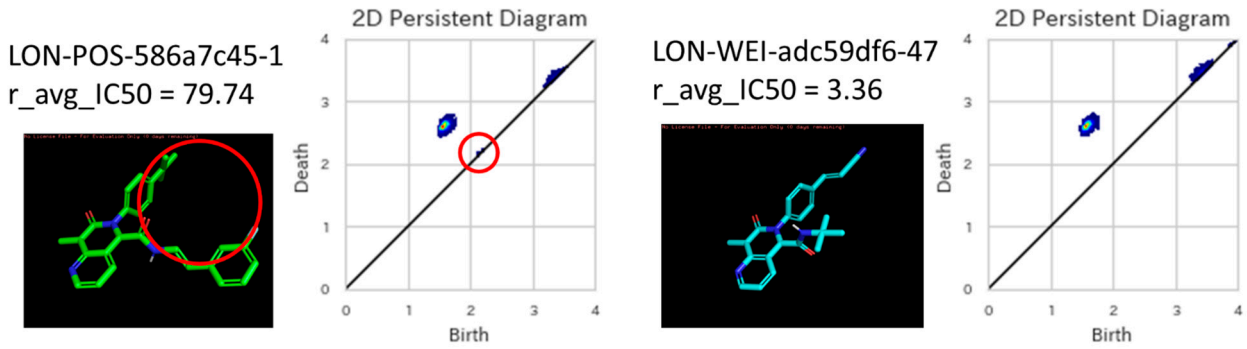
196　タンパク質の阻害に寄与する3次元構造の探索方法開発	江崎　剛史

【目的】時間と費用を削減して効率的に創薬を進めるため、薬の化学構造と生理活性の関係性を見出す研究への期待は大きく、近年は特に人工知能を始めとした計算科学手法が注目されている。昨年末から世界中で脅威をふるっている新型コロナウィルスは治療薬の早急な開発が求められており、承認までの早さから既存薬の転用に向けた検討が行われているが、順調とは言えない。問題として「患者数が少なく、治療薬の十分な効果と安全性を確認できない」ことがあるため、医薬品候補を探す計算科学手法の重要性が改めて認識されてきている。薬の性質はタンパク質との立体的な相互作用が大きく関わっているが、機械学習手法に使用される記述子は、ほとんどが 2 次元の情報である。そこで、従来の構造記述子と 3 次元の位相的な距離を特徴量として使用し、Mpro の阻害性を予測する手法の検討を行った。

【方法】COVID19 のメインプロテアーゼである Mpro の阻害性試験のデータを収集し、50%阻害濃度の実験値が存在する 659 化合物のデータを分析対象とした。部分構造を表す構造特徴量として、4 つの Fingerprint を算出し、Mpro の阻害予測に最も適した特徴量を選択した。また、構造特徴量だけでは見つけることができない情報を取得するため、Mpro の立体構造を取得してドッキングシミュレーションを行い、化合物の 3 次元結合座標を取得した。発生させた化合物の 3 次元構造から位相的な特徴量を算出する Persistent Homology を実装した。選択した構造特徴量と位相的特徴量を合わせて、Random Forest によって判別モデルを構築した。

【結果】構造特徴量の検討により MACCS と Morgan が適した構造特徴量であると決定し、位相的特徴量と合わせて Random Forest による判別モデルを構築したところ、多くの特徴量で 80%を超える正解率を出した。著しい精度の違いはなかったが、Morgan が同じであるにも関わらず、阻害性が大きく異なる 2 つの化合物を比べたところ、Persistent Homology によって立体的な特徴量を得ることができたことが確認できた。タンパク質との立体的な結合を表すためには、従来使用していた構造特徴量だけでなく、位相的特徴量も組み合わせて検討を行うことが望ましいことが示唆された。

位相的距離を用いた化合物の立体構造の特徴量化による重要構造の探索



196. タンパク質の阻害に寄与する 3 次元構造の探索方法開発

江崎　剛史

滋賀大学 データサイエンス教育研究センター

Key words：構造活性相関，構造記述子，機械学習

緒　言

1 つの薬が承認されるまでにかかる期間は 15 年から 20 年、研究開発費は 2,000 億円以上とも言われている。時間と費用を削減して効率的に創薬を進めるため、薬の化学構造と生理活性の関係性を見出す研究への期待は大きく、近年は特に、人工知能を始めとした計算科学手法が注目されている［1]。昨年末から世界中で脅威をふるっている新型コロナウィルスの治療薬の早急な開発が求められており、承認までの早さから既存薬の転用に向けた検討が行われているが、順調とは言えない。問題として「患者数が少なく、治療薬の十分な効果と安全性を確認できない」ことがあり、医薬品候補を探す計算科学手法の重要性が改めて認識されてきている。

薬の性質はタンパク質との立体的な相互作用が大きく関わっており、その関係を 2 次元の情報のみを使って表すことは困難である。だが、多くの機械学習手法を適用する際に使用する化合物の情報として使われている構造記述子は、ほとんどが 2 次元情報のみである。その理由として、タンパク質と結合する際の立体構造を特徴量に組み込むことが困難であることが挙げられる。そこで、従来の構造記述子と 3 次元の位相的な距離を特徴量として使用し、Mpro の阻害性を予測する手法の検討を行った。本研究により、従来の構造記述子だけでは得ることのできなかった 3 次元的な構造情報を特徴量に組み込むことができ、タンパク質の阻害を予測するために 3 次元構造が有効であることが示唆された。

方　法

1. データセット

COVID19 のメインプロテアーゼである Mpro の阻害性試験のデータは、随時更新・公開されているため（https://discuss.postera.ai/ai/）、収集してサーバーに格納し、本研究で使用するデータセットとした。本データセットには Mpro 阻害試験のデータが 1,495 化合物分、その中でも 50%阻害濃度の平均値が存在する 659 化合物のデータを分析対象とした。化合物の特徴量を検討する際に、膨大な特徴量が発生して計算時間がかかるため、特徴量の検討用に 200 化合物をランダム選択した（検証用データセット）。選んだ特徴量で Mpro 阻害性を予測するモデルの構築を行った際には、全データを用いた（予測用データセット）。

2. 構造特徴量の算出

部分構造を表す特徴量の算出方法として、MACCS Fingerprint、Morgan Fingerprint、Atom-Pair Fingerprint、Pharmacophore Fingerprint の 4 手法を検討し、Mpro の阻害予測に最も適した特徴量を選択することとした。これらの構造特徴量は、SCIQUICK（ver. 1.0, FUJITSU）を使用して算出した。MACCS は予め用意された 166 の部分構造を持つか否かを 2 値で示す 166 のベクトルを算出する方法であり、最も広く使用されている。Morgan は化合物内の指定した原子からある半径を持つ円を考え、その円に入る構造を 2 値でベクトル化する。本研究では、半径を 4、ベクトルの長さを 2,048（radial＝4、bits＝2,048）とした。Atom-Pair は原子間の結合を特徴量とする方法であり、8,388,608 のベクトルを算出した。Pharmacophore は薬理学の分野において有効であると想定されている原子構造同士の距離を特徴量とするものであり、310 のベクトルを出力した。

3. ドッキングシミュレーション

Mpro は立体的な構造を持つタンパク質であり、そのタンパク質に結合する化合物も立体構造を確認することが重要であると考えられる。構造特徴量だけでは見つけることができない情報を取得するため、Protein Data Bank［2］より Mpro の立体構造を取得し（PDB id: 6LU7）、結合部位における化合物のドッキングシミュレーションを行った。化合物の 3 次元化は RDKit（ver. 2021.09.5）を使用し、結合座標の確認のために Protein Plus Server［3］を、ドッキングシミュレーションには AutoDock（ver 4.2.6）を使用した［4］。

4. 位相的特徴量の算出

発生させた化合物の 3 次元構造を用い、位相的な特徴量を算出する Persistent Homology を実装するために、risper（ver. 0.6.2）を使用した。Persistent Homology は各原子を中心とした半径の円を考え、徐々に半径を大きくしていくことで生じる空間と、さらに大きくすることで消滅する空間の情報を特徴量とする手法である［5］。タンパク質の構造推定などにも活用されつつあり、タンパク質との結合状態を分析することを目指す本研究において、有益な情報となると考えられた。

5. 判別モデルの構築

Mpro の阻害性を予測するために、予測用データセットに対して特徴量の算出を行った。発生させた特徴量から分散のない特徴量を除外するため、scikit-learn（ver. 1.0.2）の関数 VarianceThreshold（threshold＝0）を使用した。その後、残った特徴量を用い、Random Forest によって判別モデルを構築した。パラメータはグリッドサーチ（GridsearchCV、cv＝10）によって最適なパラメータを選択した。

結果および考察

1. 構造特徴量の算出

検討を行った構造特徴量では、Fingerprint の算出時間に大きな差があることが分かった。検証用データセットの 200 化合物に対して MACCS と Morgan が 0.5 秒以内で完了したのに対し、Pharmacophore では約 1,000 倍、Atom-Pair にいたっては、2,000 倍程度の時間がかかることを確認した。機械学習を用いた特性予測では、多くの特徴量と前処理、そして機械学習手法の検討を行う必要がある。特徴量の算出に膨大な時間をかけることは、望ましいことではないため、本研究では MACCS と Morgan が適した構造特徴量であると決定した。

2. 位相的特徴量の算出

検討を行った構造特徴量に、新しい情報を追加するため、位相的特徴量の算出を行った（図 1）。位相的特徴量は Fingerprint とは異なる立体構造を想定していたが、構造の特徴を反映させた想定に近い結果を得ることができた。

予測用データセットをトレーニングデータ（70%）、テストデータ（30%）に分割し、トレーニングデータで分散の小さい変数を除外後、Random Forest による判別モデルを構築した。テストデータに対する予測を行ったところ、多くの特徴量で 80%を超える正解率を出した（表 1）。著しい精度の違いはなかったが、Persistent Homology は MACCS と Morgan と比較して、阻害性が高いと予測する傾向にあることが分かった。そして、MACCS や Morgan と一緒に使用することによって、阻害の予測に有効な情報となりえることが示唆された。

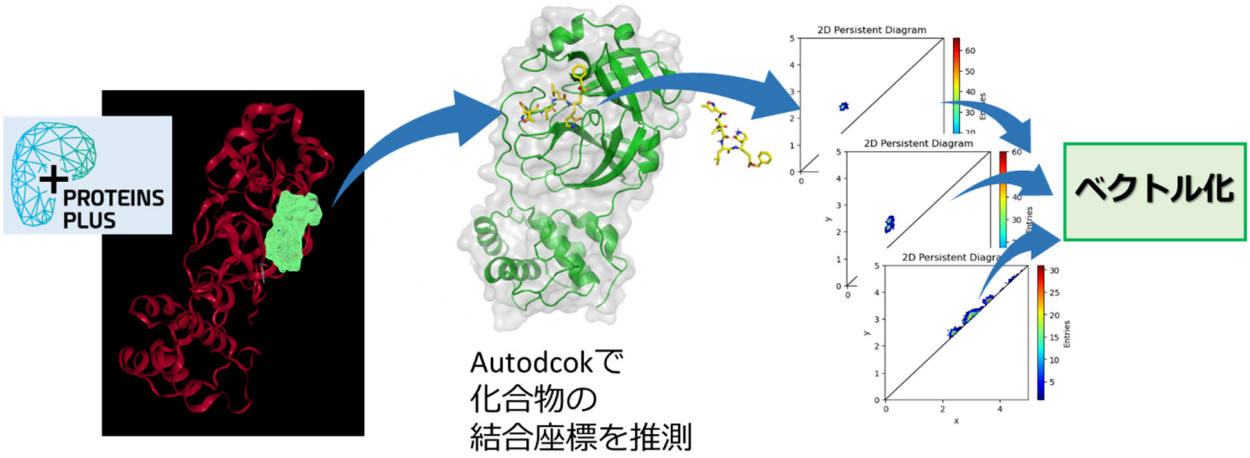


図 1. 位相的特徴量の算出イメージ

Protein Plus で Mpro の阻害に関わる結合部位の推定を行い、AutoDock で化合物の 3 次元結合座標を推測する。得られた 3 次元構造から Persistent Homology を実施し、特徴量としてのベクトルを算出した。

表 1. 阻害性の判別予測結果

	Accuracy	Precision	Sensitivity	F-score
MACCS	0.79	0.82	0.94	0.88
Morgan	0.80	0.83	0.93	0.88
Persistent Homology	0.77	0.77	1.00	0.87
MACCS + Persistent Homology	0.80	0.80	0.98	0.88
Morgan + Persistent Homology	0.80	0.82	0.94	0.88

MACCS は MACCS Finger print、Morgan は Morgan Finger print。MACCS＋Persistent Homology、Morgan＋Persistent homology は、それぞれの Finger print に Persistent Homology を追加して特徴量として使用した。

3. 特徴量の違いによる予測への影響

予測の結果からは、位相的特徴量である Persistent Homology は、予測のために明確な有効性を示すことができなかった。しかし、Fingerprint はある部分構造があるか否かを 2 値で表すものであり、数や立体的な構造は抜けている。実際、Morgan が同じであるにも関わらず、阻害性が大きく異なる 2 つの化合物を比べたところ、Persistent Homology によって立体的な特徴量を得ることができたことが確認できた (図 2)。タンパク質との立体的な結合を表すためには、構造特徴量だけでなく、位相的な特徴量も組み合わせて検討を行うことが望ましいことが推測される。このことを確認し、さらなる精度向上を目指すため、現在はデータ数の拡充を行い、大規模なデータで検討を実施している。

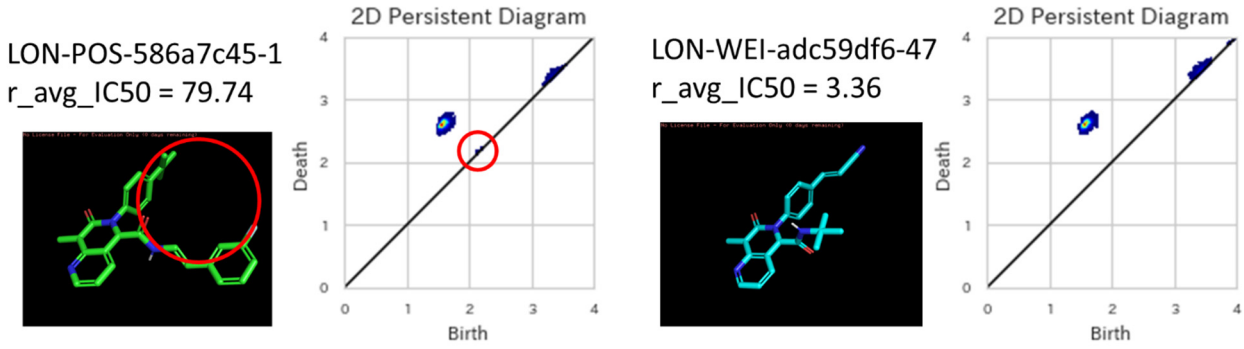


図 2. Persistent Homology によって算出できた特徴的な構造

構造特徴量では同じ特徴量が算出された化合物の位相的特徴量の違いを比較した。LON-POS-586a7c45-1（左図）は立体的に広がった構造を持ち、Persistent Homology で特徴量として捉えられた（赤丸）。その一方で、LON-WEI-abc59df6-47（右図）は閉じた構造を持ち、Persistent Homology では特徴量として算出されなかった。

謝　辞

本研究を行うにあたりご支援を賜った公益財団法人上原記念生命科学財団に深く感謝申し上げる。

文　献

1) Ma, J. Sheridan, R.P. Liaw, A. Dahl, G.E. Svetnik, V. Deep neural nets as a method for quantitative structure–activity relationships, J. Chem. Inf. Model. 2015, 55, 2, 263–274. DOI: 10.1021/ci500747n PMID: 25635324

2) Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235-42. DOI: 10.1093/nar/28.1.235. PMID: 10592235

3) Schöning-Stierand K, Diedrich K, Fährrolfes R, Flachsenberg F, Meyder A, Nittinger E, Steinegger R, Rarey M. ProteinsPlus: interactive analysis of protein–ligand binding interfaces. Nucleic Acids Research. 2020 Jul Volume 48, Issue W1, 02: W48–W53, PMID: 32297936 DOI: 10.1093/nar/gkaa235

4) Santos-Martins D, Forli S, Ramos MJ, Olson AJ. AutoDock4(Zn): an improved AutoDock force field for small-molecule docking to zinc metalloproteins. J Chem Inf Model. 2014 Aug 25;54(8):2371-9. DOI: 10.1021/ci500209e. Epub 2014 Jul 18. PMID: 24931227

5) Xia K, Wei GW. Persistent homology analysis of protein structure, flexibility, and folding. Int J Numer Method Biomed Eng. 2014 Aug;30(8):814-44. DOI: 10.1002/cnm.2655. Epub 2014 Jun 24. PMID: 24902720