BMEG 3105 Fall 2025

Data analytics for personalized genomics and precision medicine

Data & Python & Sequence

Lecturer:Yu LI(李煜) from CSE

Liyu95.com,liyu@cse.cuhk.edu.hk

Friday, 5 September 2025

Student Feedback

- Students appreciated the clarity, background information, and the professor's enthusiasm.
- Some suggested more rest time, clearer rating scales, and simpler definitions for complex terms.
- General sentiment: the lecture was engaging and a good introduction to the course.

Types of Data in Biomedical and Daily Contexts

- **Biological Data:** Genes, mutations, gene expression, proteome, metabolome, molecular/cellular networks, microbiome, organs (imaging), hospital/ECG tests, demographics, medical histories, doctor notes, lifestyle, diet, family history, social media, environment, travel history.
- **Data Structures:**
    - **Sequential Data:** Ordered lists (e.g., time series, sequences).
    - **Data Matrix:** Rows = objects (e.g., people), columns = attributes (e.g., height, weight).
    - **Spatial Data:** Includes location/geography.
    - **Temporal Data:** Time-based (e.g., medical records over time).
    - **Graphs/Networks:** Objects and their interconnections (e.g., social networks).
    - **Text:** Short and long forms (from texts to documents).
    - **Multi-modality:** Combination of types (e.g., video = images + audio +

transcript; EHR = matrix + images + text).

Introduction to Python Programming

- **Programming:** Communicating with computers to perform tasks (like messaging a friend with requests).
- **Python:** The tool/language/software for this communication.
- **Libraries:** Numpy, Scipy, Pandas extend Python's capabilities for data analysis.
- **Syntax Basics:**
    - Lists use [], function calls use ().
    - print() displays output.
    - Variables store data for reuse.

Practical Python Example

- To calculate the mean:

```
import numpy as np
a = [1, 2, 3]
a_mean = np.mean(a)
print(a_mean)    # Output: 2
```

Sequence Data in Biology

- **Types:**
    - DNA (A, T, C, G; double-stranded; ~3B base pairs)
    - RNA (A, U, C, G)
    - Protein (20 amino acids)
- **Acquisition:** Sequencing technologies (e.g., short/long reads for DNA, nanopore for DNA, mass spectrometry for proteins).
    - Nanopore: long reads, higher error rates.
    - Mass spectrometry: fragments proteins, then reconstructs sequence.

Sequence Data Processing

- **DNA:** Quality control, aligning (mapping) reads to a reference genome, variant calling, phenotype association.

- **Protein:** Sequence comparison, multiple sequence alignment, function/structure prediction, evolutionary analysis.

Sequence Alignment and Scoring

- **Purpose:** Determine similarity, infer function, or evolutionary relationships.
- **Alignment:** Insert gaps as needed to maximize similarity.
- **Scoring:**
    - Matches, mismatches (substitutions), and gaps have defined scores (using a scoring matrix).
    - Example matrix: match = 2, mismatch = -5 to -7, gaps = -10.
- **Challenge:** Enumerating all alignments is computationally infeasible for long sequences.
- **Solution:** Use **dynamic programming** for efficient alignment.

Course Logistics & Resources

- Assignments use Google Colab notebooks.
- Further resources:
    - Learn Python the Hard Way
    - Colab
- Post-lecture survey for feedback.