

BMEG3105: Data analytics for personalized genomics and precision medicine

Lecture 2 Scribing [Data & Python], Sep 5 (Fri)

Outline

- Questions and comments
- Recap
- Introduction to different data types
- Introduction to Python programming and fundamental concepts
- Resource

A. Review of Comments

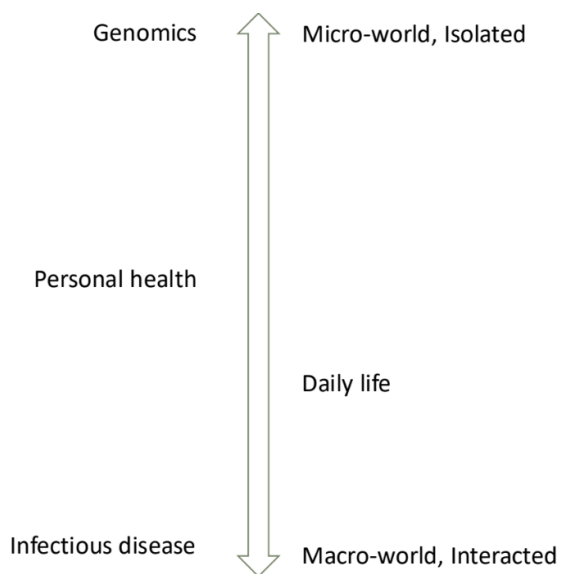
Student Feedback:

- Lecture was fun, comprehensive, and clear
- Many terms and concepts are still complex, simple definitions suggested
- Lecturer received praise for a humorous, enthusiastic approach
- Questions about the meaning of the evaluation scale

B. Recap of Biomedical Data

- Gene and mutations
- Transcriptome: gene expression
- Proteome: protein content
- Metabolome: metabolic profiles

- Molecular and cellular networks
- Microbiome: germs in oral and gut
- Biomedical imaging
- Hospital tests (blood, etc.)
- Electrocardiography (ECG)
- Demographics (age, gender, location)
- Drug and disease history
- Personal statements and doctor's diagnosis
- Exercise
- Diet
- Family history
- Communications/social media
- Environmental factors (pollution)
- Travel history (pandemic)



*** We can analyze the fundamental multiomics data (for example the genetic sequence, protein-protein interaction network, etc.) → data about personal health (for instance the patient history, lifestyle, etc.) → pandemic-related diseases (for example travel history, etc.)

Thus, data is important to interpret the connection between the isolated and interacted world.

C. Types of Data in Biomedicine

Sequential Data:

- DNA, RNA and protein sequences, etc.
 - e.g. ATTGAGGCCGATACG (the sequence of DNA is composed of A,T,G and C)

Data Matrix:

- Structured tables (n by m):

Consists of a collection of records that have a fixed set of attributes. For an n (row) by m (column) matrix, n can represent different objects while m represents the attributes. The shuffle of arbitrary row or column will not change the data itself.

- E.g., height/weight for different people

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

Spatial Data:

- Contains geographic locations and spatial information

Temporal Data:

- Data involving time

Graph/Network Data:

- Objects linked by different lines (studying the relationship between different objects), e.g., social networks, protein interaction networks

p.s. Difference of linkage definition will give various results

Text Data:

- Short phrases, sentences, or long documents

Multimodality Data:

- Data that consists of more than one above-mentioned data
 - E.g., videos (temporal images, audio, transcript) and health records (matrix, images, text)

D. Introduction to Python Programming

Programming as Communication:

- For simply definition: Programming = Communicating with a computer (a friend) to ask him/her to do some tasks.
- Code: the message that we send to the friend.

What is Python?

- Formal: Python is an interpreted high-level general-purpose programming language.
- But it is simply just like a communication software (WeChat).

Core Concepts:

We need to learn the programming language so that the computer can know what we hope them to perform. They cannot understand human language. Python will then translate the programming codes to the codes that the machine knows.

- Using plug-ins (numpy, scipy, pandas) to make Python powerful – must load them before use.
- Adopt the programming language rules. (Just like the grammatic rules in English)
- `print`: command to display values or results

Example:

- Finding the mean of some values:
 - `import numpy`
`a_mean=numpy.mean([1,2,3])`
`print(a_mean)`

p.s. The variable should be meaningful.

E. Resource to learn programming

- HW 0

<https://colab.research.google.com/drive/1cdbEwWRqw2QUtITqvBj7u5m8NRgnY89Y>

- Tutorial 1

https://colab.research.google.com/drive/18ozsp_1TOdWuqFuOACsYxZwYLJtXQqoH?usp=sharing