CSPC54_PROJECT

TABLE OF BASE PAPERS

- 106123094, 106123102

PROJECT TITLE: De-Novo Protein Function Prediction from Primary Sequences using a Bidirectional Recurrent Neural Network with Attention.

Paper	Problem Solved	Algorithm Used	Advantages	Drawbacks	Constraints
Liu, J. (2017)	De-novo protein function prediction for novel sequences where alignment-based methods fail.	Bidirectional Long Short-Term Memory (Bi-LSTM) Recurrent Neural Network.	- Does not require sequence alignment. - Effectively captures long-range dependencies.	- Acts as a "black box," making it difficult to interpret. - Less accurate than newer multi-modal methods.	- Performance depends on the quality/quantity of training data. - Relies exclusively on primary sequence.
You, R., et al. (2017)	Protein function prediction, framed as a translation problem from sequence to function.	RNN-based Neural Machine Translation (NMT) model with an encoder-decoder architecture.	- Novel conceptualization using powerful NLP techniques. Can capture hierarchical relationships between GO terms.	- Abstracting sequences into k-mers might lose information. -Performance is tied to the abstracted "language."	- Requires careful construction of "ProLan" and "GOLan" vocabularies. - Still a sequence-only approach.
Meng, S., & Wang, J. (2024)	Improving prediction accuracy by integrating multiple data modalities.	A hybrid framework (TAWFN) fusing a Graph Convolutional Network (GCN) and a Convolutional Neural Network (CNN) with attention.	- State-of-the-art performance by combining sequence and structure. - Uses attention to capture long-range dependencies.	- High model complexity. - Requires accurate 3D protein structure data, adding a preprocessing step.	- Dependent on the availability and accuracy of protein structures. - Computationally more intensive.