Retraining the ADOPT Model Using ESM2 Embeddings

Background: Some proteins or protein regions do not fold into a fixed 3D structure but remain flexible or “disordered.” These intrinsically disordered regions play key roles in regulation, signaling, and molecular interactions. Furthermore, novel proteins that do not share ancestry with existing proteins (*de novo* proteins*)* are likely to be disordered. Understanding the determinants of disorder would be a key to understanding the evolution of these proteins.

ADOPT (Redl *et al.*, 2023) is a linear (lasso) regression model that predicts intrinsic protein disorder using embeddings (numerical representations of protein sequences) from the ESM1b protein language model. It was trained using a set of protein sequences whose structures were analysed using NMR (CheZOD, Nielsen and Mudler, 2019). ADOPT showed that contextual embeddings alone can capture key biophysical properties.

The next generation of these models, *ESM2*, provides improved sequence representations and broader coverage of protein sequence space. Retraining ADOPT using ESM2 embeddings may improve the accuracy and generalization of intrinsic disorder predictions.

Objective: The student will retrain the ADOPT model using ESM2 embeddings instead of ESM1b and compare the predictive performance. The project will assess how newer embeddings influence linear predictability of disorder and whether certain ESM2 model sizes or layers perform better.

Methods:

- Generate per-residue embeddings using ESM2.
- Retrain the lasso regression model to predict experimental disorder scores.
- Compare results with the original ESM1b-based ADOPT model using performance metrics such as Pearson correlation and RMSE.

Requirements:

- Interest in protein structure and bioinformatics.
- Basic experience in Python.

Supervision:
Dr. Bharat Ravi (b.ravi@uni-muenster.de), Molecular Evolution and Bioinformatics Group, IEB

Relevant literature:

- Redl *et al.* (2023) *ADOPT: intrinsic protein disorder prediction through deep bidirectional transformers. NAR Genomics and Bioinformatics* 5(2): lqad041 .
- *Nielsen J. T., Mulder F. a. A. (2019). Quality and bias of protein disorder predictors. Sci. Rep. 9, 5137. 10.1038/s41598-019-41644-w*
- Lin Z. *et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science*. 17;379(6637):1123-1130.