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　６ 年　 ６ 月　 ５ 日現在

機関番号：８３９０１
研究種目：若手研究
研究期間：2022～2023
課題番号：２２Ｋ１８００３

研究課題名（和文）機械学習とベイズ推論の融合による免疫受容体タンパク質の設計手法の開発

研究課題名（英文）Design of immunoreceptor protein through the integration of machine learning and Bayesian inference

研究代表者
郭　中梁（Guo, Zhongliang）

愛知県がんセンター（研究所）・システム解析学分野・研究員

研究者番号：２０８７５８１９

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：タンパク質間の結合能を正確に見積もることは，タンパク質の機能を理解し，新しいタンパク質を設計し，病気の治療につながる．しかし，実験によるタンパク質間結合能の測定は時間と費用がかかる．本研究課題では，タンパク質の設計における既存の結合能予測モデルの問題点を掘り下げ，マルチモーダル学習を利用し，タンパク質の立体構造とアミノ酸配列の双方の情報を統合し，高精度かつ高速なタンパク質結合能予測手法を開発した．ベンチマークデータでの予測精度は，従来手法に比べ，Pearson相関係数が0.684から0.904に向上した．またモデルの解析を通じ，マルチモーダル学習の有効性を確認した．

研究成果の学術的意義や社会的意義
近年，ウイルスやがん細胞に結合する抗体またT細胞受容体の設計が注目され，臨床を含め，多くの研究が行われてきた．しかし，設計されたタンパク質とターゲット分子の結合能を実験で測定するには時間と費用がかかる．また，既存の結合能予測モデルの予測精度が実用化に至っていないことも十分に認識されていない．本研究課題で提案した高精度かつ高速な結合能予測モデルは，深層学習を用いたタンパク質設計手法と組み合わせることで，効率的にターゲット分子と結合するタンパク質を発見できることが期待される．タンパク質間相互作用はタンパク質機能の基礎であり，結合能を正確に予測することは生命現象の理解につながる重要なステップとなる．

研究成果の概要（英文）：Accurate protein-protein binding affinity prediction is essential for understanding protein function, designing new proteins for treating diseases. However, experimental measurement of protein binding affinity is time-consuming and expensive. In this study, we addressed the overlooked issues in current binding affinity prediction models in protein design. We developed a high-accuracy, fast prediction method that integrates information from both protein 3D structures and amino acid sequences using multimodal learning. The performance of our model on benchmark datasets showed a significant improvement over existing methods, with the Pearson correlation coefficient increasing from 0.684 to 0.904. Additionally, through model analysis, we confirmed the efficacy of multimodal learning in predicting protein binding affinity.

研究分野：機械学習

キーワード：タンパク質間相互作用　機械学習　タンパク質大規模言語モデル　トポロジカルデータ解析　タンパク質設計　TCR

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

がん免疫において，免疫細胞の一種である T 細胞が重要な役割を果たしている．がん細胞の表面にヒト白血球抗原（HLA）と呼ばれるタンパク質が発現し，がん細胞特異的なタンパク質由来のペプチドが HLA と結合して細胞表面に提示される．T 細胞の表面にあるT細胞受容体分子（TCR）はこのペプチドと HLA の複合体（pHLA）に結合し，がん細胞を認識して攻撃する．近年，がん免疫療法の研究では，TCR や免疫チェックポイントである PD-1/PD-L1 分子の役割が注目されている．特に PD-1 と PD-L1 の結合は免疫反応を抑制するので，PD-1/PD-L1 阻害剤はこの抑制を解除し，きわめて有効ながん免疫療法として実証された．しかし，PD-1/PD-L1 阻害剤が効かない患者も多数存在し，その原因の一つとしてがん細胞に対する TCR の認識能が不十分であると考えられる．そのため，がん細胞認識能を高めた TCR を T 細胞に導入する TCR 遺伝子改変 T 細胞輸注療法（TCR-T 療法）が注目されている．この治療法は，高親和性の TCR を患者から採取した T 細胞に導入し，再び患者に戻すことで，がん細胞を効率的に認識して攻撃することを目指すものである．しかし，患者ごとに最適な TCR を設計する必要があり，従来のような実験によるスクリーニングは時間とコストがかかるため，個別化医療には向かない．

この問題に対して，研究開始時点では，タンパク質の設計問題を逆問題として数理的にとらえ，がん抗原と TCR の結合能を予測する機械学習モデルにサンプリング手法を組み合わせ，TCR 空間において効率的に高い結合能を有する TCR の発見する手法を提案した．またがん抗原と TCR の結合能の予測モデルが複数存在したが，予測値と実測値の Pearson 相関係数が約 0.6 で，正確な予測ができていなかった．上記モデルを利用したサンプリングでは，偽陽性の TCR が数多く見つかり，後続の実験に支障をきたしていた．そのため，我々はタンパク質設計に利用できるタンパク質結合能予測モデルの条件を吟味し，高精度かつ高速な結合能予測モデルを開発することにした．

２．研究の目的

本研究の目的は，機械学習を利用して，高精度かつ高効率な T 細胞受容体分子（TCR）の設計手法を開発することである．近年，深層生成モデルの発展により，高精度のタンパク質生成モデルが開発されたが，ターゲット分子と高い結合能を有するタンパク質を発見するため，高精度な結合能予測モデルの開発が喫緊の課題となっている．本研究では，タンパク質設計に利用できるタンパク質結合能の予測モデルの開発を目指す．特に，以下の 3 つの性質を持つモデルが求められる．

１．高い予測精度．近年，機械学習分野においてタンパク質の設計問題が非常に注目されているが，設計されたタンパク質について十分に検証されなかったり，検証に利用される結合能予測モデルの精度が低いなどの問題が存在している．後続の実験のコストを考慮し，精度の高い結合能予測モデルを利用し，設計されたタンパク質から有望なものを選ぶ必要がある．

２．ハイスループット．機械学習モデルやサンプリング手法によって大量なタンパク質の立体構造が短時間で生成可能となっているため，結合能予測モデルもこれに対応して高速であることが望ましい．

３．汎用性：学習データに含まれていない新しく設計されたタンパク質に対しても予測が可能であること．

これらの性質をもつ結合能予測モデルを実現できれば，サンプリング手法や深層生成モデルと組み合わせ，高精度かつ高効率な TCR 設計が可能になると期待される．

３．研究の方法

がん抗原と TCR の結合能予測モデルのほとんどは TCR のアミノ酸配列から結合能を予測している．またグラフニューラルネットワークを代表とする深層学習モデルは，タンパク質の立体構造を入力として結合能を予測する．我々はマルチモーダル学習を利用し，タンパク質のアミノ酸配列と構造両方から有用な情報を統合し，高精度なタンパク質結合能予測モデルを開発した（図1）．具体的には，事前学習されたタンパク質の大規模言語モデルを利用し，アミノ酸配列からタンパク質の配列特徴量を抽出した．タンパク質の構造特徴量はトポロジカルデータ解析の手法を利用した．タンパク質の相互作用表面に存在する原子間の距離を Persistent Diagram という図の形で記述し，畳み込みニューラルネットワークで特徴量を抽出した．このように得られた配列特徴量と構造特徴量を統合し，勾配ブースティング法を用いて結合能予測モデルを構築した．勾配ブースティング法は，複数の弱学習器（決定木）を組み合わせて強学習器を生成する教師あり学習アルゴリズムである．タンパク質複合体の配列特徴量と構造特徴量を入力として受け取り，結合能を出力する．提案されたモデルはベンチマークデータセットでモデルを訓練し，評価を行った．

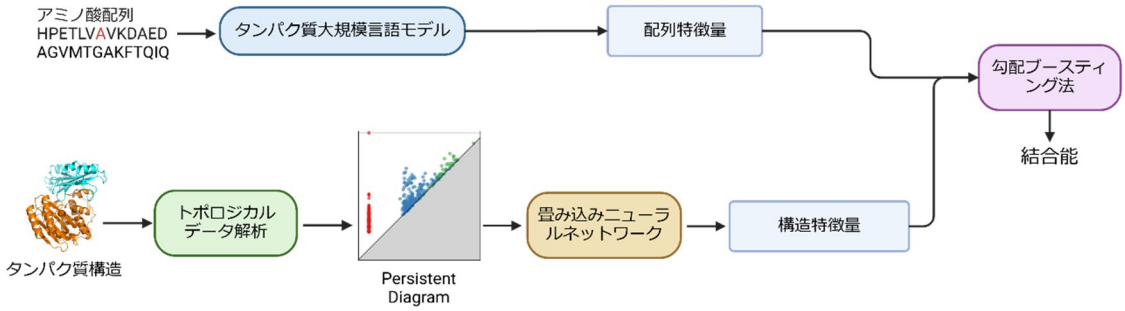

図 1. マルチモーダルなタンパク質結合能予測モデルのアーキテクチャ

４．研究成果

（１）タンパク質設計におけるタンパク質結合能予測モデルの課題

タンパク質の結合能予測は，タンパク質間相互作用の理解と機能解明に重要であるだけでなく，タンパク質デザインにも大きな役割を果たしている．設計されたタンパク質がターゲット分子と結合することを実験で確認するには時間と費用がかかるため，結合能予測モデルを利用することで，有望なタンパク質を選び出し，後続の実験に進むことができる．しかし，研究開始時点では，既存の予測モデルの精度が低く，TCR のデザインでは偽陽性や偽陰性が大きな問題となっていた．そこで，我々はタンパク質結合能予測モデルの構築に利用できるデータセット，既存のモデル，そしてタンパク質設計に必要な性質をまとめた（Guo and Yamaguchi, 2021）．特に，データセットはタンパク質の立体構造を含むものとアミノ酸配列のみのものがあり（10x Genomics, 2020; Jankauskaitė et al., 2019），それぞれに基づいて構造ベースと配列ベースの予測モデルが提案されている（Wang et al., 2020; Fischer et al., 2020）．しかし，各予測モデルは異なるデータセットで異なる方法で評価されており，モデルを公正に比較することが困難だった．私たちは，複数のデータセットで k-分割交差検証や類似性に基づくデータ分割での交差検証を提案した．また，利用可能なデータセットのサイズが小さく，高精度なモデルを構築しにくい問題に対しては，構造予測モデルを利用したデータ拡張の可能性を議論した．

（２）高精度かつ高速な結合能予測モデル

本研究では，既存の結合能予測モデルを検証し，予測値と実測値の Pearson 相関係数が約 0.6 であることを確認した．予測精度を向上させるため，タンパク質のアミノ酸配列と構造の両方の情報を統合して予測するマルチモーダルモデルを開発した（Guo et al., 2023）．二つのベンチマークデータセットで 10 分割交差検証と Leave-one-structure-out 交差検証を行い，その予測性能を検証した．特に Leave-one-structure-out 交差検証では，一つのタンパク質複合体の野生型と変異型の結合能データすべてをトレーニングデータから除外し，テストセットとして利用する．この方法は，新しいタンパク質の結合能データが利用できない状況を模倣しており，新たに設計されたタンパク質に対する予測モデルの精度を反映している．

抗体抗原複合体のデータセットでは，我々が提案したマルチモーダルモデルは Pearson 相関係数（Rp）0.904 を達成し，従来のモデル（Rp=0.684）を大幅に上回った（図 2）．また，Leave-one-structure-out 交差検証でも，提案手法は他のモデルよりも高い精度を維持したが，Rp 値は 0.832 とやや低くなった．これは，機械学習モデルの一般的な課題であり，外挿領域において予測精度が低下することが知られている．提案手法は外挿領域でも高い精度を維持しているため，タンパク質設計に利用できると考えられる．さらに，複数の種類のタンパク質複合体を含むデータセットでは，提案手法は Rp=0.849 を達成したが，抗体抗原複合体のデータセットよりパフォーマンスがやや低下した．これはデータセットに多様なタンパク質ファミリーや相互作用タイプが存在しているためであり，トレーニングデータが限られている場合，多様なデータに対して高精度な一般化モデルが構築することが難しいことを示唆している．

さらに，アブレーション研究を行い，構造情報と配列情報のそれぞれの有用性を確認した．具体的には，配列情報のみ，構造情報のみ，そして両方の情報を組み合わせた場合のモデルの予測性能を比較し，その結果，

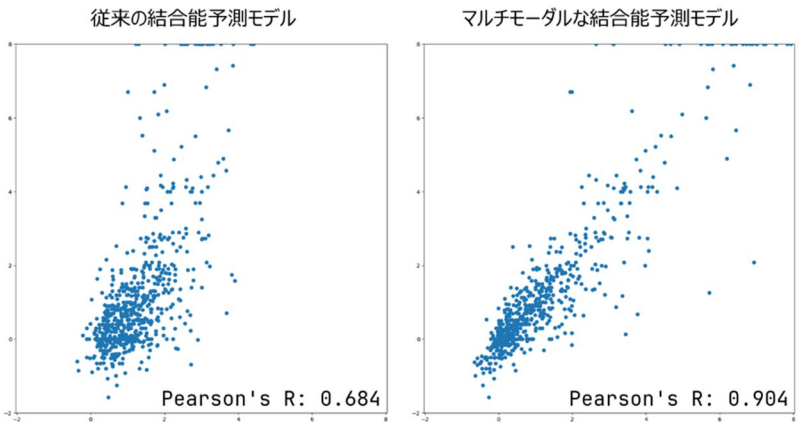

図 2. タンパク質結合能予測における従来モデルとマルチモーダルな予測モデルの予測精度．

ほとんどの検証において両方の特徴を使用することで予測性能が向上した．また，異なるデータに対しても予測精度がロバストになることも確認した．提案手法は高精度であり，一つのタンパク質複合体の結合能を予測する速度は約 10 秒で，スーパーコンピューターを用いた並列実行により，タンパク質のバーチャルスクリーニングにも利用できる．今後，本手法を活用してタンパク質の設計を進める予定である．

＜引用文献＞

① Guo, Z.; Yamaguchi, R. Machine Learning Methods for Protein-Protein Binding Affinity Prediction in Protein Design. Front. Bioinform. 2022, 2, 1065703. https://doi.org/10.3389/fbinf.2022.1065703.

② 10x Genomics. A new way of exploring immunity: Linking highly multiplexed antigen recognition to immune repertoire and phenotype. 2020. Available at: https://www.10xgenomics.com/resources/document-library/a14cde

③ Jankauskaitė, J., Jiménez-García, B., Dapkūnas, J., Fernández-Recio, J., and Moal, I. H. Skempi 2.0: An updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. Bioinformatics 2019, 35, 462-469. doi:10.1093/bioinformatics/bty635

④ Wang, R., Fang, X., Lu, Y., and Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J. Med. Chem. 2004, 47, 2977-2980. doi:10.1021/jm030580l

⑤ Fischer, D. S., Wu, Y., Schubert, B., and Theis, F. J. Predicting antigen specificity of single T cells based on TCR CDR3 regions. Mol. Syst. Biol. 2020, 16, e9416. doi:10.15252/msb.20199416

⑥ Guo, Z.; Muto, O.; Fukushima, Y.; Demachi-Okamura, A.; Ota, M.; Yoshida, R.; Matsushita, H.; Yamaguchi, R. A Multimodal Framework Combining Sequence and Topological Features for Accurate Protein-protein Binding Affinity Prediction. 32nd International Conference on Genome Informatics (GIW/ISCB-Asia), 2023.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Zhongliang Guo and Rui Yamaguchi	４．巻 2
２．論文標題 Machine learning methods for protein-protein binding affinity prediction in protein design	５．発行年 2022年
３．雑誌名 Frontiers in Bioinformatics	６．最初と最後の頁 1065703
掲載論文のDOI（デジタルオブジェクト識別子） 10.3389/fbinf.2022.1065703	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 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Zhongliang Guo, Osamu Muto, Yasunori Fukushima, Ayako Demachi-Okamura, Motonori Ota, Ryo Yoshida, Hirokazu Matsushita, Rui Yamaguchi
２．発表標題 A multimodal framework combining sequence and topological features for accurate protein-protein binding affinity prediction
３．学会等名 GIW ISCB ASIA 2023（国際学会）
４．発表年 2023年

１．発表者名 Zhongliang Guo, Osamu Muto, Rui Yamaguchi
２．発表標題 An integrated approach using sequential and structural features for precise prediction of protein-protein binding affinity
３．学会等名 IUPAB Congress 2024（国際学会）
４．発表年 2024年

１．発表者名 郭 中梁、武藤 理、山口 類
２．発表標題 A Sequence and topological feature integration for accurate protein-protein binding affinity estimation
３．学会等名 第6回日本メディカルAI学会学術集会
４．発表年 2024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
研究協力者	山口　類 (Yamaguchi Rui)		
研究協力者	武藤　理 (Muto Osamu)		

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