6ZG-02

タンパク質のリガンド結合部位の特徴量化手法の改善に関する研究

塩澤　理紗 †　　安尾　信明 ‡
東京工業大学　情報理工学院 †　　東京工業大学　物質・情報卓越教育院 ‡
関嶋　政和 §
東京工業大学　情報理工学院 §

1　導入

1つの薬を開発するのに約10から15年の時間、約26億ドルもの資金がかかるとされており[1]、これらのコストを削減するために計算機科学を応用する流れが活発化している。特に化合物探索の段階で, 近年タンパク質とリガンドとの結合部位, いわゆる「ポケット」が重要視されつつある。

DeeplyTough[2] はポケット構造同士の関係性を考慮しながら、ポケット部位の立体情報をCNN で畳み込むことによってn次元記述子として特徴量化できるモデルである。本研究ではこの DeeplyTough モデルをベースに構造の柔軟性や結合部位とリガンドとの関係性を考慮した特徴量を抽出できるようにすることを目標としている.

2　手法

本研究では DeeplyTough モデルの主に損失関数とネットワークへの入力情報を変更することによって精度の向上を図った. 特に,TOUGH-M1 データセット [3],Vertex データセット [4] に対する精度を維持しつつ,ProSPECCTS[5] における P2-P6.2 の 7 つの指標を上げることを目指す.

表1　ProSPECCTS データセットの構成概要

番号	評価内容
P2	NMR 構造, 結合部位に対する感度および柔軟性の評価
P3,P4	decoy 構造のデータセット, 物理化学的・形状的特性の異なる結合部位の差別化
P5,P5.2	Kahraman データセット, 同一リガンドと補酵素に結合する結合部位の分類
P6,P6.2	Barelier データセット, 同一もしくは類似のリガンドを持つが無関係の結合部位ペアの分類

本研究で注目した ProSPECCTS[5] データセットの評価内容は表1のようになる.

2.1　損失関数の変更

DeeplyTough モデル [2] ではペアワイズネットワークを用いることで学習している. 学習には TOUGH-M1 データセット [3] を用いており, それにおける正のペアと負のペアに対してそれぞれ別の損失関数を適用していた. しかし, 既存の損失関数では負のペアにあまり損失値が付加されず, 正のペアについて過学習されてしまうような設計がなされているのではないかと考えた. そこで負のペア側の損失関数を式 (1) にすることで正のペア側の過学習が抑えられないか

Study on Improvement of Feature Extraction Method for Protein-Ligand Binding Sites
† Lisa Shiozawa, Tokyo Institute of Technology
‡ Nobuaki Yasuo, Tokyo Institute of Technology
§ Masakazu Sekijima, Tokyo Institute of Technology

を検証した.

$$f(x) = \max(-\ln(\frac{x}{margin}), 0) \qquad (1)$$

2.2 MD を用いた Data Augmentation

DeeplyTough モデル [2] では同一アミノ酸配列のタンパク質につき 1 つの立体構造のみを学習に用いていたために, 本来生体内で揺らいでいるタンパク質の構造の柔軟性をあまり考慮できないモデルになっていたと考えた. そこで、Amber[6] を用いた Data Augmentation を行うことで構造の柔軟性を考慮できるようになり ProSPECCTS[5] の P2 の指標を上げられるのではないかと考えた.

2.3 表面情報の付加（チャンネル構成の変更）

ポケット部位の幾何学性を捉えることができればより精度が向上するのではないかと考えた. チャンネル構成に 1 チャンネル追加し, タンパク質表面点座標に関する情報を付加した. チャンネル値は該当座標から半径 5×10^{-5}Å 以内に表面点が存在すれば 1, そうでなければ 0 とした.

3 結果と考察

負のペア側の損失関数を式 (1) にすることで正のペア側の過学習が抑えられ各データセットの検証において全体的に精度が向上した. MD を用いた Data Augmentation を行い複数構造を学習に用いることで ProSPECCTS[5] の P2 の指標が大きく向上したため構造の柔軟性を考慮できるようなモデルになった. 表面情報を付加することで, 損失関数と Data Augmentation 方針を変更した上で P3-P5.2 の精度を維持もしくは向上させることができたため, ポケット部位の形状をより捉えられるようになり物理化学的および幾何学的類似性ややポケット部位とリガンドとの関係性が認識できるようになった.

4 結論

損失関数,Data Augmentation, チャンネル構成を変えることでオリジナルの DeeplyTough モデルよりも包括的に高精度なポケット部位の比較解析が行えるようなモデルを学習することができた.

参考文献

[1] Paul, et al. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, Vol. 9, No. 3, pp. 203–214, Mar 2010.

[2] Martin Simonovsky and Joshua Meyers. Deeplytough: Learning structural comparison of protein binding sites. *Journal of Chemical Information and Modeling*, Vol. 60, No. 4, pp. 2356–2366, 2020. PMID: 32023053.

[3] Rajiv Gandhi Govindaraj and Michal Brylinski. Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. *BMC Bioinformatics*, Vol. 19, No. 1, p. 91, March 2018.

[4] Chen, et al. Prediction of protein pairs sharing common active ligands using protein sequence, structure, and ligand similarity. *Journal of Chemical Information and Modeling*, Vol. 56, No. 9, pp. 1734–1745, 2016. PMID: 27559831.

[5] Christiane Ehrt, Tobias Brinkjost, and Oliver Koch. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (prospecccts). *PLOS Computational Biology*, Vol. 14, No. 11, pp. 1–50, 11 2018.

[6] D.A. Case and others (2020). Amber 2020, University of California, San Francisco.