Journal Club (2023 年 5 月 15 日) まとめ

担当: 清水 秀幸

発表論文:

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, Huimin Zhao
Enzyme function prediction using contrastive learning
Science. 2023 Mar 31;379(6639):1358-1363
doi: 10.1126/science.adf2465

研究目的および概要:

UniProt に登録されているタンパクのアノテーションがついているものはごく一部にとどまり、他の多くのものは機能が不明である。アミノ酸配列から機能を正確に予測するため、酵素を題材に深層学習モデルを構築したい。

UniProt に登録されている酵素について、タンパク質言語モデルの 1 つ ESM-1b を使いアミノ酸から数値ベクトルに変換した後に EC 番号を教師とした supervised constrastive learning を行い、同じ EC 番号のアミノ酸配列が潜在空間において近接するように訓練を行った (図 1)。

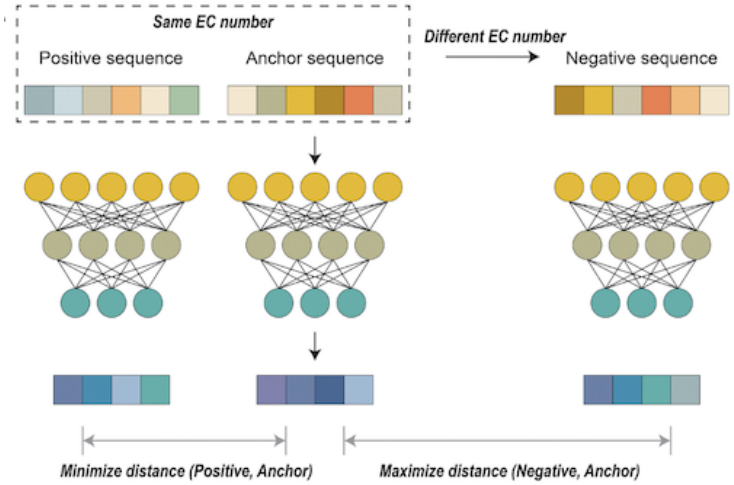

図 1

既存の手法である BLASTp, DeepEC, ProteInfer 等より未知のデータに対する F1 等の指標が優れることを示した後、annotation がまだ十分にはついていない halogenase についてアミノ酸配列から EC 番号を予測し、その活性があるのかを HPLC や mass spectrometry を使って実証した（図 2）

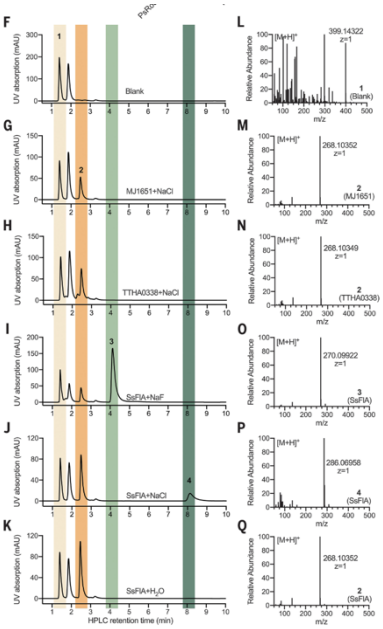

図 2

先行研究と比べて何がすごい？ 技術やアプローチのキモはどこ？:

- アミノ酸配列を数値に埋め込む際に、タンパク質言語モデルだけでなく EC 番号そのものを教師データとした supervised な対照学習を行っているところ。これにより機能を重視した埋め込み表現が得られ、先行研究と比べて SOTA を達成

どうやってこの手法/仮説の有効性を検証したのか:

- 本手法（CLEAN）を提案後に UniProt に加えられた新しいタンパクを使ったデータで、既存手法らとともに EC 番号を予測し、precision, reall, F1 などの指標で性能比較
- Halogenase を題材に EC 番号を予測しその活性があるのかを HPLC と massspectrometry などで実証。

その他、議論した内容（ネガティブコメントや limitation も):

- アノテーションがついているタンパクは UniProt のごく一部だけであり、機能未知タンパクの機能を推定できるポテンシャルが評価されて Science に掲載された
- 一番大事なのは最初のモデルを作るところで、その作り方はタンパク質言語モデルと supervised-contrastive learning だが、それはともに既存の報告になる。知識をしっかりと仕入れているのは大事なこと
- タンパク質は構造予測から機能予測が重要なトピックスになってきているが、予測で終わると Science には通らない。最後の Figure で実験で示しているのはとても impressive。

この研究をさらに発展させるとしたら:

- EC 番号は階層性になっているため、単なる EC 番号 vs その他という構図ではなく階層性を意識した埋め込み学習ができるとよりよいものができそうだ。