Data pre-processing in R

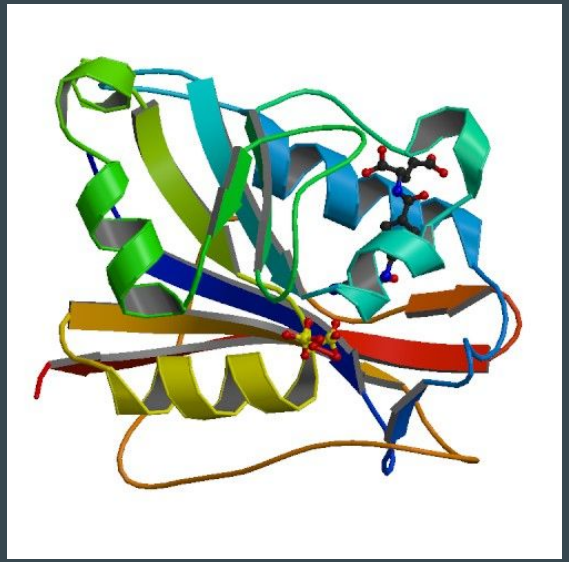
Handling missing data

...

Associate Professor Dr. Chanin Nantasenamat

DHFR dataset

- **DHFR** = **Dihydrofolate reductase**
- Dataset comprises of **325 compounds (rows)** and **229 variables (columns)**
- Of the **229 variables**, one is the target variable **Y** which represents the biological activity. It is classified as being **active** or **inactive**
- Of the **229 variables**, the remaining 228 variables are the molecular descriptors that describes their physicochemical properties (e.g. charge, molecular connectivity, solubility, etc.)



What we will learn today?

1. Loading the DHFR data
2. Check for missing data
3. If data is clean, randomly introduce NA to the dataset
4. Check again for missing data (Hint: there should now be missing data)
5. Handling the missing data. There are 2 options, decide and choose only 1
    a. Simply delete all entries with missing data
    b. Imputation
        i. Replace missing values with the column's mean
        ii. Replace missing values with the column's median

Let's get started!

Fire up your RStudio or RStudio.cloud