4ZK-01

Graph Convolutional Network を用いた蛋白質表面データに基づく蛋白質・リガンド結合予測

吉田　武司†
神戸大学†
システム情報学研究科

大川　剛直‡
神戸大学‡
システム情報学研究科

1. はじめに

蛋白質はリガンドと呼ばれる低分子化合物と相互作用し、ポケット領域と呼ばれる窪んだ構造でリガンドと結合することで機能を発揮する[1,2]。ポケット構造が類似する蛋白質は通常、似た機能やリガンドに結合することが知られている。この結合の分析は難病の治療薬や新しい治療法の提案に繋がる可能性がある。

このような結合分析の手法として Convolutional neural network (CNN) を用いた手法[3]が挙げられる。しかし、CNN を用いた手法では三次元データを扱う際の回転への対応といった課題がある。また、アミノ酸配列など一次元データを扱う 1DCNN[9]も存在するが、三次元情報を補完できないといった課題がある。

これらを根本から解決するモデルとして、Graph Convolutional Network (GCN) が挙げられる。GCN は三次元のデータに対して、回転の影響を受けずにその特徴学習が可能である。

また、異なる分子構造を持つ蛋白質が似たポケット構造を持ち、同じリガンドと結合するような事例も確認されている[4]。しかし既存の蛋白質リガンド結合予測手法では、分子骨格を扱ったものが多く、蛋白質表面データを扱った事例は少ない。。

そこで、本研究では、蛋白質表面データを基に GCN を用いた蛋白質リガンド結合予測手法 PROLIPS(PROtein LIgand binding Predictor by Surface)を提案する。PROLIPS はデータの回転に関わらず、より効果的に蛋白質の表面データからポケット構造に関連する特徴を抽出できる。

2. 提案手法

PROLIPS の概要を Fig.1 に示す。

2.1 データセット

本研究で使用するデータは、蛋白質の表面形状と特性を提供するオープンデータベース eF-site[4]から取得する。

A method of predicting protein-ligand binding using Graph Convolutional Network for protein surface data
†Takeshi Yoshida, Graduate School of System Informatics, Kobe University
‡Takenao Ohkawa, Graduate School of System Informatics, Kobe University

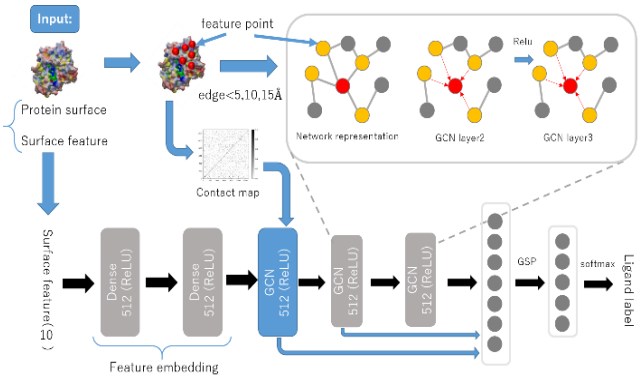

Fig. 1. The overview of the PROLIPS.

eF-site では分子表面の化学的特性(静電性、疎水性、温度因子)と幾何学的情報(点座標、法線ベクトル)が三次元ポリゴンデータの形で格納されている。

また、本研究では、PDBbind version 2017[6]を蛋白質リガンド結合ペアのベンチマークとして使用する。このベンチマークに登録された蛋白質とリガンドのセットを参照し、eF-site から蛋白質表面点データの抽出を行った。その後、データ拡張等を行い、166 種類の結合対象リガンドと 8352 の蛋白質表面点データが得られた。

2.2 蛋白質表面データの特徴点

eF-site で得た蛋白質表面点データには結合に関与していない部分も過剰に含まれており、これらを用いることは、機械学習の観点から学習のノイズに、そして計算効率の観点から非効率である。

そこで、本研究では、蛋白質表面点データからポケット領域を形成する特徴点と呼ばれる点を抽出する。湾曲部分を特徴とするポケット構造を主に抽出することを目指して、本研究では、分子表面点群の特徴データ「曲率」を利用して、湾曲が顕著なポイントを特徴点として抽出する。具体的な特徴点の抽出法は以下のとおりである。

1. eF-site の表面点データの特徴データから、最大曲率 A と最小曲率 B を抽出し、各点の曲がり具合の大きさ C を数値化する。
$$C = \sqrt{A^2 + B^2}$$
2. 曲がり具合 C の大きさが上位の点の何点か(後述)を特徴点とし、抽出する。

2.3 Graph Convolutional Network

GCN は、機械学習とグラフ理論を組み合わせた Graph Neural Network（GNN）の派生物である[8]。GCN は、蛋白質機能予測の分野で優秀な結果を出した DeepFRI[7]のように、様々な分野で優れた結果を達成している。

本研究では、蛋白質表面の特徴抽出において GCN が有効かどうか検討する実験を行う。本研究では DeepFRI の研究を基に、GCN のモデルを構築する。DeepFRI では、アミノ酸残基のC_α原子を基に、特徴行列、隣接行列が作られるが、本稿では蛋白質表面上の L 個の特徴点データ間の隣接関係を表す隣接行列$\mathrm{A} \in \mathbb{R}^{L \times L}$、$L$個の特徴点データの特徴を表す特徴行列$\mathrm{X} \in \mathbb{R}^{L \times A}$($A$ = 特徴数)を用いて、蛋白質表面グラフを表現する。特徴行列を構成する特徴データは、法線ベクトル、静電性、疎水性、温度因子、曲率である。これらの入力から GCN を通して得られた蛋白質構造隠れ表現を、活性化関数として softmax 関数を持つ全結合層に通すことで、各リガンドへの結合確率予測の出力を得る。

3. 実験

3.1 実験内容

PROLIPS の有用性を蛋白質リガンド結合予測実験で検証した。比較手法として、蛋白質リガンド結合親和性予測のための cpi-cnn [9] を使用した。cpi 予測モデルは、アミノ酸配列から 1DCNN で特徴抽出を行い、リガンド smiles 配列から GNN で特徴抽出し、それらを統合して結合親和性予測を行う。

また、実験で使用したデータセットは、PDBbind の既知の蛋白質リガンド結合ペアで構成された eF-site のデータである。これらのデータから特徴点抽出等(L = 2000,1700,1500,1300)の加工を行い、隣接行列での点毎の隣接判定を行う際の閾値として 5,10,15Åを用いて、グラフデータを作成した。

cpi-cnn モデルでは、同じペアでリガンド smiles 配列と蛋白質のアミノ酸配列を準備した。ここでは、データ数のクラス間不均衡に対応するため、各蛋白質に結合するリガンド 1 種類、結合しないリガンド 3 種類を学習データとして与えている。

評価方法は、AUC 及び AUPR を使用した。本研究では、クラス間のデータ数の不均衡に配慮して、データセットを収集しているが、ある程度の偏りは生じるため、macroAUC を利用している。

3.2 実験結果と考察

PROLIPS と cpi-cnn との、AUC および AUPR 値を Table 1 に示す。Table 1 の結果から明らかなように、PROLIPS の AUC および AUPR 値は、cpi-cnn のそれを上回っている。この結果から、提案された手法が優れた性能を発揮していることが示される。

これは、三次元蛋白質表面データのグラフ表現を学習することが、蛋白質-リガンド結合親和性を予測するために効果的であることを示唆している。

Table 1. The result of PROLIPS and cpi-cnn.

	PROLIPS	cpi-cnn
AUC	0.946	0.593
AUPR	0.642	0.334

4. 結論と今後の展望

本論文では、蛋白質リガンド結合予測手法である PROLIPS を提案し、GCN を用いて蛋白質表面の性質と構造に焦点を当てた。PROLIPS を評価するため 1DCNN を用いたアプローチと比較し、蛋白質とリガンドの結合予測の精度の検証を行った。結果は、PROLIPS が高い AUC および AUPR 値を達成し、蛋白質表面情報に焦点を当てることが効果的であることを示している。今後の研究の方向性として、Grad-CAM[10]などの機械学習の解釈可能性を向上させる手法を組み込み、モデルが蛋白質表面のどの部分に焦点を当てているかの視覚化を目指すことなどが考えられる。

参考文献

[1] 藤博幸, “タンパク質の立体構造入門”, 講談社.2010.

[2] 藤博幸, “はじめてのバイオインフォマティクス”, 講談社,2006.

[3] Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. “Protein-Ligand Scoring with Convolutional Neural Networks. ”J Chem Inf Model, vol. 57, no. 4, pp942-957, 2017.

[4] Kinoshita, K., Nakamura, H., 2004. "eF-site and PDBjViewer: database and viewer for protein functional sites." Bioinformatics, vol. 20, pp. 1329–1330.1, 2004.

[5] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.,"The protein data bank." Nucleic Acids Research, vol. 28, pp. 235–242, 2000

[6] Wang, R.; Fang, X.; Lu, Y.; Wang, S. "The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures." Journal of Medicinal Chemistry, vol. 47, issue 12, pp. 2977-2980, 2004.

[7] Gligorijević, V., Renfrew, P.D., Kosciolek, T. et al. "Structure-based protein function prediction using graph convolutional networks." Nature Communications, vol. 12, article number: 3168, 2021.

[8] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907, 2016.

[9] Tsubaki, M, Tomii K, and Sese J. "Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences." Bioinformatics 35.2 : 309-318, 2019.

[10] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision, pp. 618-626, 2017.