81. 深層学習を用いたタンパク質－リガンド結合親和性予測

清水　謙多郎

東京大学 大学院農学生命科学研究科 応用生命工学専攻 生物情報工学研究室

Key words：タンパク質，リガンド，結合親和性予測，深層学習

緒　言

タンパク質とリガンド（低分子化合物）の結合親和性（affinity）の解析は、薬剤および有用酵素の設計において、基盤となる重要な技術である。結合親和性をもとに、特定のタンパク質に結合するリガンドの探索、あるいは、特定のリガンドに結合するタンパク質の探索が行われるが、こうしたスクリーニングを広範に行うには、生化学実験や構造解析は多くの時間と労力を必要とするため、生命情報工学による予測が期待されている。スクリーニングの基盤となるタンパク質とリガンドの結合予測（結合するかどうかの予測）は、タンパク質の配列から予測する手法が古くから提案されてきたが、予測精度は低く、結合親和性の値まで予測することは困難である。構造情報を用いた予測では、従来、タンパク質－リガンドの相互作用エネルギー計算に基づく方法が中心であったが、大量の構造情報が利用できるようになってきた近年は、機械学習が用いられるようになってきた。しかしながら、構造は、潜在的に大きな情報量をもつため、人間が事前に構造特徴を絞り込む必要が生じ、学習手法の限界もあって、十分な精度向上が達成できないでいる［1～3］。最近、深層学習を用いて構造を直接的に学習させる結合予測手法が報告されているが［4，5］、これらは、相互作用エネルギー計算に基づく方法より高い精度を達成しているものの、現実のスクリーニングに用いるには精度が十分でなく、結合親和性の値も予測できていない。精確な結合親和性は、スクリーニングに加えて、リガンドの作用機序、薬剤においては薬効や副作用の解析を行う上でも非常に重要である。

本研究では、タンパク質とリガンドが結合する可能性のある部位の 3 次元（3D）構造から、深層学習によって、それらの結合親和性の値を高精度で予測するシステムを開発する。実験によって得られた結合構造を学習させ、ドッキング予測などで得られた候補構造に対して予測を行う。予測には、結合構造あるいは候補構造の周辺の 3D 空間をグリッドに区切り、グリッド内の原子の属性（原子種など）を入力とする。このような予測には高度な深層学習が不可欠であり、本研究では、画像認識の分野で実績のある convolutional neural network（CNN）を 3D 空間に適用できるよう拡張した手法 3D-CNN を開発する。

方　法

1.　学習データセットの取得

学習に用いるデータは、タンパク質－リガンドの結合構造と実験によって計測された親和性の値を収集した PDBbind データベース（http://www.pdbbind.org.cn/）から取得した。性能評価は、PDBbind のデータのうち、信頼度の高いものを選択して行った。PDBbind は、現在、利用可能な結合親和性のデータベースとしては最大のものであり、人手によってキュレーションが行われている。本研究では、タンパク質－リガンドの 17,679 個（version 2019、2020 年 3 月の時点で最新版を利用）の結合構造と親和性の値のうち、PDBbind データベースが“refined set”と定義している精度が高いもの 4,825 個を選び、さらに配列冗長性を排除して 315 個を選択し、学習および予測のためのデータセットとした。また、後述するように、他の手法との性能比較のために、PDBbind “core set”（285 個のデータ）を利用した。

2. 入力データの処理

タンパク質－リガンドの結合構造を深層学習に入力するため、タンパク質－リガンド結合部位の 1 辺 24 Å の立方体（ボックス）を抽出し、リガンドに最も近い原子を特定して、その原子をもつアミノ酸残基をボックスの中心に置いた。タンパク質－リガンド結合部位の配向に依存しない予測を実現するため、各ボックスは、CA-N および CA-C 結合によって形成される平面が xy 平面を形成し、CA-C$_\beta$結合が正の内積をもつ直交方向が正の z 軸となるように配置する。

ボックスは、0.5 Å のボクセルに分割し、そこにタンパク質とリガンドそれぞれにおける原子種（炭素、酸素、窒素、硫黄）とその結合状態に対応して（アミノ基の窒素、カルボキシル基の酸素、芳香族の炭素など）11 個のチャネルを設定した。各原子の電子を割り当てることで擬似的なタンパク質の電子密度を再現した。

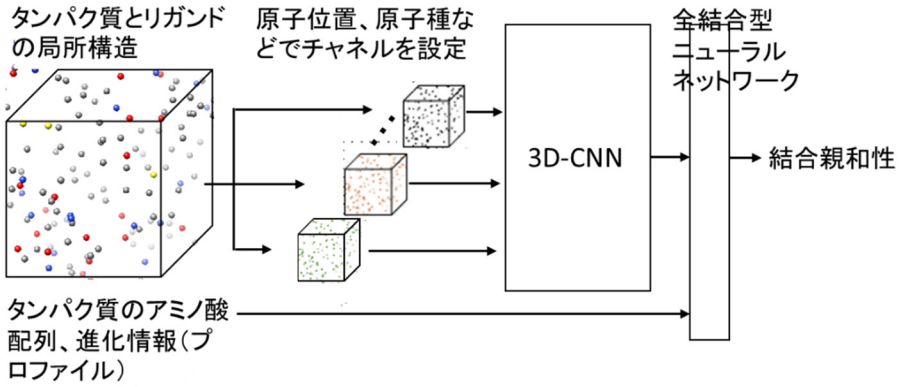

図 1. タンパク質－リガンド結合親和性予測システムの基本構成

タンパク質－リガンド結合部位の 3D 空間を 0.5 Å のボクセルに区切り、ボクセル内の原子をもとに電子密度を形成し、深層学習 3D-CNN の入力とする。この 3D-CNN による 3D 構造とあわせて、タンパク質のアミノ酸配列、類縁配列によって構成されるプロファイル、ドメイン情報を学習させ、タンパク質－リガンドの結合親和性を予測する。

また、構造と合わせて、PSI-BLAST を用いて類縁タンパク質のマルチプルアラインメントを構築し、プロファイルを作製した。プロファイルはネットワークの最後のステージ（3. に記載）に入力することで、進化情報を加味した予測を行った。

3. 深層学習による予測プログラム

タンパク質－リガンド結合部位の周辺を直接学習させるため、CNN を 3 次元に拡張した 3D-CNN を開発した。3D-CNN は、3D 構造そのものを特徴とすることができ、あらかじめ人手で特徴を設定する必要がないという利点がある。また、例えば、原子間距離に基づく 2D 表現に変換したりする必要もない。図 2 に、3D 構造を学習するネットワークの構成を示す。

畳み込みにより周辺構造の特徴が把握され、プーリングにより入力のダウンサンプリングを実行して特徴が濃縮される。畳み込みとプーリングの繰り返しにより、結合構造のすべての位置にわたる情報が統合される。次に、統合された情報が Softmax 分類器に送られ、最終的な予測が行われる。結合親和性（Kd 値）は 10 段階に離散化して予測する。トレーニングでは、確率的勾配降下法とバックプロパゲーションを使用して多項ロジスティック損失の最小化を行った。

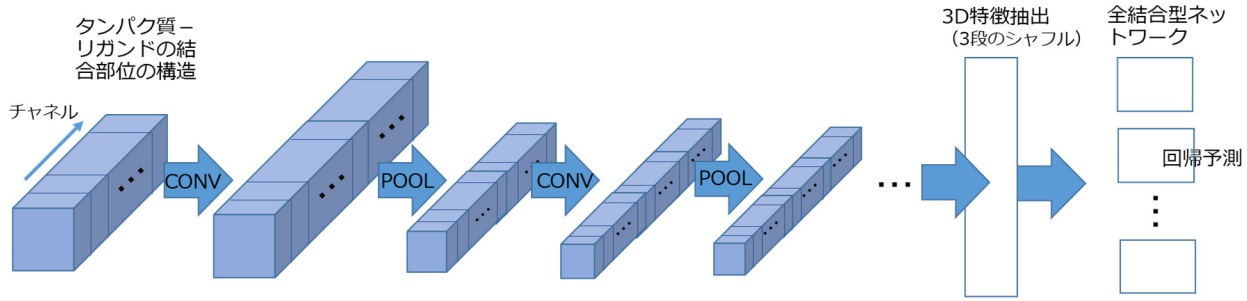

図 2. ネットワークの構成

11 個のチャネルの 3D 情報を深層学習 3D-CNN に入力する。3D-CNN は、畳み込み層（CONV）、プーリング層（POOL）の 4 段の繰り返し（図ではその一部を示す）である。最後の段は回帰予測を行うブロックである。

4. 予測の実行とパラメータの調整

1. の学習データセットを対象にテスト予測を行い、3D-CNN の畳み込み層およびプーリング層の構成、フィルターの設計、その他各種パラメータの最適化を行った。

結果および考察

1. 3D 構造の表現

3D 空間を 0.5 Å のボクセルに分割し、電子を配置して 3D-CNN の入力とした。図 3 は、実際のリガンド結合タンパク質（PDB ID : 1A0Q）で電子密度がどのように再現されているかを示したものである。原子の位置と原子種を入力とする従来の方法［5］に比べて、より多くの情報を多数のチャネルに分けて表現することができる。

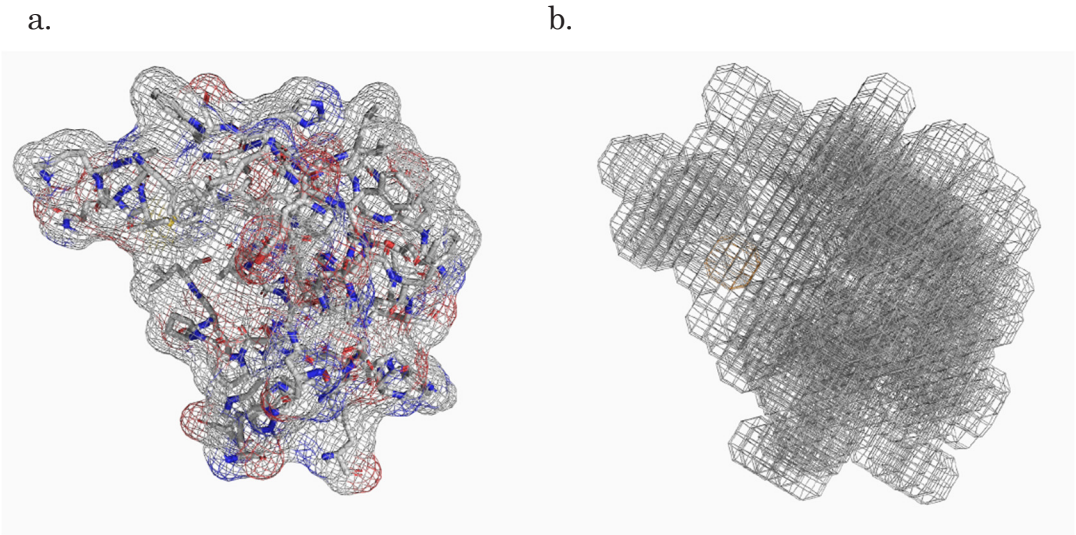

図 3. タンパク質（PDB ID : 1A0Q）の表現

a） 実際の電子密度。

b） 3D-CNN に入力するためのボクセル表現（この図では 1 Å のボクセルを示す）。

2. モデル構造の評価

親和性予測については、3-fold cross validation を用いて評価を行った。ターゲットは訓練セットとの類似性を避けるため、ターゲットをクラスタ化することにより、訓練およびテストデータセットを構築した。予測誤差の指標としては、Mean Absolute Error（MAE）、Root Mean Square Error（RMSE）および回帰予測の結果の評価値として実験値との相関係数を用いた。

3. 予測結果

図 4 は、1. で述べた PDBbind のデータに対して予測した親和性と実験によって得られた親和性（どちらも pK 値）の関係を示したものである。相関係数は 0.782 であり、3D 構造と従来の配列ファイルのみを用いた比較的簡易な方法で非常に高い予測性能が得られていることがわかる。

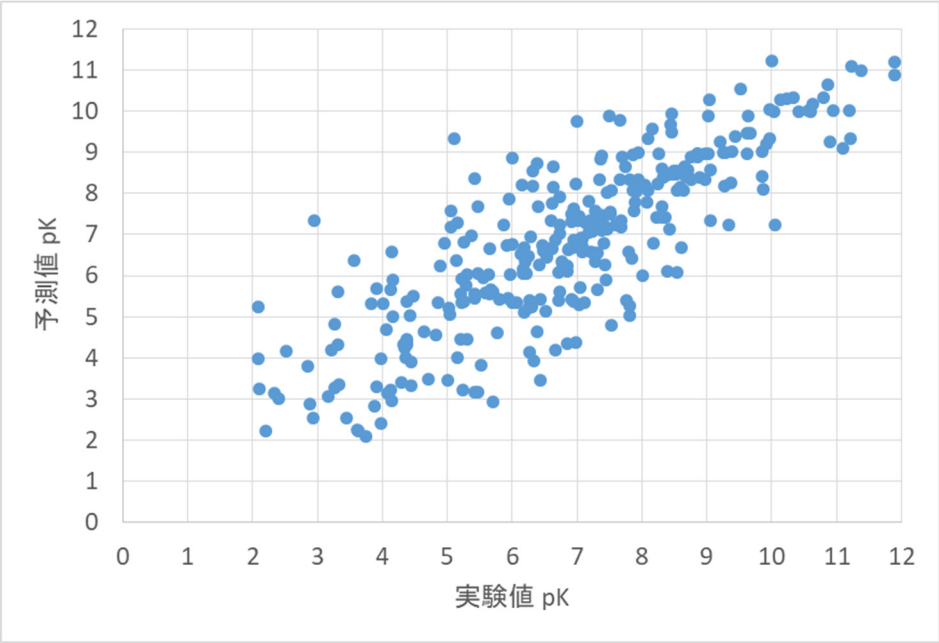

図 4. 結合親和性予測の結果と実験値との比較

横軸は PDBbind に登録された結合親和性（pK 値）、縦軸は本手法で予測した結合親和性を示す。

表 1 に、PDBbind core set を用いて性能評価を行った結果を示す。また、従来の手法のうち、とくに高い精度を実現した Deep Atom［6］の性能を合わせて示す。実験値との相関係数は、上記と異なるデータセットを用いているため、図 4 と値は異なる。表 1 より、本研究は Deep Atom より高い精度で予測できていることがわかる。本研究の手法は、Deep Atom に比べてチャネル数は少ないが、電子密度をきめ細かく学習に取り込んでおり、予測精度の向上に貢献しているものと考える。

表 1. 予測性能

	MAE	RMSE	相関係数
本手法	1.007	1.304	0.798
DeepAtom［6］	1.039	1.318	0.807

MAE は、Mean Absolute Error、RMSE は、Root Mean Square Error、相関係数は、Pearson correlation coefficient を示す。数値は 5 回実行した結果の平均値である。

4. 今後の展開

本研究で開発した高精度の結合親和性予測は、化合物および標的タンパク質の探索、ドッキング予測のポーズ選択に有用と考えられ、今後、これらの応用に取り組んでいきたいと考えている。また、現在、敵対的生成ネットワーク (GAN) を用いて、リガンドのデザインをタンパク質との結合部位とあわせて行う手法を開発している。

共同研究者・謝辞

本研究の共同研究者は、東京大学大学院農学生命研究科生物情報工学研究室の松本知也と尹軒宇、農研機構の Cao Wei である。

文　献

1) Ertl P, Lewis R, Martin E, Polyakov V. In silico generation of novel, drug-like chemical matter using the LSTM neural network. arXiv preprint. 2017 Dec 20. arXiv:1712.07449
2) Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic chemical design using a data-driven continuous representation of molecules. ACS Central sci. 2018 Feb 28;4(2):268-276. Epub 2018 Jan 12. PMID: 29532027 DOI: 10.1021/acscentsci.7b00572
3) Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular De-Novo Design Through Deep Reinforcement Learning. J Cheminform. 2017 Sep 4;9(1):48. PMID: 29086083 DOI: 10.1186/s13321-017-0235-x
4) Wójcikowski M, Ballester PJ, Siedlecki P. Performance of Machine-Learning Scoring Functions in Structure-Based Virtual Screening. Sci Rep. 2017 Apr 25;7:46710. PMID: 28440302 DOI: 10.1038/srep46710
5) Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-Ligand Scoring With Convolutional Neural Networks. J Chem Inf Model. 2017 Apr 24;57(4):942-957. Epub 2017 Apr 11. PMID: 28368587 DOI: 10.1021/acs.jcim.6b00740
6) Yanjun L, Mohammad AR, Chenglong L, Xiaolin L, Dapeng W. DeepAtom: A Framework for Protein-LigandBinding Affinity Prediction. 2019 Dec 1. arXiv:1912.00318