Báo cáo Dự án Cuối kỳ: Dự đoán Chức năng Protein – CAFA 6 Challenge

Nguyễn Minh Đức
Mã số sinh viên: 23020049
Lớp: INT3405_7 - Học máy
Đại học Công nghệ, ĐHQGHN
Tên nhóm Kaggle: ABC_INT34057
Giảng viên hướng dẫn: ThS. Tạ Việt Cường

***Tóm tắt nội dung*—Dự án này tập trung giải quyết bài toán dự đoán chức năng protein trong cuộc thi CAFA 6 (Critical Assessment of Functional Annotation) trên nền tảng Kaggle. Mục tiêu là xây dựng hệ thống gán các thuật ngữ Gene Ontology (GO terms) cho các trình tự protein chưa được chú thích. Kết quả thực nghiệm cho thấy phiên bản tốt nhất (v2) đạt vị trí xếp hạng 326/1073 trên Public Leaderboard với điểm số F-max là 0.312. Phương pháp tiếp cận hiệu quả nhất là sự kết hợp (ensemble) giữa mô hình Deep Learning sử dụng ESM-2 embeddings và các nguồn dữ liệu tương đồng (Homology).**

***Từ khóa*—CAFA 6, Protein Function Prediction, Gene Ontology, ESM-2, Ensemble Learning.**

I. Giới thiệu

Việc xác định chức năng của protein là một bài toán cốt lõi trong tin sinh học. Trong khuôn khổ môn học INT3405 – Học máy, dự án này tham gia cuộc thi CAFA 6 nhằm phát triển các mô hình tính toán có khả năng dự đoán chính xác các GO terms.

Độ chính xác của mô hình được đánh giá dựa trên chỉ số F-max (Weighted F1-score). Thách thức lớn nhất nằm ở tính chất phân cấp của Gene Ontology và sự mất cân bằng dữ liệu lớn.

Kết quả đạt được của dự án: (12/12/2025)

- **Vị trí xếp hạng:** 326/1073.
- **Public Score cao nhất:** 0.312.
- **Phiên bản tốt nhất:** Version 2 (v2).

Mã nguồn dự án được lưu trữ tại: GitHub Repository.

II. Phương pháp đề xuất

Mô hình đạt hiệu quả cao nhất (0.312) được xây dựng dựa trên một pipeline kết hợp Deep Learning và Homology-based inference.

A. Tổng quan luồng thực thi

Quy trình xử lý bao gồm 4 bước chính:

1) Trích xuất Đặc trưng (Feature Extraction): Sử dụng mô hình ngôn ngữ protein **ESM-2** (Evolutionary Scale Modeling) để chuyển đổi trình tự protein thành các vector embedding. ESM-2 nắm bắt tốt các đặc trưng cấu trúc và tiến hóa của protein mà không cần so gióng trình tự (alignment).

2) Huấn luyện (Training): Một mạng nơ-ron (Neural Network) đơn giản được huấn luyện để ánh xạ từ không gian embedding ESM-2 sang xác suất của các GO Terms. Việc giữ mô hình đơn giản giúp tránh overfitting trên tập dữ liệu hạn chế.

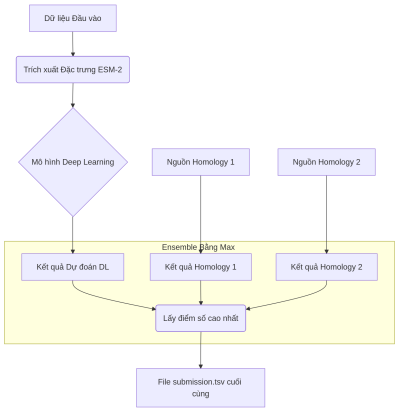
3) Ensemble - Kết hợp dự đoán: Đây là bước then chốt mang lại sự đột phá về điểm số. Chiến lược sử dụng là **Max Pooling Ensemble**:

$$P_{final} = \max(P_{DL}, P_{Homology\_1}, P_{Homology\_2}) \quad (1)$$

Trong đó kết quả từ mô hình Deep Learning (P_{DL}) được kết hợp với hai nguồn dự đoán dựa trên độ tương đồng (Homology) mạnh mẽ khác. Phương pháp này tận dụng được khả năng tổng quát hóa của Deep Learning và độ chính xác cao của Homology đối với các trình tự đã biết.

4) Tạo Submission: Dữ liệu được định dạng lại theo yêu cầu của cuộc thi để tạo file `submission.tsv`.

B. Sơ đồ hệ thống



Hình 1. Sơ đồ luồng xử lý của pipeline đạt điểm số 0.312.

Hình 1 minh họa luồng dữ liệu, trong đó việc sử dụng embedding tính toán sẵn giúp tiết kiệm tài nguyên tính toán đáng kể.

III. Thực nghiệm và Thảo luận

Quá trình phát triển trải qua nhiều vòng lặp (Rounds) để tìm ra giải pháp tối ưu.

A. Round 1: Xây dựng Baseline

Bắt đầu với việc tái lập các notebook public (EDA + Model cơ bản). Thử nghiệm ban đầu với T5 embeddings cho thấy hiệu quả kém hơn so với ESM-2.

- *Bài học:* Lựa chọn embedding là yếu tố tiên quyết.

B. Round 2: Đột phá với Ensemble (Best Score)

Tại vòng này, notebook `01_Baseline_and_Ensemble_Best_Score` được phát triển.

- *Phương pháp:* Kết hợp ESM-2 Model + 2 nguồn Homology.
- *Kết quả:* Public Score đạt **0.312**.
- *Đánh giá:* Chiến lược ensemble đơn giản nhưng hiệu quả vượt trội so với mô hình đơn lẻ.

C. Round 3: Xử lý hậu kỳ nâng cao

Tập trung vào việc tích hợp kiến thức sinh học (Gene Ontology Structure) vào mô hình.

- *Kỹ thuật:* Cài đặt module **Hierarchy Propagation** (lan truyền phân cấp) và **Negative Propagation**.
- *Kết quả:* Điểm số đạt khoảng 0.257 thấp hơn Round 2
- *Nguyên nhân:* Do chủ động loại bỏ 1 nguồn Homology (có score 0.3) để kiểm tra hiệu quả thực sự của thuật toán lan truyền mà không bị nhiễu bởi nguồn dự đoán mạnh bên ngoài.
- *Ý nghĩa:* Giảm sự phụ thuộc vào các mô hình có sẵn, tăng tính làm chủ thuật toán.

D. Round 4: Các hướng tiếp cận thay thế

Thực hiện các thử nghiệm bổ sung:

1) Multi-Ontology Architecture: Thử nghiệm huấn luyện 3 mô hình riêng biệt cho 3 nhánh Ontology (BPO, CCO, MFO).

IV. Kết luận

A. Tổng kết

Dự án đã hoàn thành mục tiêu xây dựng hệ thống dự đoán chức năng protein với thứ hạng khả quan (Top 30%). Bài học lớn nhất rút ra là sức mạnh của phương pháp Ensemble Learning khi kết hợp các cách tiếp cận khác nhau (Deep Learning và Homology).

B. Hướng phát triển

Hướng đi tiềm năng nhất để cải thiện kết quả là kết hợp ưu điểm của Round 2 và Round 3:

1) Sử dụng nền tảng là pipeline Ensemble (DL + Homology) của Round 2 để đảm bảo điểm số cơ sở cao.
2) Áp dụng các thuật toán xử lý hậu kỳ (Hierarchy/Negative Propagation) của Round 3 lên kết quả cuối cùng để tinh chỉnh độ chính xác và đảm bảo tính nhất quán sinh học.

Tài liệu

[1] CAFA 6 Kaggle Competition, "Critical Assessment of Functional Annotation,"https://www.kaggle.com/competitions/cafa-6.

[2] GitHub Repository: CAFA-6-TEAM-ABC_INT34057, https://github.com/nmd29io/CAFA-6-TEAM-ABC_INT34057.