Predicting Possible Oligomerization States of Protein Sequences

Agney K Rajeev & Joel Joseph K B

- **Idea:** We attempt to develop an ML algorithm that predicts possible oligomerization states given a FASTA sequence of a particular protein chain.
- **Dataset:** We curate our own dataset by scraping the RCSB Protein Data Bank
- **Relevant Papers:**

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. a. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
2. Shen, H., & Chou, K. (2009). QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. Journal of Proteome Research, 8(3), 1577–1584. https://doi.org/10.1021/pr800957q
3. Simeon, S., Shoombuatong, W., Anuwongcharoen, N., Preeyanon, L., Prachayasittikul, V., Wikberg, J. E. S., & Nantasenamat, C. (2016). osFP: a web server for predicting the oligomeric states of fluorescent proteins. Journal of Cheminformatics, 8(1). https://doi.org/10.1186/s13321-016-0185-8

- **Work Divison:**
  - Agney K Rajeev: Literature review on various ML implementation, preprocessing of data, database cleaning
  - Joel Joseph KB: Literature review on Protein oligomerization, feature extraction, and identifying important features in training data
  - Both: Database curation, Slides, Reports, and Implementation of Algorithms
- **Algorithms to be implemented:**
  - k-Nearest Neighbour, Random Forest and Neural Network
- **Midway Targets:**
  - Data curation and organization
  - Research about possible embedding and preprocessing of data
  - Implementation of 1 or 2 above-mentioned algorithms
- **Expected Results:**
  - We expect improved results compared to the existing models in the literature due to our extensive database($\sim$150,000) and advanced algorithms implemented.