Case Study Overview 5100 by Tareque Muzahid Zamee and Md Mehadi Hasan Sohag

**Using ChatGPT to develop the functional R code for the visualization of proteomic data volcano plot**

Volcano plots are used to measure the statistical significance of experiments comprising large datasets of replicated data. Our first research question is “What is the purpose of a proteomics data volcano plot?”. The volcano plots are widely used in proteomics studies to identify significantly differentially expressed proteins that are synonymously known as proteomics data volcano plots. Then, our next research question is “How close the chat GPT is providing the functional R code to create the actual proteomic data volcano plot?” To develop this case study idea, we chose an article entitled “Total Proteome Analysis Identifies Migration Defects as a Major Pathogenetic Facto in Immunoglobulin Heavy Chain Variable Region (IGHV)-unmutated Chronic Lymphocytic Leukemia” by Eagle et al., (2015). As the proteomics data of this study are publicly available as well as a biochemistry web blog (R for Biochemists) generated the original code to visualize this data volcano plot shared at <https://rforbiochemists.blogspot.com/2015/06/drawing-proteomic-data-volcano-plot.html>, we saved the original R code and validate the code in R console (R Core Team 2024). At first, we were asking ChatGPT for the functional code to generate the proteomics data volcano plot by feeding the dataset. We asked for different prompts as well as the specified prompt regarding the functional code over and over again and saved all the responses of the ChatGPT into a .txt file. Besides, the code sequences provided by the ChatGPT were saved into a .R file. To make the case study idea successful, we had to ask ChatGPT through verbatim prompts sixteen times and got the functional code in our last attempt. All the required files are shared at <https://github.com/tmzamee/Bio-5100-case-study>. In conclusion, we can say that though it takes time to get the functional R code by ChatGPT, providing feedback with exact error responses could be a convenient solution for getting the original R code. So, this case study demonstrates that elaborative and specific verbatim prompts can help an inexperienced coder to get the functional code by using the large language model (ChatGPT).

**References**

1. Eagle, G. L., Zhuang, J., Jenkins, R. E., Till, K. J., Jithesh, P. V., Lin, K., ... & Pettitt, A. R. (2015). Total Proteome Analysis Identifies Migration Defects as a Major Pathogenetic Factor in Immunoglobulin Heavy Chain Variable Region (IGHV)-unmutated Chronic Lymphocytic Leukemia\*[S]. Molecular & Cellular Proteomics, 14(4), 933-945

2. Alsagaby, S. A., Khanna, S., Hart, K. W., Pratt, G., Fegan, C., Pepper, C., ... & Brennan, P. (2014). Proteomics-based strategies to identify proteins relevant to chronic lymphocytic leukemia. Journal of proteome research, 13(11), 5051-5062

3. <https://rforbiochemists.blogspot.com/2015/06/drawing-proteomic-data-volcano-plot.html>

4. R Core Team. 2024. R: A Language and Environment for Statistical Computing. from <https://www.r-project.org/>

5. Ye, H., Shen, X., Li, Y., Zou, W., Hassan, S. S. U., Feng, Y., ... & Zhu, W. (2024). Proteomic and metabolomic characterization of bone, liver, and lung metastases in plasma of breast cancer patients. PROTEOMICS–Clinical Applications, 2300136.

6. Jenkins, S. F., Magill, C. R., & McAneney, K. J. (2007). Multi-stage volcanic events: A statistical investigation. Journal of Volcanology and Geothermal Research, 161(4), 275-288.

7. Perkel, J. M. (2023). Six tips for better coding with ChatGPT. Nature, 618(7964), 422-423.