Prediction of membrane protein types using protein language models - Project plan

Danmarks Tekniske Universitet - DTU
Department of Health Technology, Bioinformatics

Marius Thrane Ødum, MSc. Bioinformatics & Systems Biology

Henrik Nielsen (Assoc. Prof., supervisor)
Felix Teufel (PhD AI Novo Nordisk, co-supervisor)

September 21st, 2022

Motivation and aim of the project

The goal of this project is to develop a multi-class prediction tool to determine if a protein is membrane-bound or not, and if so, which of the following membrane-bound classes it likely belongs to; *transmembrane*, *lipid-anchored* or *peripheral*. Membrane proteins are highly associated with a large range of important cell functions. Compartmentalization of organelles and cell structures through membrane separation allows cells to control which interactions that can take place between exterior proteins/signal molecules and intracellular macromolecules. As a consequence of these selective interactions, membrane proteins displays a huge role in cell communication, internal gene regulation in response to chemical messengers, membrane transportation and various other functions. Due to these properties, membrane proteins are essential for understanding and determining cell functionality and biological pathways, and therefore also continues to be prime targets for drug development. The importance of developing and attaining accurate predictors to identify and classify novel membrane proteins can therefore not be underestimated for the reasons mentioned. For an overview and estimation of time management for the duration of the project the associated tasks can be split into the following phases.

Data curation: Collection of data containing protein sequences from UniProtKB with experimentally annotated data regarding subcellular localization for training and testing. The dataset includes appx. 11,600 and 20,500 sequences for the membrane-bound and soluble classes respectively.
Model architecture: Design of base model including pretrained protein *language models* (LM) (e.g. ESM-1b/2, ProtT5, ProtBERT or ProGen2).
Training and testing model: Evaluating the various LMs and optimization of the base model architecture can be considered part of this phase.
Model comparison: Compare the performance of the developed model to other state-of-the-art membrane class predictors. These include the models of the following publications *iMem-Seq* (Xiao et al. 2015), *HHT + SVM* (Han et al. 2014), *VFI* (Farman et al. 2015), *Mem-ADSVM* (Shibiao et al. 2016), *Toot-M* (Alballa et al. 2019) and *MKSVM* (Wang et al. 2019)
Integration with DeepLoc 2.0: If time allows and the project shows promising results, the aim and final task of the project will be to fully integrate the developed tool into the existing DeepLoc 2.0 predictor.

Time management

22. aug - 23. sep: Curated dataset finalized, Overview of literature
20. sep - 21. oct: Design standard architecture of model for membrane type predictions and implement ESM and ProtT5 into it (additionally ProtBERT, ProGen2 or other protein LMs might be implemented)
21. oct - 18. nov: Train/evaluate models - set up different experiments (eukaryotes vs. prokaryotes vs. combined)
19. nov - 10. dec: Other exams/final course projects - ”Break” from thesis
10. dec - 13. jan: (Continued) Optimize experiment with standard architecture/Train and evaluate models and set up different experiments (eukaryotes vs. prokaryotes vs. combined)
14. jan - 27. jan: Implement and compare performance to other state-of-the-art
28. jan - 28. feb: Main writing period
1. mar - 20. mar: Integration with DeepLoc2.0 (optionally, if time allows)
21. mar - 9. apr: Buffer
10. apr: Deadline thesis hand-in

Bibliography

Alballa, Munira and Gregory Butler (2019). “Integrative approach for detecting membrane proteins”. In: *BMC Bioinformatics* 21.

Farman, Ali and Hayat Maqsood (2015). “Classification of membrane protein types using Voting Feature Interval in combination with Chou’s Pseudo Amino Acid Composition”. In: *Journal of Theoretical Biology* 384, pp. 78–83.

Han, Guo-Sheng, Zu-Guo Yu, and Vo Anh (2014). “A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou’s PseAAC”. In: *Journal of Theoretical Biology* 344, pp. 31–39.

Shibiao, Wan, Mak Man-Wai, and Kung Sun-Yuan (2016). “Mem-ADSVM: A two-layer multi-label predictor for identifying multi-functional types of membrane proteins”. In: *Journal of Theoretical Biology* 398, pp. 32–42.

Wang, Hao, Yijie Ding, Jijun Tang, and Fei Guo (2019). “Identification of membrane protein types via multivariate information fusion with Hilbert–Schmidt Independence Criterion”. In: *Neurocomputing* 383, pp. 257–269.

Xiao, Xuan, Hong-Liang Zou, and Wei-Zhong Lin (2015). “iMem-Seq: A Multi-label Learning Classifier for Predicting Membrane Proteins Types”. In.