Scribing by SAHA, Anujit (SID 1155163076)

# BMEG3105 Fall 2023 Lecture 6: Data exploration and data cleaning (taught on 23rd September, 2023)

## 1.Notable questions and comments from previous and current lecture

- Mostly positive response about the previous lecture clarifying dynamic programming
- Requested more genome assembly practice – Yu's response was to follow the recommended reading for more examples
- Requested more details for scRNA sequencing- responded that this topic is to be covered in details in the later lectures
- Video recordings were requested- professor commented coming to class would be the safest way to learn

## 2.Recap from last lecture

- Topic transitioning from sequence data to data matrix
- Our outcome is RNAseq to be able to obtain a gene matrix with a gene count within different samples
- For genome assembly method, we MUST include both the overlap and non-overlapping region of the short reads with the reference genome as illustrated below; both needed to be able to extend the length (referred back to the jigsaw puzzle that is an example of assembly)

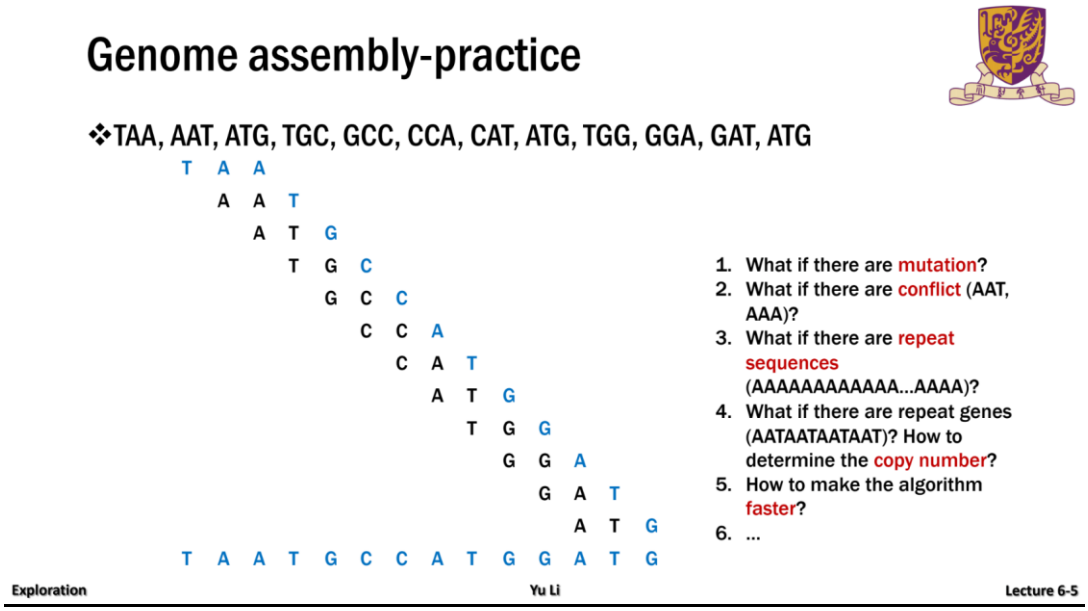


Figure 1

- **Mapping** is done by sliding each read along the genome, and calculating the difference (can be optimized by using dynamic programming); an example of performing it manually is shown below

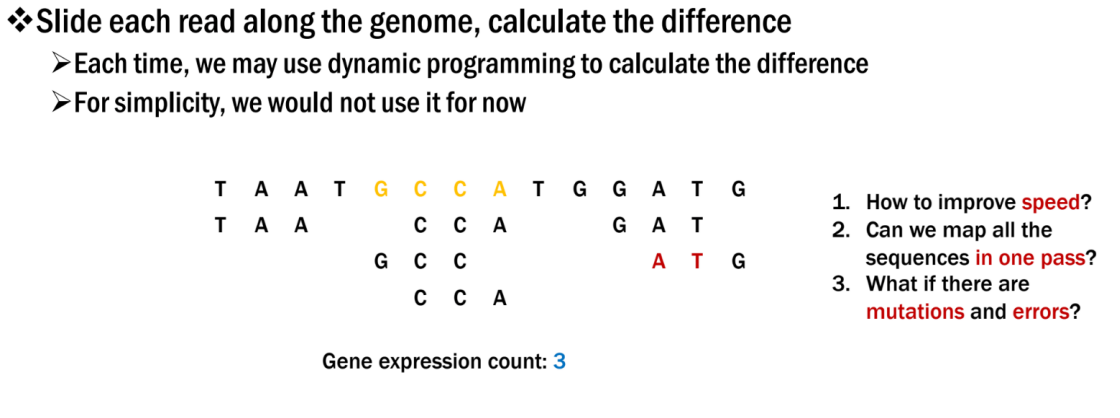



Figure 2

- The mapping can also be achieved through advanced techniques of mRNA isolation followed by Illumina sequencing for faster and more efficient conversion to the data matrix from sequencing data

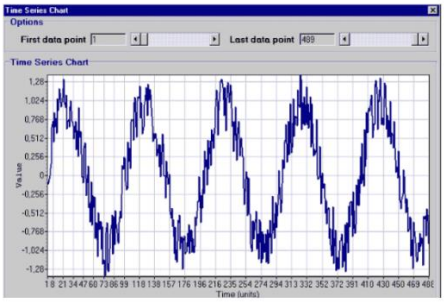
## 3.Today's agenda

### ❖ Data Cleaning

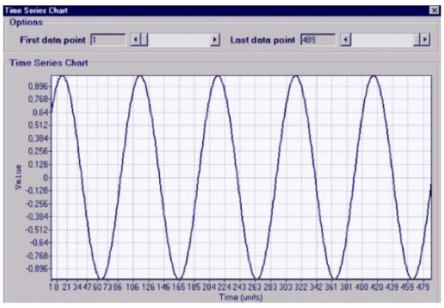
- **Data matrix review**
  - Data that consists of a collection of records, each of which contains a fixed set of attributes
  - It is usually expressed as an n rows by m column matrix

**What kinds of data quality issues we might face? Listed and explained below**

1) **Noise**: It is the modification of original values due to some background signal and interference. Figure 3 shows an example of a sine wave with noise(sharp edges) while a denoised(or attenuated) wave is at the right



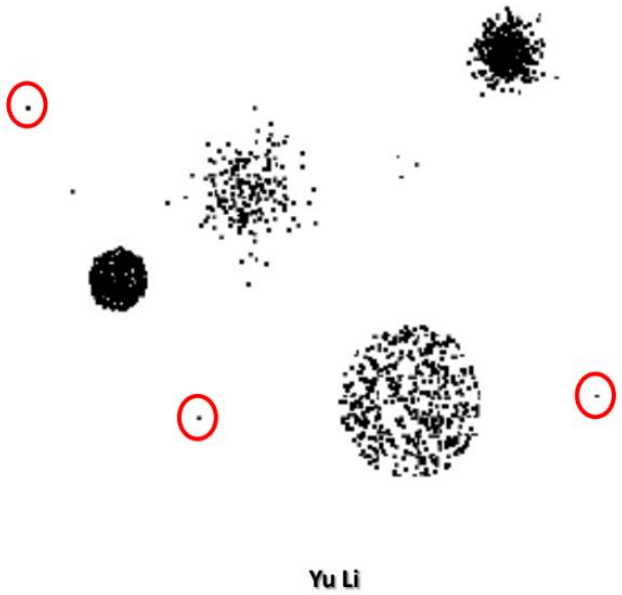

A sine wave with noise




The denoised sine wave

Figure 3

2) **Outliers**: data objects that are considerably different than most of the other data objects in the data set( for eg the below figure shows in red values in clustered data that are clearly outside the similar clusters( will be discussed more in the next lecture)



Yu Li

Figure 4

3) **Missing values**: caused due to information not being collected or the data type is not applicable to a certain demographic

To be able to analyze the data set, we need to be able to handle the missing values for proper statistical analysis

Some ways include :eliminate data objects, estimate missing values, ignore the missing value during analysis and replace with all possible values based on their weighted probabilities

4) **Duplicate data**: exists especially when merging different databases and cause a duplication of the same data values

5) **Unnormalized data**: an issue arising from attributes on the data matrix belonging to different level of measurement; an example the range of height and weight range being vastly different within the dataset below and hence weight values dominate the difference in attributes between the persons

Usually solved by two methods as illustrated below:

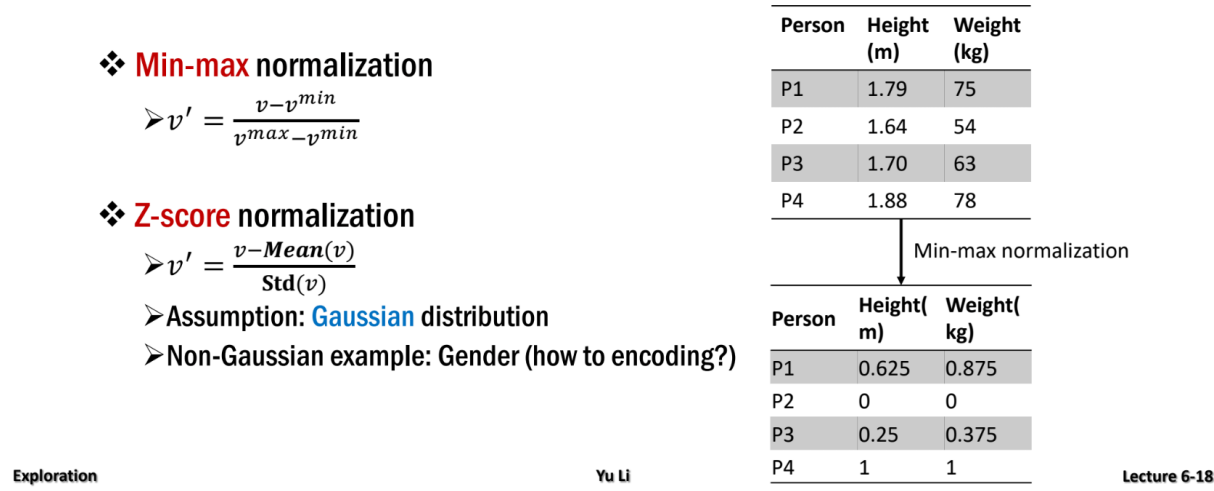



Figure 5

Both normalizations depend on the idea of finding the distance of each entry from the expected value; Min-max normalization leads to a value between 0 and 1 to exist in all entries.

We also need to decide whether to normalize along each row or each column

- As discussed in the A1 assignment set, direction depends on the requirement of the analysis; a sample wise normalization can help us deduce if the difference

across conditions are significant; where as gene wise normalization can help us deduce the prevalence of of each gene universally along the samples used

6) **Categorical data**: Data that do not have a numeric value; they are often converted to either 0 or 1 based on **one-hot encoding** as shown below

## Categorical data

| Person | Height(m) | Weight(kg) | Gender |
|---|---|---|---|
| P1 | 0.625 | 0.875 | Male |
| P2 | 0 | 0 | Female |
| P3 | 0.25 | 0.375 | Female |
| P4 | 1 | 1 | Male |

⇒

| Person | Height(m) | Weight(kg) | Male | Female |
|---|---|---|---|---|
| P1 | 0.625 | 0.875 | 1 | 0 |
| P2 | 0 | 0 | 0 | 1 |
| P3 | 0.25 | 0.375 | 0 | 1 |
| P4 | 1 | 1 | 1 | 0 |

Computers are better on handling numbers
For categorical data, we can use one-hot encoding

Figure 6

## ❖ Data Exploration

**Summary statistics** refers to numbers that summarize properties of the data such as:

1) Measures of location: mean is the most common measure of the location of a set of points defined as:

$mean(x) = \frac{1}{m}\sum_{i=1}^{m} x_i$

**We have to note that is mean is very sensitive to outliers**

The median is another measure of the central location defined as:

$$median(x) = \begin{cases} x_{(r+1)} & if\ m\ is\ odd, i.e.,\ \ m = 2r+1 \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & if\ m\ is\ even,\ \ i.e.,\ m = 2r \end{cases}$$

2) Whereas, range is the difference between max and min; but a even better measure of the range is variance defined as:

$$variance(x) = \frac{1}{m-1}\sum_{i=1}^{m}\left(x_i - mean(x)\right)^2$$

**Such variance calculation is often sensitive to outliers as well, for which we introduced more measures such as:**

**Median absolute deviation (MAD)**

- $median(|x_1 - mean(x)|, \dots, |x_m - mean(x)|)$

**Interquartile range**

- $x_{75\%} - x_{25\%}$

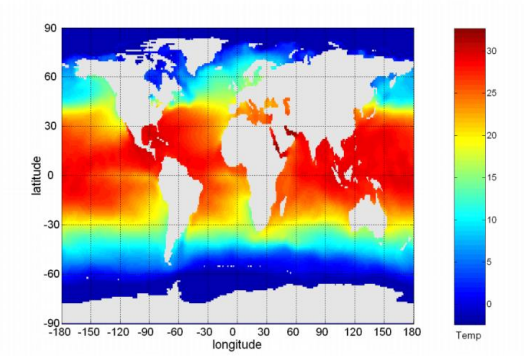
3) Percentiles: For an ordinal continuous attribute x and a number p between 0 and 100, the p-th percentile is a value of x such that p% of the observed values of x less than xp

   Emphasis given by professor regarding noting that percentile is defined as p% of the observed values of x **less than xp; need to be careful when defined that x is higher than xp**

4) Frequency and mode: frequency registers the percentage of time a particular value occurs within a data set; whereas the mode of an attribute is the most occurring attribute value; they are useful when dealing with categorical data

Exploratory visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and relationships among data items can be analysed

An example below illustrates how visualization can help us detect changes in temperature along the latitudes and longitudes in a manner that is much more efficient than viewing each attribute in a csv file




**Exploration** **Yu Li** **Lecture 6-33**

Figure 7

An example of a common visualization technique is **histogram** that can show the distribution of values of a single variable by diving the values into bins(intervals)- height of each bar indicates the number of objects whereas shape of the histogram depends on the number of bins. A major advantage of this technique is the ability to spot outliers- as seen below the value 11 for height is an outlier as it occurs for a very low frequency compared to other values

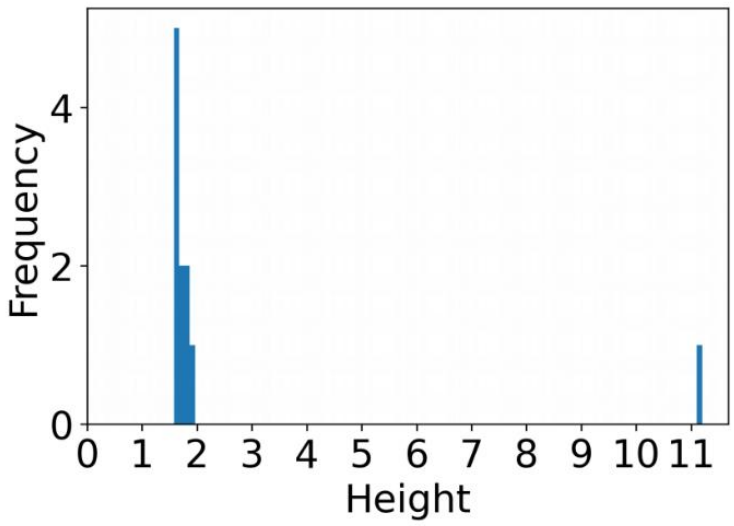



Figure 8

Histograms can be extended to **Two-Dimensional histograms** to show the distribution/ relation between two different attributes. The figure below shows that with petal length, petal width increases as well.

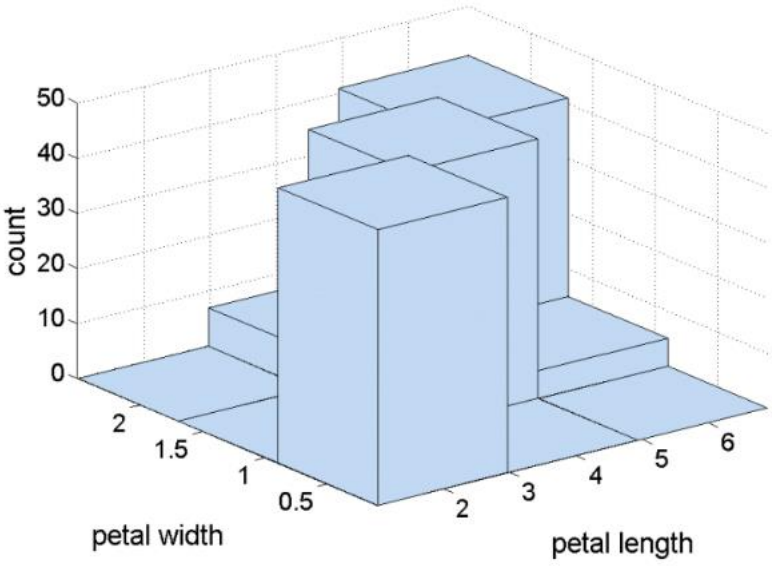



Figure 9

Another way of displaying of data is the **Box Plots** which clearly defines the median, the 75$^{th}$ and 25$^{th}$ percentiles and the min and max of a particular attribute as illustrated below. It helps us compare along the different attributes of a data matrix and helps determine the skew and interquartile range

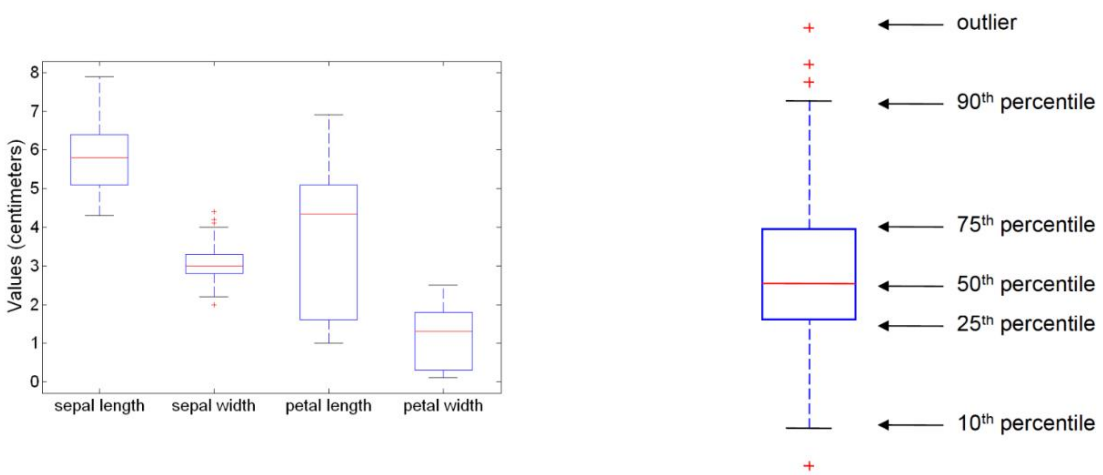



Figure 10

Data exploration example was given for further practice:
https://colab.research.google.com/drive/1z0_IEP-tOZgt7auZqEMtLvmafVuHRP0d?usp=sharing

The next lecture will be dealing with more data visualization techniques such as clustering.