Project 2
pLM-Enhanced CATH Classification

Background

Classifying protein domains into structural and evolutiotnary groups is fundamental for understanding protein function and evolution. While Pfam focuses on sequence-based family classification, CATH (Class, Architecture, Topology, Homology) provides a hierarchical classification based on both sequence and structural properties [1]. CATH organizes protein domains into a four-level hierarchy: Class (secondary structure composition), Architecture (arrangement of secondary structures), Topology (connectivity of secondary structures), and Homology (evolutionary relationships). Traditional approaches for CATH classification rely on structural alignment and sequence profiles, but deep learning methods like those developed for Pfam [2,3] could potentially improve coverage and accuracy. Recent advances in protein language models (PLMs) like ProtT5 [4] have produced powerful pre-trained embeddings that capture protein sequence semantics and evolutionary information. Meanwhile, new domain segmentation tools like Chainsaw [5] have significantly improved the accuracy of domain boundary prediction using convolutional neural networks. This project applies the ProtENN2-inspired approach to CATH classification, replacing one-hot encoding with ProtT5 embeddings as input to CNN layers, aiming to predict CATH domains from sequence alone with improved performance compared to existing methods.

Hypothesis/Objective

Objective: Develop a protein domain classification model using ProtT5 per-residue embeddings as input to a ProtENN2-style architecture, exploring whether pre-trained embeddings can effectively predict CATH domain classifications at various hierarchical levels, and compare performance against state-of-the-art domain prediction methods like Chainsaw.

Hypothesis: Pre-trained protein language model embeddings provide a rich representation that can capture both sequence and implicit structural information needed for CATH classification. ProtT5 embeddings should allow CNN layers to more effectively discriminate between CATH domains compared to one-hot encoding, particularly at the Topology and Homology levels where both sequence and structural information are important. We hypothesize that certain CATH domains that are challenging to classify from sequence alone might be better resolved using embeddings that encode evolutionary context and long-range dependencies, potentially approaching the performance of structure-based methods like Chainsaw.

Datasets

1. **CATH subset:** Select a manageable subset of the CATH database [1] for the classification task, focusing on 50-100 homology groups (H-level) with sufficient sequences per group. Consider separate classification tasks for different CATH levels (C, A, T, or H).

2. **Training/validation/test split:** Create a 70/10/20 split, ensuring no significant sequence identity overlap between the sets to test generalization.

3. **ProtT5 embeddings:** Use the ProtT5-XL-UniRef50 model [4] to generate per-residue embeddings (1024-dimensional vectors) for all sequences.

4. **Baseline dataset:** For comparison, prepare one-hot encoded sequences (21-channel vectors per position) of the same proteins for a comparative baseline model following the original ProtENN2 architecture.

Method/Task

1. **Data preparation:**

    ○ Extract domain sequences from the selected CATH homology groups
    ○ Precompute and save ProtT5 per-residue embeddings for efficient reuse
    ○ Prepare residue-level labels for each sequence (CATH classification at different levels)
    ○ Create mapping between sequence positions and domain boundaries

2. **Per-residue classification with ProtT5 embeddings:**

    ○ Following ProtENN2's approach but replacing the one-hot encoding input with ProtT5 embeddings
    ○ For each protein sequence, obtain the per-residue embeddings from ProtT5 (L × 1024 tensor, where L is sequence length)
    ○ Pass these embeddings through residual CNN layers similar to ProtENN2
    ○ Apply a fully connected classification layer to predict CATH class probabilities for each residue
    ○ Implement thresholding and post-processing to identify continuous regions of domain predictions

3. **Baseline ProtENN2-style model:**

    ○ Implement a simplified version of the original ProtENN2 approach with one-hot encoded inputs
    ○ Use the same convolutional architecture and residual layers as the ProtT5 version
    ○ Apply identical classification layer and post-processing
    ○ Compare performance and training efficiency with the embedding-based model

4. **Comparison against Chainsaw and other domain predictors:**

    - Benchmark against Chainsaw [5], which represents the state-of-the-art in domain boundary prediction using fully convolutional neural networks
    - Include other domain prediction methods such as UniDoc, PUU, and SWORD for comprehensive comparison
    - Use the same evaluation metrics as in the Chainsaw study: intersection-over-union (IoU), proportion of correctly parsed domains, and domain boundary distance

5. **Hierarchical classification exploration:**

    - Develop models for different CATH hierarchical levels (C, A, T, H)
    - Explore multi-task learning approaches that predict multiple levels simultaneously
    - Analyze which CATH levels benefit most from the ProtT5 embeddings versus one-hot encoding

6. **Advanced experimentation (if time permits):**

    - Ensemble learning: Create an ensemble of models trained on different subsets of the data
    - Domain embedding analysis: Analyze how embeddings cluster by CATH classification
    - Cross-database prediction: Train on CATH and test on Pfam (or vice versa) to explore relationship between sequence and structure classification

Expected Outcomes

1. **Effective CATH classification:** The ProtT5-embedding-based CNN model should achieve strong performance on predicting CATH domains from sequence alone, particularly at the Homology level which has both sequence and structural components.

2. **Comparative performance against Chainsaw:** While Chainsaw uses structural information directly for domain boundary prediction (achieving 78% accuracy on CATH domains), our sequence-only approach with ProtT5 embeddings should approach this performance level for domains with strong sequence signals. Understanding the performance gap between sequence-only and structure-based approaches will provide insights into when structural information is essential.

3. **Comparative performance across CATH levels:** Analysis of how performance varies across the CATH hierarchy, with potentially better results for Class and Homology levels than for Architecture and Topology levels which are more purely structural.

4. **Input representation insights:** Quantification of how much ProtT5 embeddings improve performance over one-hot encoding for structural classification tasks, providing insights

into how much implicit structural information is captured in sequence embeddings.

5. **Domain boundary detection:** Analysis of how well the residue-level predictions aggregate into accurate domain predictions, including boundary detection quality compared to existing methods such as Chainsaw, UniDoc, and SWORD.

6. **Structure-sequence relationship insights:** If cross-database prediction is implemented, insights into the relationship between sequence-based (Pfam) and structure-based (CATH) classification systems, potentially identifying where these systems align and diverge.

References

[1] Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., ... & Orengo, C. A. (2021). CATH: increased structural coverage of functional space. Nucleic Acids Research, 49(D1), D266-D273.

[2] Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., ... & Colwell, L. J. (2022). Using deep learning to annotate the protein universe. Nature Biotechnology, 40(6), 932-937.

[3] Ponamareva, I., Andreeva, A., Bileschi, M. L., Colwell, L., & Bateman, A. (2024). Investigation of protein family relationships with deep learning. Bioinformatics Advances, 4(1), vbae132.

[4] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., ... & Rost, B. (2021). ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 7112-7127.

[5] Wells, J., Hawkins-Hooker, A., Bordin, N., Sillitoe, I., Paige, B., & Orengo, C. (2024). Chainsaw: protein domain segmentation with fully convolutional neural networks. Bioinformatics, 40(5), btae296.

[6] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. Structure, 5(8), 1093-1109.