Predicting Protein Function with Recurrent Neural Networks in an Empirical Kernel Space

Author: Alexis Powell Email: poa@seas.upenn.edu

Advisor: Junhyong Kim Email: junhyong@sas.upenn.edu

Abstract

This thesis will investigate the use of segment-level alignment scores computed against a curated set of reference proteins to form an empirical kernel space for protein function classification. By aligning each protein subsequence to these reference templates, we derive landmark-based feature vectors that are then fed into a recurrent neural network (RNN), specifically a Long Short-Term Memory (LSTM) network. Building on techniques originally developed for protein fold recognition, this approach segments each amino acid sequence into fixed-length windows and aggregates the resulting alignment scores. Segmentation, as opposed to full-length alignments, can capture local motifs indicative of function, while also reducing length-related biases. We will validate the method using known functional labels, such as Gene Ontology (GO) terms, and evaluate performance across diverse species. This is to assess whether the resulting classifier can generalize beyond training-domain constraints. Ultimately, this work aims to extend the Protein Empirical Structure Space (PESS) framework into a tool for protein function prediction using machine learning approaches.

Introduction

Proteins, composed of amino acid sequences, are central to most biological processes, with each protein's function determined by its structure and interaction patterns. Accurately predicting protein function is a significant focus in computational biology, supporting insights into pathways, drug targets, and systems-level behavior. Traditional sequence-based methods or structure-based algorithms can struggle when data is sparse; at the same time, purely data-driven machine learning approaches often require extensive labeled examples that are unavailable for many protein families. To address these limitations, landmark-based methods, where unknown proteins are compared (threaded) against a set of reference templates, have emerged by encapsulating essential structure–function information into a smaller feature space. Building on this concept, the Protein Empirical Structure Space (PESS) framework was developed to tackle fold recognition [3]. PESS coordinates are created by threading each protein/domain against a fixed set of template structures, reducing dimensional complexity and improving generalization. Recent advances extend PESS by segmenting protein sequences into fixed-length overlapping windows, thereby capturing localized structural or functional motifs and mitigating biases associated with variable sequence lengths.

RNNs are a family of deep learning models designed for sequential inputs, making them well-suited for amino acid sequences. Their hidden state mechanism allows them to capture context from previous elements in the sequence, while LSTM units reduce issues, such as exploding or vanishing gradients, through a gating system. This capability to preserve and update long-range dependencies is relevant when analyzing protein subsequences for potential functional signals, which may span hundreds of residues. By translating aligned subsequences into compressed feature vectors and then feeding them into an RNN, it effectively transforms the structure–function mapping problem into a learned sequence-processing task. Moreover, feeding segment-level features into RNNs can uncover long-range dependencies relevant to protein function. By aligning each segment to reference templates, constructing PESS vectors, and integrating them via RNNs, this approach aims to identify functional classes (such as GO terms) with improved accuracy and cross-species robustness.

Motivation

Accurate protein function prediction is essential for understanding biological processes, identifying drug targets, and accelerating biomedical discovery. However, the high complexity and diversity of protein sequences often lead to sparse or noisy data, demanding robust computational methods. By using landmark-based feature representations with sequential neural networks, we can better capture sequence patterns that drive protein function.

Technical Scope and Depth

In this thesis, we will adapt the LSTM-based landmark approach used for fold classification and apply it to protein function prediction. First, protein sequences will be segmented into fixed-length overlapping windows (e.g., 100 amino acids, stride 10). Each segment will then be aligned (via a tool such as CNFalign_lite) against a curated set of reference proteins to generate the empirical kernel space (like the PESS). These alignment-score vectors form a landmark-based embedding that an LSTM network can process sequentially. The final hidden state of the LSTM will map to a fully connected layer for classification into functional categories, with training implemented in PyTorch. Model performance will be measured using standard multi-class classification metrics plus top-k accuracy if multiple function labels are relevant. We may also assess cross-species transfer by training on one dataset and testing on a non-overlapping set from different organisms.

Expected Outcomes and Analyses

This project aims to produce an LSTM-based classifier that utilizes segment-level alignment scores (i.e., empirical kernel coordinates) for accurate protein function prediction. Evaluation will include multi-class metrics such as accuracy, precision, recall, F1 scores, and confusion matrices. In scenarios with imbalanced functional labels, additional measures—like the Matthews Correlation Coefficient (MCC) or area under the ROC curve (AUC)—can offer more nuanced insights. Moreover, testing on cross-species datasets will help gauge generalizability beyond the training domain. By examining which segments and reference alignments influence classification decisions, we can assess whether the learned representations align with known biological motifs.

Schedule

- January 15–27, 2025: literature review on landmark-based methods for fold recognition and their applicability to protein function prediction, also read literature comparing GRUs and LSTMs for protein sequence tasks, identify suitable protein function datasets, confirm availability of function annotations, assess the current curated landmark set (CNFpred-based) for potential updates, including the possibility of integrating AlphaFold structures for broader functional coverage, explore alignment tools for segment-level scoring, draft + submit thesis proposal by Monday, January 27, 2025.

- January 28–February 14, 2025: refine goals based on feedback from the proposal, develop thesis outline, specifying methodology (segmentation, alignment, LSTM design), implement a minimal pipeline for segmenting protein sequences and generating empirical kernel coordinates on a small test set, investigate switching from CNFpred threading scores to ESM-based similarity measures (cosine distance in ESM2 embeddings) to reduce CPU overhead, submit thesis outline by Friday, February 14, 2025.

- February 15–28, 2025: finalize data acquisition, ensuring the protein sequences and reference templates are prepared, set up the development environment and alignment pipeline (CNFalign_lite), schedule and attend a Zoom meeting with Professor Kim and course coordinator(s) by Friday, February 28, 2025.

- March 1–7, 2025: implement the baseline LSTM model for function prediction using the segment-level feature vectors, train + evaluate initial runs on a subset of the dataset, document progress for the mid-course draft, submit the mid-course thesis draft by Friday, March 7, 2025.

- March 8–21, 2025: integrate feedback from the mid-course draft and advisor meeting, train the finalized LSTM model on the complete dataset, including cross-species subsets if applicable, perform hyperparameter tuning (number of LSTM units, learning rate, etc.), attend the second Zoom meeting with Professor Kim and course coordinator(s) by Friday, March 21, 2025.

- March 22–April 12, 2025: top-k, perform comprehensive performance evaluations and generate visualizations to assess interpretability, identify which segments or reference alignments drive classification, refine as needed.

- April 13–May 9, 2025: write the remaining sections of the thesis, incorporate final suggestions, polish figures and discussion of results, submit the final thesis by Friday, May 9, 2025.

References

[1] Chen, J., Z. Gu, L. Lai, and J. Pei. In silico protein function prediction: the rise of machine learning-based approaches. *Medical Reviews*, vol. 3, no. 6, 2023, pp. 487–510, doi:10.1515/mr-2023-0038, PMID: 38282798, PMCID: PMC10808870.

[2] Jang, Y.J., Qin, QQ., Huang, SY. et al. Accurate prediction of protein function using statistics-informed graph networks. *Nat Commun 15*, 6601 (2024), https://doi.org/10.1038/s41467-024-50955-0.

[3] Kuang, Da, Dina Issakova, and Junhyong Kim. Learning Proteome Domain Folding Using LSTMs in an Empirical Kernel Space. *Journal of Molecular Biology*, vol. 434, no. 15, 2022, p. 167686, ISSN 0022-2836, doi:10.1016/j.jmb.2022.167686.

[4] Kulmanov, M., Guzmán-Vega, F.J., Duek Roggli, P. et al. Protein function prediction as approximate semantic entailment. *Nat Mach Intell 6*, 220–228 (2024), https://doi.org/10.1038/s42256-024-00795-w.

[5] Liu, Xueliang. Deep Recurrent Neural Network for Protein Function Prediction from Sequence. *ArXiv*, 2017, eprint:1701.08318, https://arxiv.org/abs/1701.08318.

[6] Soleymani, F., E. Paquet, H. Viktor, W. Michalowski, and D. Spinello. Protein-protein interaction prediction with deep learning: A comprehensive review. *Computational and Structural Biotechnology Journal*, vol. 20, 2022, pp. 5316–5341, doi:10.1016/j.csbj.2022.08.070, PMID: 36212542, PMCID: PMC9520216.

[7] Yu, Guo-Xian, Huzefa Rangwala, Carlotta Domeniconi, Guoji Zhang, and Zili Zhang. Predicting Protein Function Using Multiple Kernels. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 5, 2015, pp. 219–233, doi:10.1109/TCBB.2014.2351821.

Undergraduate Courses

Fall 2021: CHEM 053, CHEM 101, CIS 110, MATH 104, WRIT 089
Spring 2022: CBE 160, CHEM 054, CHEM 102, MATH 114, MUSC 005, PHYS 150
Fall 2022: CBE 2300, CHEM 2410, FREN 0100, MATH 2400, PHYS 0141
Spring 2023: CBE 2310, CBE 3500, CHEM 2510 (Biological Chemistry), EAS 2030, MATH 2410
Fall 2023: CBE 3600, CBE 4790 (Biotechnology & Biochemical Engineering), CIS 1200, JPAN 0103
Spring 2024: CIS 1600, CIS 2400, HIST 1169, NETS 2120, PHYS 0051
Summer 2024: CIS 1210 Transfer Credit (XCAT)
Fall 2024: CIMS 1800, CIS 2620, CIS 4190 (Applied Machine Learning), CIS 4360 (Computational Biology & Biological Modeling), SPAN 0100
Spring 2025: CIS 3200, CIS 3500 (Software Design/Engineering), CIS 4980, JPAN 0105