様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 ２１ 日現在

機関番号：１２６０８
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１１５２４

研究課題名（和文）３次元畳み込みニューラルネットワークによる構造ベース化合物活性予測

研究課題名（英文）Structure-based ligand activity prediction using 3-dimensional convolution neural network

研究代表者
石田　貴士（Ishida, Takashi）

東京工業大学・情報理工学院・准教授

研究者番号：４０５０８３５５

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：多くの計算資源を必要とするドッキング計算を用いず、新規のタンパク質に対しても適用可能な、機械学習を用いた構造ベースの化合物活性予測手法の開発を行った。タンパク質のリガンド結合ポケット構造をグラフで表現し、グラフ畳み込みニューラルネットワークを利用することでエンドツーエンドでの学習を行うことで、既存のタンパク質の配列情報のみを利用した予測手法に比べてより高精度な予測を達成した。またAutoDock Vinaによるドッキング計算による予測と比較してもより高速な予測と良好な予測精度を実現した。

研究成果の学術的意義や社会的意義
新規のタンパク質に対しても利用可能なタンパク質構造と化合物構造を入力とした深層学習ベースの化合物活性予測手法を新たに開発した。これにより、実験情報のない新規のタンパク質に対しても化合物活性予測の適用が可能となり、応用可能な範囲が広がった。しかし、残念ながらその予測精度はまだ不十分であり、より実用的な利用にはさらなる今後の改良が必要となっている。

研究成果の概要（英文）：We developed a novel machine-learning based compound activity prediction using binding pocket information. The method converts a binding pocket structure of a target protein to a structure-graph and uses a graph convolutional neural network for end-to-end learning. The proposed method achieved better accuracy than a method only using protein sequence information. Additionally, the proposed method was more accurate and fast than a docking calculation using Autodock Vina.

研究分野：バイオインフォマティクス

キーワード：深層学習　化合物活性予測　ヴァーチャルスクリーニング　タンパク質立体構造　リガンド結合ポケット

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

創薬に必要とされるコストは年々増加の一途を辿っており、現在では十数年に渡る開発期間と三千億円とも言われる膨大な費用が必要となっている。そのため、計算を用いた薬剤開発プロセスの効率化には大きな期待が寄せられているが、その中でも精力的に技術開発が行われている分野の一つが標的タンパク質に対して薬剤活性のあるヒット化合物の探索である。この計算機によるヒット化合物の探索はヴァーチャル・スクリーニングとも呼ばれるが、現在提案されている手法は、機械学習により既知の化合物活性情報から予測モデルを構築するリガンドベース手法と物理化学エネルギー関数により結合の強さを推定するドッキング計算手法の２つに大別される。

機械学習によるリガンドベース予測手法では、十分な既知データがあれば、リガンド化合物のみを入力として高い精度での予測が可能であり、よく知られた標的については既に実用的なレベルの予測モデルが提案されている。しかし、教師あり機械学習を用いるため、対象となるタンパク質についての化合物活性情報が大量に必要となり、新たな標的に対しては十分なデータが利用できないためその予測精度が大きく低下してしまう。一方、ドッキング計算による構造ベース予測手法では、標的タンパク質について実験による既知の活性情報がなくとも適用が可能で、予測される化合物も、機械学習による手法に比べると多様なものが得られることが多い。しかし、ドッキング計算では化合物の内部自由度も考慮した広大な構造空間から最適な結合状態を探索する必要があり、その計算量が大きいという問題がある。また、その予測精度も機械学習による予測に比べると多くの場合劣った結果となっている。

このように現在のヴァーチャル･スクリーニング技術では、どんな標的に対しても適用が可能で高速、高精度な予測手法が存在しない。実用的には適用可能性の問題は特に重要で、予測精度の低さは理解しながらもドッキング計算が利用されている。一方、現在公的なデータベースに蓄積される実験情報はどんどんと増加しており、データベースに登録された実験情報の数は現在では 1000 万件を超えている。そのため、直接標的となるタンパク質についての活性情報がない場合でも、これらの情報を流用することが可能となれば機械学習を利用した高精度での予測も可能となると考えられる。

２．研究の目的

本研究では化合物だけでなく、標的となるタンパク質の立体構造を機械学習の入力とすることで、新規の標的タンパク質に対しても適用可能で、ドッキング計算よりも高速、高精度な「機械学習による構造ベース」の予測手法の開発を目指す。

化合物活性予測の際にタンパク質の立体構造情報を機械学習の入力に取り込む試みは既に幾つかの研究でも行われているが、それらは対象を一つのタンパク質ファミリーに限定し、結合ポケット内のアミノ酸の変異情報を入力ベクトル化して用いるといった汎用性の低いものであった。これは、タンパク質立体構造の情報を直接機械学習の入力としてしまうと入力の次元が高くなりすぎ学習が困難となるため、単純なベクトル列への変換が必要であったためである。

それに対して、本研究では近年注目を集める深層学習を用いることで、タンパク質構造情報を効果的に機械学習の入力として取り込むことを目指す。

３．研究の方法

本研究では近年注目を集める深層学習の一手法である３次元畳み込みニューラルネットワーク（3D-CNN）を用いることで、タンパク質構造情報を効果的に機械学習の入力として取り込むことを目指す。画像認識で大きな成功を納めた２次元の畳み込みニューラルネットワークを３次元に拡張した 3D-CNN は物体認識などの分野で応用が始まっており、層が進むに従って高次の特徴を学習する表現学習が可能となる技術である。生体高分子に対する応用は既にドッキング結果の評価など幾つかの研究が発表されており、我々のグループも結合ポケット予測に利用することで、良い予測精度が達成できることを確認している。

本研究の独自性は、この 3D-CNN によるタンパク質立体構造情報と化合物の構造情報をマルチモーダルな情報としてニューラルネットワークの入力とすることで予測モデルを構築することであり、従来では困難であったタンパク質立体構造情報の直接的な取り込み、学習を実現する点である。また、本研究が扱う生体高分子に対する 3D-CNN の利用技術が進展すると、核酸、ペプチドなどの他の生体高分子についても同様の応用が可能となり、立体予測構造モデルの評価などの他の問題にも利用が可能になると考えられ、その波及効果は大きなものであると考えられる。

４．研究成果

研究当初はタンパク質立体構造の表現として原子種や物理化学的特徴を負荷したボクセルを利用し、データを処理するための深層学習の構造としては３次元畳み込みニューラルネットワークを用いることを想定していたが、利用可能なデータ数に対してネットワークが複雑になりすぎることが判明し、立体構造の表現としてグラフ構造を利用し、それに対してグラフ畳み込みネットワークを用いることで少ないデータからでも十分な学習を可能とした。また、３次元畳み込みニューラルネットワークを用いたタンパク質立体構造データの処理については、結合予測での利用は困難であったが、予測タンパク質立体構造の品質評価に応用し、そのタスクにおいて非常に良い成果が得られた。

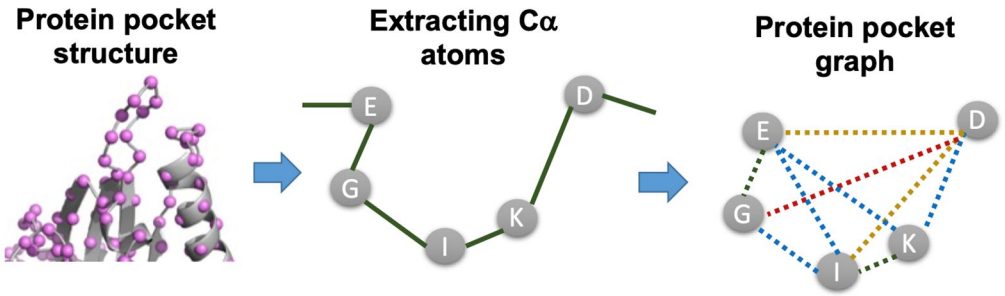

図１．グラフによるタンパク質結合ポケットの表現（Tanebe and Ishida, 2021）

(1) グラフ畳み込みを用いたエンドツーエンドでの化合物活性予測

新規の標的タンパク質への化合物の活性の予測のため、タンパク質構造情報と低分子化合物の双方を機械学習モデルの入力とし、タンパク質ポケット構造情報をアミノ酸残基をノードとし、距離情報をエッジとしたグラフとして捉え、グラフ畳み込みニューラルネットワークを適用したエンドツーエンド表現学習によって活性予測を行う手法を開発した。活性化合物と非活性化合物との距離を調整し単純な機械学習での予測を困難とした MUV データセットを用いた評価で、提案手法はアミノ酸配列情報のみを用いた既存手法に比べ精度の向上を示し、また、AutoDock Vina を用いたドッキング計算比べてもより高精度であることが示された（Tanabe and Ishida, 2019）。

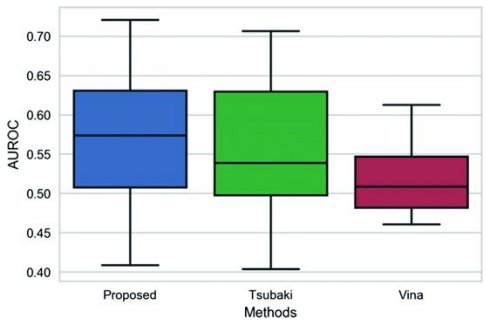

図２．提案手法による予測精度の向上（Tanabe and Ishida, 2019）

(2) ３次元畳み込みニューラルネットワークを用いた予測タンパク質立体構造品質評価

化合物活性予測のために開発された３次元畳み込みニューラルネットワークを用いた解析技術を転用し、ホモロジーモデリング等によって生成されたタンパク質立体構造モデルの品質評価手法を開発した。提案手法ではタンパク質の各残基についてその環境を切り出して深層学習により評価することで、先行研究に比べより高精度な品質評価を可能とした（Sato and Ishida, 2019）。提案手法は立体構造予測コンテストである CASP11, 12 のデータを利用したベンチマークで既存のどの手法よりも良い性能を示した。また、更に２次構造予測の結果や進化的プロファイルを入力に追加することで改良を行い、Web サーバとして生物学の研究者などの

一般のユーザにも利用が可能なかたちで機能の提供を行っている。

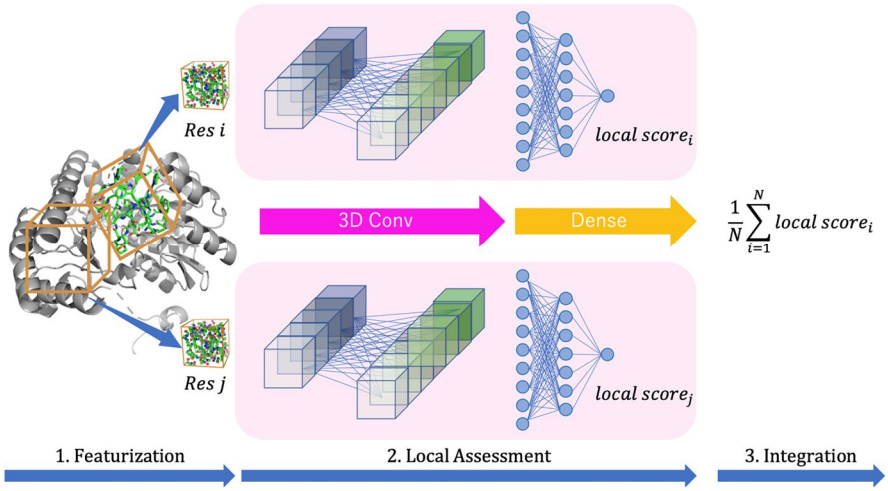

図３．予測タンパク質立体構造品質評価手法の概要（Sato and Ishida, 2019）

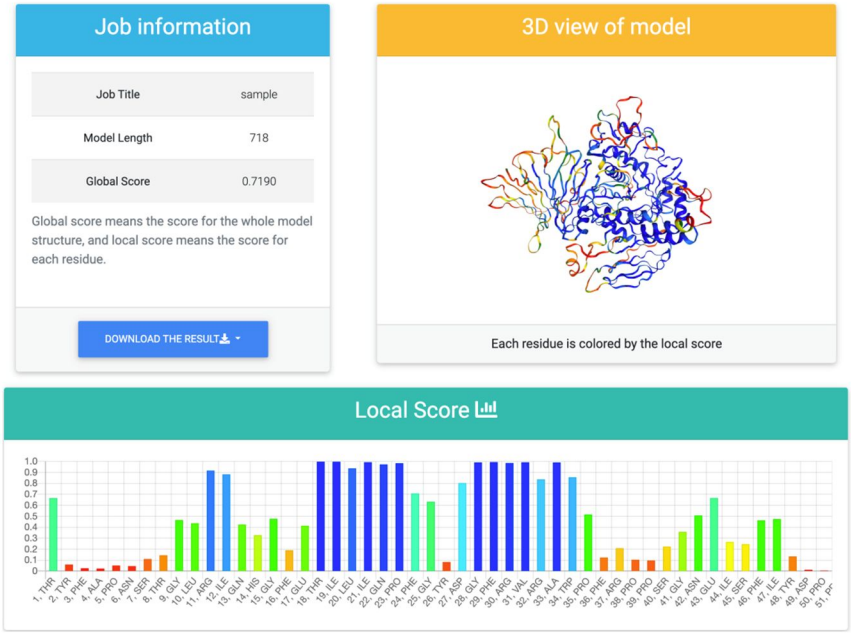

図４．タンパク質予測立体構造品質評価 Web サーバ（Takei and Ishida, 2021）

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Hasic Haris、Ishida Takashi	４．巻 61
２．論文標題 Single-Step Retrosynthesis Prediction Based on the Identification of Potential Disconnection Sites Using Molecular Substructure Fingerprints	５．発行年 2021年
３．雑誌名 Journal of Chemical Information and Modeling	６．最初と最後の頁 641～652
掲載論文のDOI（デジタルオブジェクト識別子） 10.1021/acs.jcim.0c01100	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Makigaki Shuichiro、Ishida Takashi	４．巻 18
２．論文標題 Sequence alignment generation using intermediate sequence search for homology modeling	５．発行年 2020年
３．雑誌名 Computational and Structural Biotechnology Journal	６．最初と最後の頁 2043～2050
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.csbj.2020.07.012	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Takei Yuma、Ishida Takashi	４．巻 8
２．論文標題 P3CMQA: Single-Model Quality Assessment Using 3DCNN with Profile-Based Features	５．発行年 2021年
３．雑誌名 Bioengineering	６．最初と最後の頁 40～40
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/bioengineering8030040	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Makigaki Shuichiro、Ishida Takashi	４．巻 36
２．論文標題 Sequence alignment using machine learning for accurate template-based protein structure prediction	５．発行年 2019年
３．雑誌名 Bioinformatics	６．最初と最後の頁 104～111
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/btz483	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Sato Rin、Ishida Takashi	４．巻 14
２．論文標題 Protein model accuracy estimation based on local structure quality assessment using 3D convolutional neural network	５．発行年 2019年
３．雑誌名 PLOS ONE	６．最初と最後の頁 221347～221347
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0221347	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Tanebe Toshitaka、Ishida Takashi	４．巻 11644
２．論文標題 End-to-End Learning Based Compound Activity Prediction Using Binding Pocket Information	５．発行年 2019年
３．雑誌名 2019 International Conference on Intelligent Computing	６．最初と最後の頁 226～234
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-26969-2_21	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 松村真里，石田貴士
２．発表標題 タンパク質配列情報と薬剤結合部位構造情報を用いた新規タンパク質に対する深層学習リガンド結合予測
３．学会等名 第八回生命医薬情報学連合会
４．発表年 2019年

１．発表者名 Rin Sato and Takashi Ishida
２．発表標題 GCMQA: Graph convolutional neural network for model quality assessment
３．学会等名 第八回生命医薬情報学連合会
４．発表年 2019年

１．発表者名
種部俊孝、石田貴士

２．発表標題
タンパク質ポケット構造情報を考慮した機会学習によるリガンド結合予測

３．学会等名
生命医薬情報学連合大会2018年大会

４．発表年
2018年

１．発表者名
種部俊孝、石田貴士

２．発表標題
ポケット構造情報を考慮したエンドツーエンド表現学習によるリガンド結合予測

３．学会等名
情報処理学会第57回BIO研究発表会

４．発表年
2019年

１．発表者名
佐藤倫、石田貴士

２．発表標題
グラフ畳み込みを用いたタンパク質予測立体構造の評価手法の開発

３．学会等名
情報処理学会第57回BIO研究発表会

４．発表年
2019年

１．発表者名
佐藤倫、石田貴士

２．発表標題
深層学習を用いたタンパク質予測立体構造モデルの評価

３．学会等名
情報処理学会第54回BIO研究会

４．発表年
2018年

１．発表者名 佐藤倫、石田貴士
２．発表標題 3次元畳み込みニューラルネットワークを用いたタンパク質予測立体構造の評価手法の開発
３．学会等名 生命医薬情報学連合大会2018年大会
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関