

# Biomedical Data Science 2025: Homework

## Assignment 1

**Due: March 26th (Wednesday), 11:59pm EST**

Choose to do either **MCDB & MBB (non-programming)** or **CBB & CS & S&DS (programming)** assignment, depending on your academic affiliation. No late submissions will be accepted. Submission should be done in Canvas.

### MCDB & MBB (Non-Programming)

1. **(25 pts)** Multiple sequence alignments (MSA) cannot be efficiently handled using purely dynamic programming. Choose one existing MSA software and describe how it implements MSA.  
(For example: Muscle, ClustalW, Kalign, MView, T-Coffee, etc.)

2. **(25 pts)** ChIP-seq is a common method to determine protein-DNA interaction on a genome-wide scale. The exact sites of binding must be inferred from sequence reads of the DNA that is purified along with the protein of interest. Describe an algorithm for determining protein-DNA binding sites from ChIP-seq data.

See the following citation for a list of example algorithms:

*Wilbanks, EG, Facciotti, MT (2010). Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One, 5, 7:e11471.*

3. **(25 pts)** Machine learning approaches have become extremely useful in the analysis of biological data. Read the paper referenced below and answer the following questions:

*Ghandi, Mahmoud, et al. "Enhanced regulatory sequence prediction using gapped k-mer features."*

**PLoS Comput Biol** 10.7 (2014): e1003711.

[Paper Link]

- What are the researchers trying to predict/infer?
- What information is being used for the prediction? What is the logic behind using these data?

- What preprocessing steps are used to prepare the data for machine learning?
  - What is the model the researchers use?
  - How do the researchers evaluate their predictions? Were they effective? What biological insight was gained?
4. **(25 pts)** Answer the above questions for question 4 for the following paper:

*"Building a Hybrid Physical-Statistical Classifier for Predicting the Effect of Variants Related to Protein-Drug Interactions."*

**Structure.** 2019 Sep 3;27(9):1469-1481.e3. doi: 10.1016/j.str.2019.06.001.  
[Paper Link]

## CBB & CPSC & S&DS (Programming)

This year's programming assignment is provided in an `.ipynb` file, which you can find on the course website or in Canvas files. We highly recommend running it on Google Colab. To do this:

- Download the `.ipynb` file for this assignment.
- Go to Google Colab.
- Upload the `.ipynb` file.
- Further instructions can be found in the notebook once you upload and start working on it.
- When finished, download your notebook as an `.ipynb` file (File → Download → Download `.ipynb`), and submit it on Canvas.