CS482/682 Final Project Proposal

Protein-Protein Interaction Prediction

Jacky Chen/jchen291, Alexandra Szewc/aszewc1, Shreayan Chaudhary/schaud31, Nicholas Bowen/nbowen3

1 Problem Statement

Proteins are biological molecules that dictate almost all biological functions by interacting with other molecules or proteins. Because of this, there is widespread interest to determine which protein-protein interactions (PPI) may occur.[5] Since wet-lab experiments to determine PPIs are often costly and time consuming, computational methods are highly desirable. However, current computational approaches often rely on protein structure, which is not reliably available and does not apply to intrinsically unstructured proteins.[6] Here, we aim to develop a sequence bases PPI prediction approach.

2 Dataset

HuRi[3] is a dataset of $17,500$ protein sequences and their experimentally measured interactions with each other. Given two sequences, the database contains binary labels indicating whether the two proteins interact or not. Hence the database contains $17,500^2$ interactions with $51,813$ positive pairs.

3 Background Work

Previous work uses stacked autoencoders to study PPI prediction.[1] This research will serve as one of the benchmarks used to study protein-protein interaction prediction.

Other approaches use a combination of cross-attention and self-attention to do a mixed fusion strategy to study drug-target interaction. They combine different attention and pooling layers to do a graduate, intermediate fusion strategy and report improvements over a single late or early fusion approach.

Finally, Contrastive learning is a technique that involves the use of a self-supervised learning approach to learn useful features from data without labels, achieved by optimizing Contrastive Loss [4] that encourages the model to pull similar samples closer together in the feature space measured using a distance metric and push dissimilar samples farther apart.

4 Approach

Early Fusion with MLP This approach will involve concatenating the two protein sequences together and then passing the contacted sequence into a simple MLP as our baseline.

Early Fusion with Attention This approach will involve concatenating the two protein sequences together and then passing the contacted sequence into a model that uses a self-Attention mechanism before passing it into a fully connected layer.

Mixed Attention Mechanism Next, we will use cross/mixed attention mechanism similar to [2] in order to attempt to improve upon our original attention mechanism.

Mixed Attention with Contrastive Loss This approach will involve training a model using the Attention mechanism and contrastive loss. This loss function may fit well into our training scheme by leveraging the additional information carried in a large number of negative samples.

5 Approval

Verbal approval obtained from Professor Unberath during office hours on the condition that we establish baselines approaches, which are outlined in Section 4.

6 References

[1] Sun, T., Zhou, B., Lai, L. et al. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics 18, 277 (2017). https://doi.org/10.1186/s12859-017-1700-2

[2] Zhao Q, Zhao H, Zheng K, Wang J. HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics. 2022 Jan 12;38(3):655-662. doi: 10.1093/bioinformatics/btab715. PMID: 34664614.

[3] Luck, K., Kim, DK., Lambourne, L. et al. A reference map of the human binary protein interactome. Nature 580, 402–408 (2020). https://doi.org/10.1038/s41586-020-2188-x

[4] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... Krishnan, D. (2020). Supervised contrastive learning. Advances in neural information processing systems, 33, 18661-18673.

[5] Dunham B, Ganapathiraju MK. Benchmark Evaluation of Protein-Protein Interaction Prediction Algorithms. Molecules. 2021 Dec 22;27(1):41. doi: 10.3390/molecules27010041. PMID: 35011283; PMCID: PMC8746451.

[6]Tsuchiya, Y., Yamamori, Y. Tomii, K. Protein–protein interaction prediction methods: from docking-based to AI-based approaches. Biophys Rev 14, 1341–1348 (2022). https://doi.org/10.1007/s12551-022-01032-7