Exploring Hybrid Models for Protein Secondary Structure Prediction

Joel Markapudi Tanay Grover Jennisha Martin   
[markapudi.j@northeastern.edu](mailto:markapudi.j@northeastern.edu) [grover.t@northeastern.edu](mailto:grover.t@northeastern.edu) [martin.jennis@northeastern.edu](mailto:martin.jennis@northeastern.edu)

**Project Introduction**

Our project, ‘Protein Secondary Structure Prediction’ focuses on achieving structure prediction by implementing and comparing hybrid models like Hidden Markov Model (HMM)-based and SVM algorithms, using the CB513 dataset. Protein secondary structure prediction is essential for understanding protein function. By exploring multiple algorithms, we aim to identify which approaches effectively capture the complex relationships between amino acid sequences and their corresponding secondary structures. By leveraging the sequence modeling and feature learning capabilities of different models, we aim to see how well each approach captures local and global dependencies in protein sequences.

We are particularly curious about the potential of hybrid approaches. This project not only advances bioinformatics prediction methods but also has broader implications for drug effectiveness, discovery, and disease mechanisms.

**Data**

We will utilize the CB513 dataset, a well-established benchmark in protein secondary structure prediction research. The dataset includes 513 non-homologous protein sequences. Each sequence is represented using standard single-letter amino acid codes. It also has Secondary Structure Labels. They are annotations for each amino acid, indicating helix (H), sheet (E), or coil (C). The CB513 dataset is ideal for our project due to its balance and manageability. Other powerful, versatile datasets like Protein Databank (PBD) exist but are complex to parse and have too much data. We plan to build thorough implementations with a ‘prototype dataset’, and make progress to use the full CB513 dataset, and potentially a larger dataset later.

Data Source: <https://www.kaggle.com/datasets/moklesur/cb513-dataset-for-protein-structure-prediction>, <https://huggingface.co/datasets/proteinea/secondary_structure_prediction/tree/main>.

**Algorithms and Project Plan**

**Algorithms:** HMM-CRF with Neural Nets, HMM-BiLSTM Model, SVM with Position-Specific Scoring Matrix (PSSM).

1. **Overview for Algorithm 1:** We find CRFs fascinating because they excel at modeling the conditional probability of a sequence of labels given an input sequence. In protein secondary structure prediction, CRFs can capture the dependencies between adjacent secondary structure elements—something traditional neural networks might overlook. This ability to model label dependencies addresses a critical gap in predicted structures’ accuracy. Also, HMMs are valuable for their strength in modeling sequential data through probabilistic state transitions. **Motivation:** To explore hybrid models and learn of the synergy between probabilistic models and deep learning.
2. **Overview for Algorithm 2:** The BiLSTM (Bidirectional Long Short-Term Memory) model offers a powerful approach for capturing both forward and backward dependencies in sequential data. By integrating BiLSTM with HMMs, we leverage both probabilistic state transitions and deep learning’s capacity for understanding complex patterns. BiLSTM’s ability to account for past and future contexts allows it to handle long-term dependencies, which is vital for capturing the intricate relationships we are looking to unpack. **Motivation:** The motivation is based on the need to capture both local and distant dependencies in protein sequences. BiLSTM layers ensure that context from all directions in the sequence is considered. We aim to understand how this synergy between sequence modeling and bidirectional information flow enhances secondary structure prediction accuracy.
3. **Overview for Algorithm 3:** We are interested in using Support Vector Machines (SVM) with Position-Specific Scoring Matrix (PSSM) for this project because it is a straightforward yet powerful method for predicting protein secondary structures. SVM works especially well with PSSM profiles, which are rich in evolutionary information. This setup lets SVM detect patterns to predict whether each part of a protein will form a helix, strand, or coil. Compared to deep learning models, SVM with PSSM is faster and resource-efficient, which is ideal for this dataset. **Motivation:** To explore SVM’s stability, avoiding the high computational demands. The SVM “kernel trick” also allows it to adapt to complex, high-dimensional data—perfect for handling the detailed information in PSSM profiles. Keeping the model simple, focused on essential structural patterns.
4. **Sequence of methods:** The main progression of steps we wish to follow here would be Sequence Encoding, Label Preparation, Feature Analysis and Augmentations, Architecture Selection (Neural Nets, BiLSTM, etc.), Integration of HMM and hybrid components, Training, Validation.
5. **Libraries:** torchcrf, numpy, PyTorch, Matplotlib or Seaborn, and hmmlearn, scikit-learn for SVM, PSI-BLAST (or HH-suite). Biopython for sequence handling. For HMM, we will attempt manual code given the possibility.
6. **Specific Components:** The losses -CrossEntropyLoss(), nll\_loss(), hmmlearn.hmm.GaussianHMM, MultinomialHMM from the **hmmlearn** library, Stacking torch.nn.LSTM for BiLSTM, crf\_layer, and CRF() from torchcrf library, and pystruct library.
7. **Loss Function:** The training objective will be to minimize the negative log-likelihood of the correct label sequences or using cross-entropy loss minimization.
8. **Visualizations and Expected Results**
   1. Prediction Outputs: Secondary structure predictions for test sequences in the dataset.
   2. Evaluation Metrics: Accuracy, precision, recall, and F1-score for each class.
   3. Comparative Analysis: Tables and charts comparing the performance.
   4. Visualizations: Confusion matrices, Training, and validation loss curve, etc.

**Task Allocation**

|  |  |  |
| --- | --- | --- |
| **Project Member** | **Assigned Task** | **Algorithm Focus** |
| Joel Markapudi | HMM-CRF Hybrid Model. Full Pipeline. | HMM-CRF with Neural Nets |
| Tanay Grover | Sequence Modeling with BiLSTM. Full Pipeline. | HMM-BiLSTM Networks |
| Jennisha Martin | Pattern Detection with SVMs. Full Pipeline. | SVM with Position-Specific Scoring Matrix (PSSM). |

**Timeline**

Given our available timeline (Nov-24), our working plan is as follows. In the first few days, we will complete domain understanding, dataset understanding, pre-processing, feature preparation, and base architectures (12-15 days). Midway, we will integrate the necessary hybrid layers. Followed by training the model, conducting validation, and refining based on performance metrics.

We plan to use an agile methodology, with each team member actively involved in all stages of their algorithm pipeline. Team members will collaborate, and share updates and insights frequently, enabling quick iterations, and adjustments.

**References (Academic Papers)**

Li, Z., & Yu, Y. (2016). Protein secondary structure prediction using deep convolutional neural fields. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(5), 1169-1176. doi: 10.1109/TCBB.2016.2573281

Zhou, P., Yang, L., Wang, Y., & Zeng, J. (2018). An improved model for protein secondary structure prediction based on an explicit integration of LSTM and CRF. BMC Bioinformatics, 19(1), 135. doi: 10.1186/s12859-018-2128-1

Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. Foundations and Trends® in Machine Learning, 4(4), 267-373. doi: 10.1561/2200000013

Wang, Z., Yan, W., & Oates, T. (2018). Time series classification from scratch with deep neural networks: A strong baseline. arXiv preprint arXiv:1806.06394. Retrieved from <https://arxiv.org/abs/1806.06394>

Guo, Y., Wang, B., Li, W., & Yang, B. (2018). Protein secondary structure prediction is improved by recurrent neural networks integrated with two-dimensional convolutional neural networks. Journal of Bioinformatics and Computational Biology, 16(5), 1850021. doi: 10.1142/S021972001850021X

Ismi, D. P., Pulungan, R., & Afiahayati. (2022). Deep learning for protein secondary structure prediction: pre-and post-AlphaFold. Computational and Structural Biotechnology Journal, 20, 6271-6286. doi: 10.1016/j.csbj.2022.11.012