Name: LEUNG, Chung Ki
SID: 1155175455

BMEG3105 Data analytics for personalized genomics and precision medicine
Scribing for lecture on 13/9/2023
Scriber: Leung Chung Ki, SID: 1155175455

Course arrangements

Arrangements on scribing:

- If there were future adverse weather leading to lecture cancellation, the TA will inform the affected students about the arrangements
- If one signed up for two scribing sessions, then the best one will be graded for 10%, and the second one will be graded for the bonus 1%

Arrangements on recording:

- No recording provided (as only one student requested)

Recap of last lecture

- Course arrangements (lecture, tutorials, assignments, projects, etc)
- Data available for analysis (from genes to population behaviour)

Different types of data

- Sequential data (DNA base sequences)
- Data matrix (tables)
    - Does not changed the data if the entire row/ column is shuffled
- Spatial data (pictures/ map)
    - Meaning of the value changes with the location of the data
- Temporal data (ECG/ stock prices/ EMG)
    - Values changes over time
- Graph or network (Protein networks/ social media relationships)
    - Relationship changes with the definition of the links
- Text (Wikipedia/ papers/ social media)
- Multi-modality data (video/ electronic health records)
    - E.g. Video (temporal spatial data), is a temporal sequence of images/ audio over time
    - The video transcript is a text data
- Unknown data types
    - Uncertain before the data is revealed

Introduction to programming

- (Simplified) definition of programming: A way to tell computer to do something for you
- (Simplified) explanation of Python: A programming language that you can use to communicate with the computer (only if you use Python)
- Libraries (numpy/ scipy/ pandas/ matplotlib/ etc): pre-written programs that you could use to simplify programming
- Variables: “boxes” that could be used to store values
- print(): ask the computer to show something. To show something directly use “” (double-quotes) to enclose the things to be showed directly

Sequence data & sequence alignment

- DNA sequence (double-stranded, A/T/C/G)
- RNA sequence (singles-stranded, A/U/C/G)
- Protein sequences (20 amino acid-sequences)
- Obtaining DNA/RNA sequences: sequencing (e.g. nanopore sequencing), each sequencing method has its own advantages and drawbacks; nanopore sequencing has long reads but high error rate

- Protein sequencing: separate proteins into small pieces (digestion), then use mass spectroscopy to determine their weights; put the pieces together afterwards
- Uses of sequences: variant calling, function inferences (sequence → structure → function relationship), determine common ancestors (homology), etc
- Sequence alignment: find the arrangement that maximizes the similarity between two sequences
- Assess sequence similarity by scoring (match/ mismatch/ gaps)
- Use of scoring matrix: assign different scores for match/ mismatch/ gaps, then compute the score for the alignment (highest score = best alignment)
  Note: the scoring matrix would change depending on the application (e.g. BLOSUM matrix used for scoring protein similarity)
- Pairwise alignment: the no. of cases goes up very quickly as the length of target sequence increases (C_n^{2n} cases)
- Solution: dynamic programming