BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine

Lecture 3: Data & Python & Sequence

11 September 2024

Lecturer: Yu LI

Scriber: LEE, Michelle (1155179462)

1. Different data types

a. Sequential data

- Displays the sequence of nucleic acid or amino acids (DNA, RNA, and proteins).
- Example: plasmid sequence (pSB1C3 from https://parts.igem.org/Part:pSB1C3)

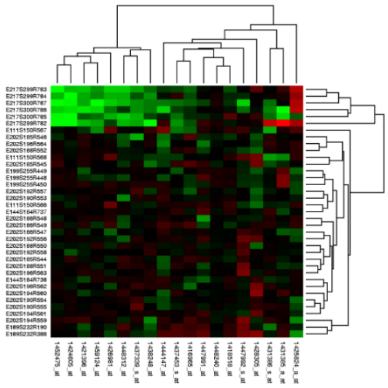
```
tactagtagcggccgctgcagtccggcaaaaaagggcaaggtgtcaccaccctgccctttttctttaaaaccgaaaagattacttcgcgttatgcaggcttcctcgctcactgactcgctgcgctcggtc
gttcggctgcggcgagcggtatcagctcactcaaaggcggtaatacggttatccacagaatcaggggataacgcaggaaagaacatgtgagcaaaaggccagcaaaaggccaggaaccgtaaaaaggccg
cgttgctggcgtttttccacaggctccgcccccctgacgagcatcacaaaaatcgacgctcaagtcagaggtggcgaaacccgacaggactataaagataccaggcgtttccccctggaagctccctcgt
gcgctctcctgttccgaccctgccgcttaccggatacctgtccgcctttctcccttcgggaagcgtggcgctttctcatagctcacgctgtaggtatctcagttcggtgtaggtcgttcgctccaagctg
ggctgtgtgcacgaaccccccgttcagcccgaccgctgcgccttatccggtaactatcgtcttgagtccaacccggtaagacacgacttatcgccactggcagcagccactggtaacaggattagcagag
cgaggtatgtaggcggtgctacagagttcttgaagtggtggcctaactacggctacactagaagaacagtatttggtatctgcgctctgctgaagccagttaccttcggaaaaagagttggtagctcttg
atccggcaaacaaaccaccgctggtagcggtggtttttttgtttgcaagcagcagattacgcgcagaaaaaaaggatctcaagaagatcctttgatcttttctacggggtctgacgctcagtggaacgaa
aactcacgttaagggattttggtcatgagattatcaaaaaggatcttcacctagatccttttaaattaaaaatgaagttttaaatcaatctaaagtatatatgagtaaacttggtctgacagctcgaggc
ttggattctcaccaataaaaaacgcccggcggcaaccgagcgttctgaacaaatccagatggagttctgaggtcattactggatctatcaacaggagtccaagcgagctcgatatcaaattacgccccgc
cctgccactcatcgcagtactgttgtaattcattaagcattctgccgacatggaagccatcacaaacggcatgatgaacctgaatcgccagcggcatcagcaccttgtcgccttgcgtataatatttgcc
catggtgaaaacgggggcgaagaagttgtccatattggccacgtttaaatcaaaactggtgaaactcacccagggattggctgagacgaaaaacatattctcaataaaccctttagggaaataggccagg
ttttcaccgtaacacgccacatcttgcgaatatatgtgtagaaactgccggaaatcgtcgtggtattcactccagagcgatgaaaacgtttcagtttgctcatggaaaacggtgtaacaagggtgaacac
tatcccatatcaccagctcaccgtctttcattgccatacgaaattccggatgagcattcatcaggcgggcaagaatgtgaataaaggccggataaaacttgtgcttatttttctttacggtctttaaaaa
ggccgtaatatccagctgaacggtctggttataggtacattgagcaactgactgaaatgcctcaaaatgttctttacgatgccattgggatatatcaacggtggtatatccagtgatttttttctccatt
ttagcttccttagctcctgaaaatctcgataactcaaaaaatacgcccggtagtgatcttatttcattatggtgaaagttggaacctcttacgtgcccgatcaactcgagtgccacctgacgtctaagaa
accattattatcatgacattaacctataaaaataggcgtatcacgaggcagaatttcagataaaaaaaatccttagctttcgctaaggatgatttctggaattcgcggccgcttctagag
```

b. Data matrix

- Collection of records; each consists of a fixed set of attributes.
- Can be represented by row(n) x columns(m) matrix, each row for each object and each column for each attribute.
- Swapping the entire column or the entire row at one time will not change the data.
- Commonly found in spreadsheets forms.
- Examples:

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

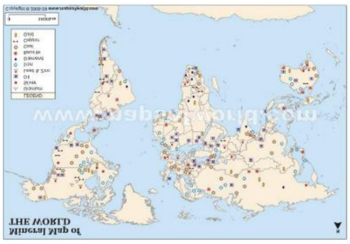
(4x2) matrix with 4 people and 2 attributes

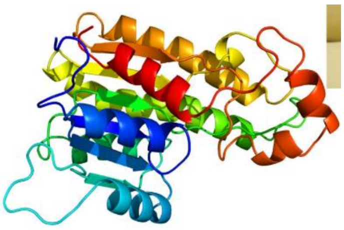


Gene expression in cells

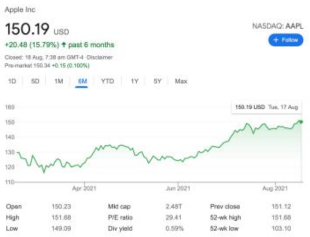
c. Spatial data

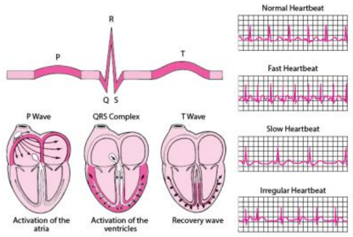
- Displays geographic locations or spatial information.
- Changing the rows or columns will change the data.
- Examples:
  - World mineral map
  - Coordinates of atoms in a protein structure (3D data)



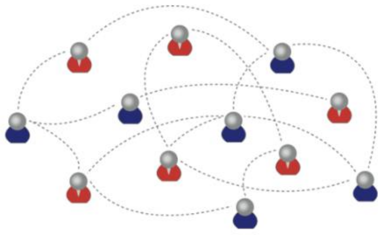


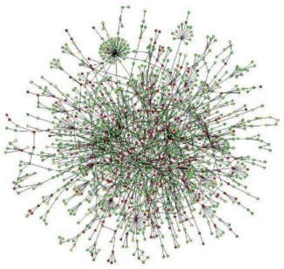
d. Temporal data
- Build-in support for data that involves time.
- Change of data over time.
- Examples:
  - Stock market graphs
  - ECG signals





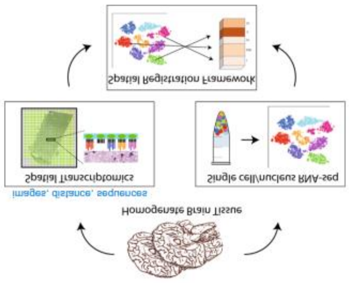
e. Graph or networks
- Displays object and its connections
- Examples:
  - Social network
  - Protein-protein interaction (PPI) networks





f. Text
- Data in text format; usually for describing something.
- Such as short and long sentences, documents.
- Examples:
  - Wikipedia articles
  - PubMed articles
  - Tweets in twitter

g. Multi-modality data
- Combinations between multiple data types.
- Examples:
  - Video; a combination of temporal images, audio, and transcript
  - Electronic health records; a combination of data matrix, images, and text
  - Spatial transcriptomics; a combination of spatial data and data matrix



h. Unknown data type
- The data has not been shown; only when data is shown, we will know the data type.
- Examples: diet and exercise.

2. Introduction to Python programming

- Programming is a **way of communicating with a computer** so that it can do something for us, and this is achieved through codes that we feed into the computer.
- Python is the **software** to send codes to the computer as well as **one of the languages** used to communicate with the computer.
- Plug-ins are available in Python to make Python more powerful; need to be loaded before its usage in the codes.

  - NumPy
  - SciPy
  - Pandas
- Code examples:
  - Calculating the mean of some values:

Codes	Returns

`import numpy`
`numpy.mean([1,2,3])`
	2.0

  - Storing values in array and calculate the mean, variance, median, max, mean, etc.:

Codes	Returns

`import numpy`
`a = [1,2,3,4,5,6,7,8,9]`
`numpy.mean(a)`
	4.916666666666667

`numpy.std(a)`	2.253084305765962
`numpy.median(a)`	5.0
`numpy.max(a)`	9
`print(a)`	[1, 2, 3, 4, 5, 6, 7, 8, 9]

  - Printing strings and variables

Codes	Returns
`print(“The a array is “, a)`	The a array is [1, 2, 3, 4, 5, 6, 7, 8, 9]

Storing values in variables:
`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)`	

Printing the stored values: `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
Printing without storing the value into a variable: `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

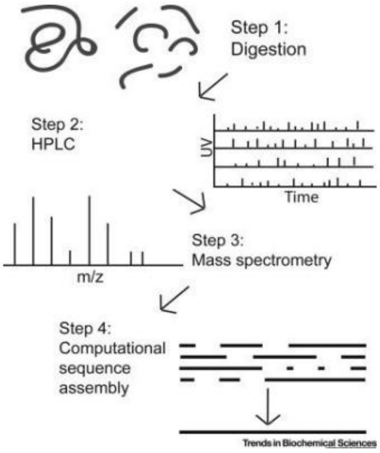
3. Sequence data

- DNA, RNA, and protein are part of the central dogma of biology.
- Genetic information is hidden in DNA sequences.
- Phenotype results from the combination of genotype (believed to be the sequences by biologists) and environmental factors.

DNA sequence	○ Composed of A, T, C, G. ○ Normally exist in double-stranded form. ○ Approximately 3 billion base pairs in the human genome.
RNA sequence	○ Composed of A, U, C, G. ○ Single-stranded.
Protein sequence	○ Usually composed of 20 amino acids.

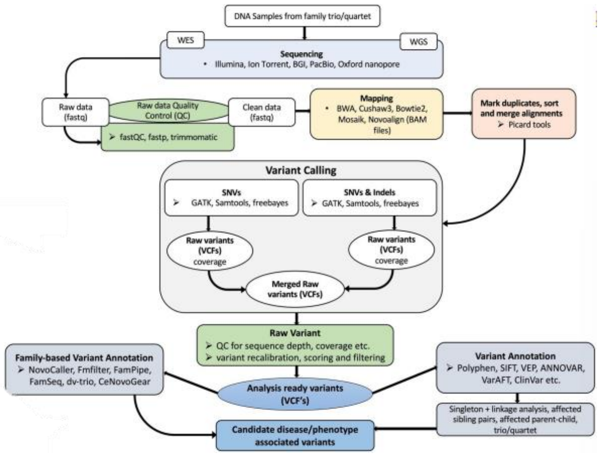
	o Multiple sequence alignment: technique to compare and analyze multiple protein sequences to study its evolution and function.

- DNA/RNA sequencing can be used to obtain short to long reads of its sequences; the technology is still under active development. Milestones in this field are the following:
    - Sanger method (1977)
    - Human mitochondrial genome sequence (1981)
    - Human genome project (1990)
    - Complete cell genome (1995)
    - Complete eukaryotic genome (1996)
    - Complete the human genome project (2001)
    - Second generation sequencer (2005-2007)
    - Research human microbiome project (2008)
    - Third generation sequencer (2011)
    - Nanospace sequencing (2014)
    - Third stage human microbiome project (2019)
- Nanopore sequencing
    - Exploits different electrical current change generated by different bases of DNA when it goes through a chemical pore.
    - Using MinION, an enzyme that unwinds DNA and feeds one strand of DNA through the protein pore.
    - Able to sequence long sequences up to 3Mb, compared to older generation technologies that can only sequence up to 1000bp.
    - The error rate is relatively high compared to the older technology (5% and 0.001% respectively).
    - Still under active development.
- Protein sequencing
    - Through the process of digestion into shorter pieces, HPLC, mass spectrometry to determine the weight, and computational sequence assembly.



Trends in Biochemical Sciences

- Raw data, such as DNA sequences, will go through quality control, followed by mapping the reads to reference genome, variant calling, and finally processed to phenotype associated variants.


DNA Samples from family trio/quartet
WES
WGS
Sequencing
• Illumina, Ion Torrent, BGI, PacBio, Oxford nanopore
Raw data (fastq)
Raw data Quality Control (QC)
➢ fastQC, fastp, trimmomatic
Clean data (fastq)
Mapping
• BWA, Cushaw3, Bowtie2, Mosaik, Novoalign (BAM files)
Mark duplicates, sort and merge alignments
➢ Picard tools
Variant Calling
SNVs
➢ GATK, Samtools, freebayes
SNVs & Indels
➢ GATK, Samtools, freebayes
Raw variants (VCFs) coverage
Raw variants (VCFs) coverage
Merged Raw variants (VCFs)
Raw Variant
➢ QC for sequence depth, coverage etc.
➢ variant recalibration, scoring and filtering
Family-based Variant Annotation
➢ NovoCaller, Fmfilter, FamPipe, FamSeq, dv-trio, CeNovoGear
Analysis ready variants (VCF's)
Variant Annotation
➢ Polyphen, SIFT, VEP, ANNOVAR, VarAFT, ClinVar etc.
Singleton + linkage analysis, affected sibling pairs, affected parent-child, trio/quartet
Candidate disease/phenotype associated variants

- Protein sequences are to be compared, for example, through multiple sequence alignment. Similar sequences correspond to similar structure and function. Similar sequences also indicate common ancestor.

4. Sequence comparison and alignment score

a. Sequence alignment and similarity

- Sequence alignment is to determine the similarity between sequences and identify the regions of similarity.
- Pairwise sequence alignment: arranging two sequences to maximize their similarity; inserting gaps is allowed.

b. Sequence alignment score

- To define sequence similarity by quantity depending on match (e.g., A with A), mismatch (e.g., G with T), or gap (e.g., C with gap).
- Using scoring matrix to calculate the score:

	A	C	G	T
A	2	−7	−5	−7
C	−7	2	−7	−5
G	−5	−7	2	−7
T	−7	−5	−7	2

Gap penalty = −10

- Examples, by enumeration:

Alignment	Score
AGGCCG ATGC_G	2 - 7 + 2 + 2 - 10 + 2 = −9
AGGCCG ATGCG_	2 - 7 + 2 + 2 - 7 - 10 = −18

- To find the best pairwise alignment, the straightforward solution will be by enumeration, calculating scores for all possible alignments and selecting the one with the highest score.
  - Problem: too many possible alignments.
  - No of possible alignments = $\binom{2n}{n} = \frac{(2n)!}{(n!)^2}$
  - For n = 300, there are 7 x 10^{88} possible alignments.
  - Solution: dynamic programming

Disclaimer: all figures are adapted from Prof. Li BMEG3105 Lecture Notes