17/09/2025,1155184178 ZHOU Wan Ieng

Recap and question:

Short reads are obtained from sequencing from the genome.

How do we get the genome sequence?

How does mapping between sequence and short reads work ?

1. Genome assembly (no reference genome available)

Why?

- *Illumina sequencing* produces **short reads** (around 200bp)
- **Human genome** is ~3 billion base pairs long.

How?

- Assembly it into longer sequence we have to find the overlap regions.

Issues:

Repetitive regions (e.g., AAAA...), it becomes **ambiguous** where a read should be placed.

This create an overlapping issue, multiple possible reconstruction ways exist :

AAAAAAAA [sequence 1]

AAA [seq. 2]

AAAAA[seq.3]

AAAAAAAA AAAAAAAA

AAA or AAA

AAAAA AAAAA

Repeat overlapping region, can create **assembly errors.**

Although accuracy is relatively high, but considering the large amount of sequencing needed to be done to reconstruct, still possible for repeat error.

Solutions:

long-read sequencing technologies, effectively reducing the amount of repetitive regions and possible errors.

2. Mapping (reference genome available)

In a genome, there are regulatory regions and coding regions.

We want to use mapping on coding region to infer DNA expression level (estimate, not exact), this can be DNA / RNA seq (gene expression quantification).

How?

mRNA by reverse transcription into cDNA , then sequenced. Amount of cDNA read map to real genome is a representation of the gene expression level, all examples in the slides are simplified only to using DNA complementary base pairing rule.

- Gene expression count (photo)

Common issue:

- Repetitive sequencing: common when short map onto long
  2 possible locations for the same match
- Sequencing error / real mutation causing mismatch

Data cleaning

Types of data:

❖Sequential Data

❖Data matrix: Special property: shift whole row / column, data information and meaning will not be changed.

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

❖Spatial data

❖Temporal data

❖Graph or networks

❖Text

❖Multi-modality data

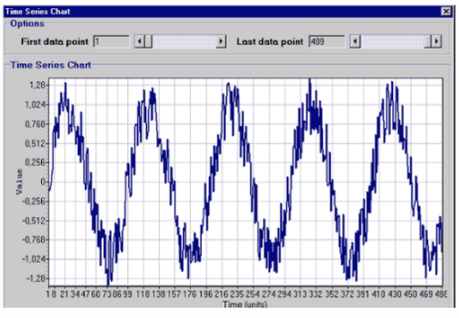
➢Video: Temporal images, audio, transcript ➢Electronic health records: Data matrix, images, text ➢Spatial transcriptomics: Spatial data, sequence, data matrix

Cleaning pipeline (orders matter!!)

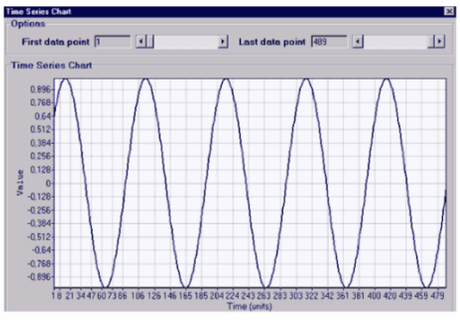
❖Denoise data (if applicable)

Raw data require cleaning because they include noise (modification of original value)

Averaging and smoothing and usually solve this issue.



A sine wave with noise



The denoised sine wave

❖Remove outliers : **significantly different from most of the other values** in a dataset, does not fit in the current pattern.

*However, whether ignore can be decided case by case since some outliers can be true and informational.

❖Handling missing data

➢Eliminate Data Objects

➢Estimate Missing Values (eg. By taking the mean)

➢Ignore the Missing Value During Analysis

➢Replace with all possible values (weighted by their probabilities)

❖Remove duplicates : may affect by dominating and change distribution of the entire dataset

❖Categorical data encoding: one hot encoding

❖Data normalization: To put attributes on the similar level of measurement, in order to make fair comparisons

1. Max / Min
2. Z score (Gaussian distribution):

Issue: Normalization direction can usually be performed across both column and row , first need to define what question is being asked.

Comparing different gene

Standardizing and eliminating variance across different samples

Data exploration

- Summary statistic : Frequency, location, spread

1. Mean: sum of all data values divided by the number of values
   **sensitive to outliers
2. Median: middle value when data are sorted
3. Range: max -min
4. Variance : average squared deviation from the mean

Eg. [2, 3, 4], mean = 3 → variance = [(2−3)² + (3−3)² + (4−3)²] / (3−1) = (1+0+1)/2 = 1.

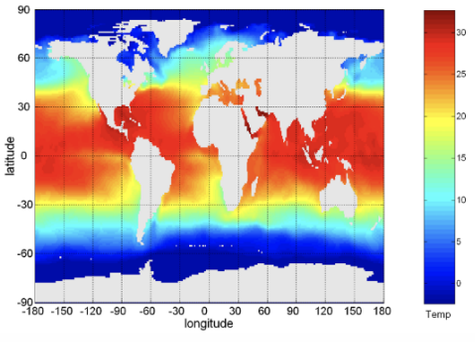
5. Percentiles: p-th percentile is a value of x such that p% of the observed values of x are less than $x\,p$

** Interquartile range: 25th percentile = Q1/ 50th percentile = Median/ 75th percentile = Q3

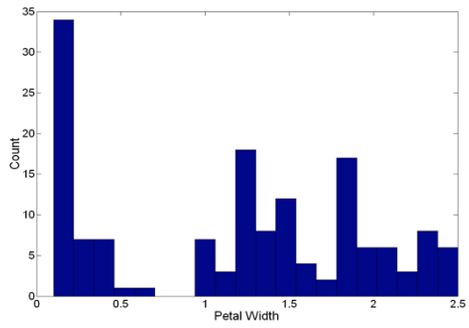
6. Frequency: % of time a specific value occurs
7. Mode : the most frequent value

Visualization (taking data matrix as an example)

1. Single figure eg. Sea surface temperature, before visualization this is a triplet, difficult to observe trend directly

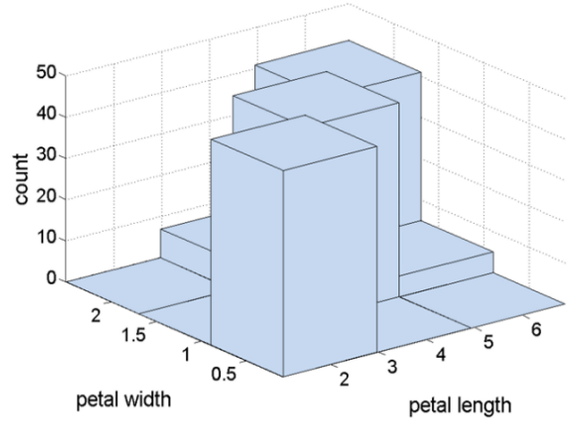


2. Histograms



3. 2D histograms (2 attributes= joint distributions)

Question: What does this tell us?



4. Box plots