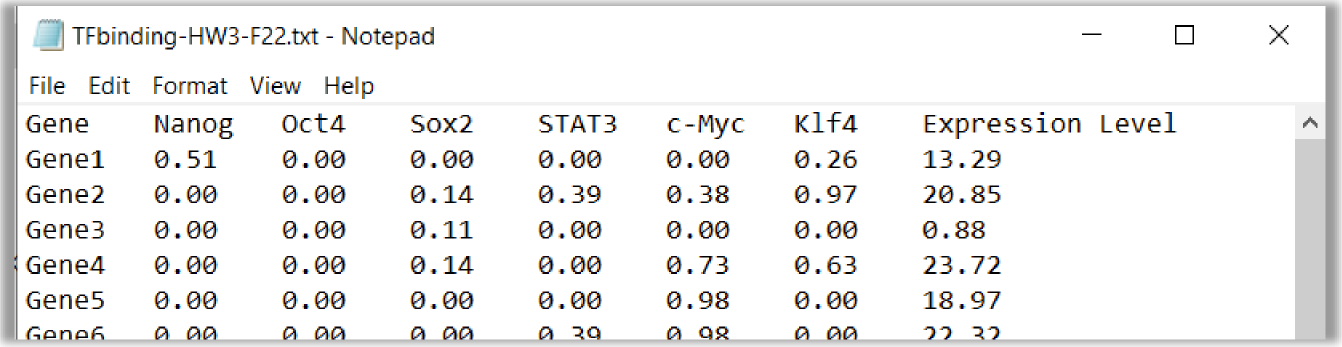
BENG215/BSB501/MBG624 – Biocomputing – Python HW3 – Jan 12th, 23:30, 2023

Transcription factors (TFs) are proteins that bind DNA. They bind to promoter regions of genes on the DNA and activate (or sometimes suppress) expression of the gene. A researcher collected data on binding level of several TFs (Nanog, Oct4 …) on promoters of genes (Gene1, Gene2 …) and expression level of the genes. The researcher believes that there is a linear relationship between TF binding level and expression level and would like to build a linear model to test this hypothesis.

The data is given in the TAB limited text file **TFbinding-HW3-F23.txt**



```
Gene    Nanog   Oct4    Sox2    STAT3   c-Myc   Klf4    Expression Level
Gene1   0.51    0.00    0.00    0.00    0.00    0.26    13.29
Gene2   0.00    0.00    0.14    0.39    0.38    0.97    20.85
Gene3   0.00    0.00    0.11    0.00    0.00    0.00    0.88
Gene4   0.00    0.00    0.14    0.00    0.73    0.63    23.72
Gene5   0.00    0.00    0.00    0.00    0.98    0.00    18.97
Gene6   0.00    0.00    0.00    0.39    0.98    0.00    22.32
```

Write a Python script that does the following analyses and charts:

1. (15 points) Use pandas or basic file operations to read the data from the file and keep them as numerical values in lists, arrays or data structures.
2. (10 points) Calculate the mean and standard deviation of the expression levels.
3. (10 points) Calculate the total binding score for each gene (Nanog + Oct4 + … + Klf4). Calculate the mean and standard deviation of the total binding scores.
4. (15 points) Draw two histograms as subplots. The first one is a histogram of the expression levels, and the second one is a histogram of the total binding scores. Use 10-20 bars and write the mean and standard deviation on the chart as a text or title.
5. (20 points) Fit a linear equation to predict gene expression level as a function of total binding score. Fitting a line means predicting values of a and b in the following equation:
   Expression_Level = a + b * Total_binding_score
   (this equation will also predict the expression level for each gene using the predicted a and b values)
6. (10 points) Predict the expression level of each gene by the linear equation obtained in step 5, then calculate the prediction error for each gene as follows:
   Error = True_expression_level – Predicted_expression_level for a gene
7. (20 points) Plot your results as follows:
   - Total binding scores should be on the x-axis.
   - True expression values (circles), predictions (line), errors (asterix) should be on the y-axis
   - Display the fitted equation on the chart
   - Add a legend to your chart.

5 bonus points for writing reusable functions.

5 bonus points for nice formatting of the charts (titles, labels etc).

5 bonus points for explanatory comments and docstrings in your script.

Every step has to be done in Python programatically. You can use any module or function, even if it was not covered in the lectures. If you manipulate the data manually, you will lose points of that section.

Good luck!