can be various)
  - Calculate the alignment score
    a. Directly observe the base pair
    b. According to the scoring matrix, add the score together corresponding to base pair (scoring matrix can be various)
    c. Optimal alignment score should be equal to the score on the right lower conner on DP
  - Sequence alignment can be used to identify sequence similarity
- Invent DP and DP process
  - Fill in the table according to the scoring matrix
  - Preserve the arrows
  - The value in the last cell is the best alignment score
  - Trace back the arrows to get the alignment.
- Further information provided by DP table
  - DP table stores answer of sub-problems and the construction path
    *[e.g. from DP which solve the problem of F(5,4), it contains the answer of F(3,3), F(4,2), etc..]*
- Concept of local alignment
  - Similar components, motifs and domains, in dissimilar sequences
  - Only care the local information between two sequencing; care the most important and the number in the cell

***Computational Analysis***:

- Using two sequences and scoring matrix
- Provide straight forward solution
- If using DP will be too much calculation
- Webserver for sequence alignment is provided at supporting link section

***Scoring Matrix***:

- Mismatch causes by mutation
- Insertion/deletion, or gene duplication due to additional insertion during transcription may cause a gap

- Scoring matrices are various
  - Different databases can build different scoring matrices
  - Different scoring matrices can aim different needs
  - There are different types of matrices including specific for DNA, RNA or protein *[e.g. Blocks Substitution Matrix (BLOSUM) is a protein scoring matrix]*
- It depends on how we define the similarity between two sequences

***Data Sequencing***:

- Purpose:
  - Reveal the genetic information which hidden in DNA sequences
- Since human genome is mostly the same, sequencing alignment can help find the differences
  - Gene expression difference is important for studying the phenotype difference

***Gene Expression Matrix***:

- Purpose:
  - Visualize the difference between different gene expressions across sample or environment, etc..
- Principle:
  - The amount of protein that is translated by a specific gene can reveal the gene expression level
  - Since the protein is hard to count, check the counts of RNA copies can also determine the gene expression due to central dogma
- Processing of building gene expression matrix:
  - Map the short read to the genome
  - Count the number of reads, which is content of gene expression matrix

***Potential Project – 1***

- A pipeline to get the gene expression matrix form reaeds
  - Find the genome
  - Find the reads
  - Map reads to reference genome
  - Count reads for each gene
  - Use Google to find the software and the data
  - Explain each step in the report to let us know you understand what you are doing

**<u>Next lecture topic</u>**:

- Where to find/ how to get reference genome
- How do we do genome assembly and mapping
- Data exploration and data cleaning

**<u>Supporting Links</u>**:

- Webserver for sequence alignment: https://www.ebi.ac.uk/Tools/psa/emboss_needle/
- Biopython: https://biopython.org/
- Post-lecture survey: https://forms.gle/4AyB35ztD7QWPdDv8

**<u>Resource and related uncovered topics</u>**:

- 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