BMSB3105 Data Analytics for Personalized Genomics and Precision Medicine-Lecture 3

Topic: Data & Python & Sequences
Lecturer: Professor Yu LI (李煜) from CSE Liyu95.com, liyu@cse.cuhk.edu.hk
Date: 13 September 2023, Wednesday

Lecture Outline

- Different data types
- Introduction to Python programming
- Sequence data
- Sequence comparison and alignment score

Different Data Types

1. **Sequential data**
   - Sequences e.g DNA and RNA

2. **Data matrix**
   - Data that consists of a collection of records, each of which consists of a fixed set of attributes
   - Data set can be represented by an n by m matrix, where there are n rows, one for each object, and m columns, one for each attribute
   - **If you shuffle the entire column or the entire row at one time, you would not change the data
   - Example: 4 by 2 matrix: 4 people, each with 2 attributes

Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

3. **Spatial data**
   - Geographic locations and spatial information involved
   - If you shuffle the entire column or the entire row at one time (image is made of pixels) , you would get a different image data
   - e.g maps and images

4. **Temporal data**
    - With built-in support for handling data involving time
    - E.g ECGs, stocks

5. **Graph or networks**
    - Objects and connections (the link)
    - Social network and PPI network

6. **Text**
    - Short sentences
    - Long sentences or documents

7. **Multi-modality data** (more than one of the above mentioned data)
    - Video: Temporal images, audio, transcript
    - Electronic health records: Data matrix, images, text
    - Spatial transcriptomics: Spatial data, sequence, data matrix

Introduction to Python Programming

1. **Programming**
    - Wiki: "Computer programming is the process of designing and building an executable computer program to accomplish a specific computing result or to perform a specific task."
    - Metaphor: Programming: to communicate with the computer (your friend), asking him/her to do something for you while code is the WhatsApp message you send to the computer (your friend)
    - Computer does not understand human language while we do not understand the computer/machine language. Therefore we need a bridge/translator to help us command the computer ➔ we learn programming language

2. **Python**
    - Wiki: Python is an interpreted high-level general-purpose programming language
    - Metaphor: the WeChat/WhatsApp software to help us communicate with the computer. E.g calculate mean/ other values

3. **Numpy, Scipy, Pandas**
    - Additional plug-in to make Python more powerful ➔ Need to import them first
    - Calculate lots of things e.g mean, variance, median, max, min

```
[2] import numpy
    a = [1,2,3,4,4,5,5,5,6,7,8,9]
    numpy.mean(a)

    4.916666666666667

[4] numpy.std(a)

    2.253084305765962

[5] numpy.median(a)

    5.0

[6] numpy.max(a)

    9

[8] print(a)

    [1, 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 9]
```

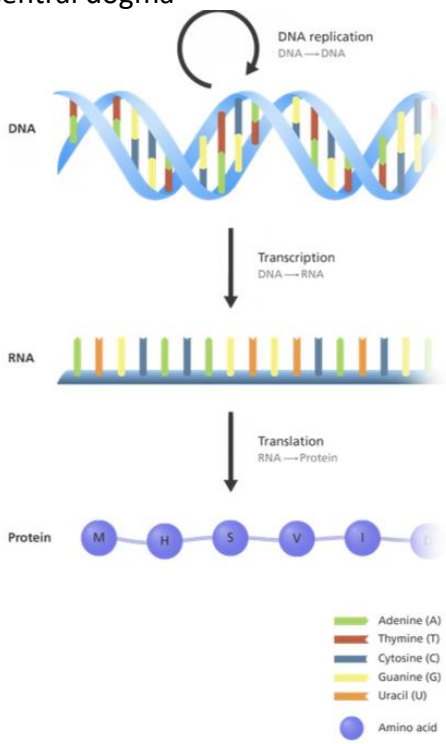
```
[9] print("The a array is ", a)

    The a array is  [1, 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 9]
```

```
[18] import numpy
     a = [1,2,3,4,4,5,5,5,6,7,8,9]
     a_mean = numpy.mean(a)
     a_std = numpy.std(a)
     a_med = numpy.median(a)
     a_max = numpy.max(a)

[19] print("The a array is ", a, "Its mean is ", a_mean)

     The a array is  [1, 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 9] Its mean is  4.916666666666667

[20] print("The a array is ", a, "Its mean is ", numpy.mean(a))

     The a array is  [1, 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 9] Its mean is  4.916666666666667
```

4. **Other Important Programming Concepts**
   - Print(): to show something on the screen, can be used for checking
   - Variable is like a box to store some data/values
   - Syntax are rules that we need to follow during programming
     - Kind of like Grammar in English

Sequence Data

1. Sequence Data Definition
   - Central dogma



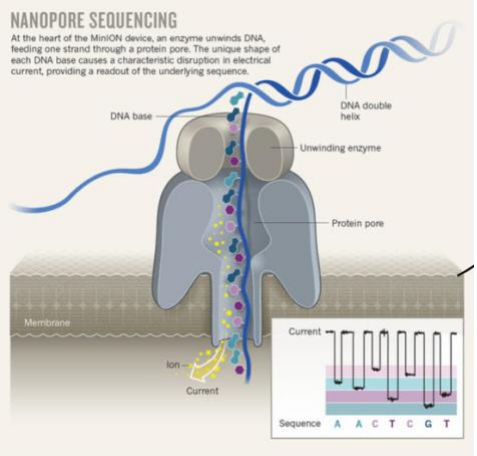
   - The genetic information in hidden in DNA sequences
   - Phenotype = Genotype (determined by sequences) + Environment

2. **Types of Sequence Data**
   - DNA sequence :
     - Composed of A,T,C,G
     - Complementary double strand

- Approximately 3 billion of these base pairs
- RNA sequence:
  - Composed of A,U,C,G
  - Single stranded
- Protein sequence:
  - Usually composed of 20 amino acids
  - Multiple sequence alignment

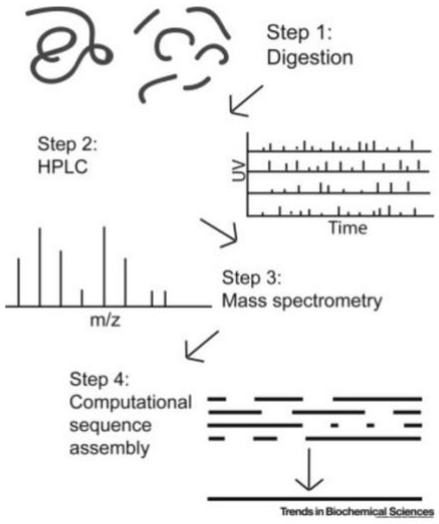
3. Obtaining Sequence Data

- DNA/RNA sequencing:
  - Still under development
  - From short reads to long reads
  - E.g Sanger sequencing, 2nd Generation sequencer (Genetic Analyzer 2), Third generation seqencer (PacBio RS),
  - Nanospace sequencer:
    - DNA goes through a chemical pore
    - Different bases -> different electrical current change
    - Sequencing by detecting current change
    - Advantages: Very long (up to 3Mb VS 1000bp)
    - Disadvantages: Error rate is high (5% VS 0.001%) ➔ Under active development

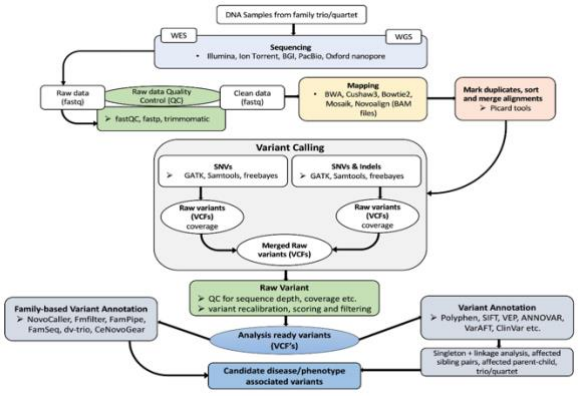


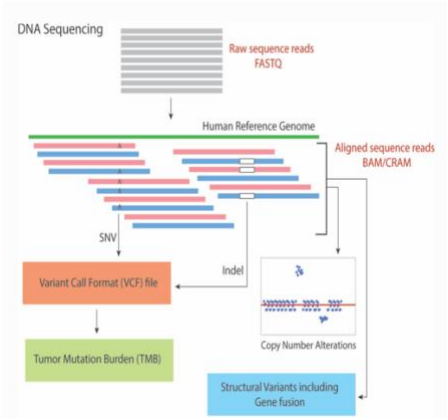
- Protein sequencing:
  - Mostly based on mass spectrometry (MS)
  - Break the long sequence into short pieces
  - Each piece can be determined by MS
  - Determined by the weight of each piece

- Assemble the short pieces into the raw sequence (like jigsaw puzzle)



4. **Making Use of Sequence Data:**
   - DNA sequences:
     - Quality check of DNA sequences to remove noises
     - Compare the sequence to the reference genome
     - Variant calling
     - Compare if the sequences are phenotype associated variants



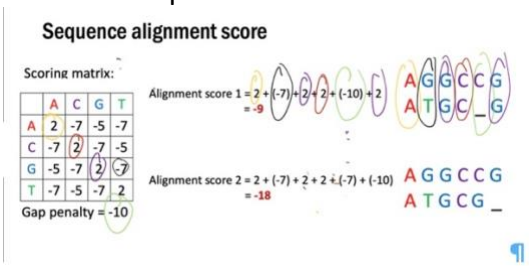


   - Protein sequences:
     - Sequence comparison
     - Multiple sequence alignment
     - Similar sequence -> Similar structure -> Similar function (The "Sequence- to-Structure-to-Function Paradigm")
     - Similar sequence -> Common ancestor ("Homology")

Sequence comparison and alignment score

1. **Sequence Alignment and Sequence Similarity**

- Sequence alignment: To determine the similarity between sequences and identify regions of similarity
- Significance:
    - Sequence-to-Structure-to-Function Paradigm, Biomolecular function and property prediction
    - Similar sequence -> Common ancestor: evolution, identifying conservative region (important region e.g histones), investigating mechanism
- Pairwise sequence alignment: Arrange two sequences to maximize the similarity between them. We can also insert gaps in the sequences
- Sequence similarity definition(score an alignment):
    - Match: A <-> A
    - Mismatch (Substitution): G <-> T
    - Gap (Insertion or deletion): C <-> _
    - Example:



    - Scoring matrix depends on cases or models

2. Finding the best pairwise alignment

- We have two sequences, scoring matrix ➔ Straightforward solution: enumeration
- Enumerate all the possible alignments between two sequences
- Calculate scores for all the alignments and select the one with the highest score (the degree of similarity)
- Problem:
- Too many possible alignments
    - e.g Align two sequences with length n:
    - Formula:
    - $\binom{2n}{n} = \frac{(2n)!}{(n!)^2}$ n = 300: $7 * 10\$\$$
    - The possible combinations is almost infinite.
    - Solution: Therefore dynamic programming is used to save time and cost for enumeration.