

令和7年度 東京情報大学総合研究所プロジェクト研究
研究実績報告書

1. 研究課題名

モデル系動植物に由来するオミクスデータを用いた分子機能解析及び分子間相互作用
予測システムの基盤技術の開発

2. 研究組織

区分	氏名	所属・職名
研究代表者	村上 洋一	総合情報学部 総合情報学科・教授
研究分担者	田中 啓介	総合情報学部 総合情報学科・准教授
	李 嘉一	総合情報学研究科・大学院生
	松井 千里	総合情報学部 総合情報学科・学部生
	栗野 浩大	総合情報学部 総合情報学科・学部生

3. 研究期間

2025 年 4 月 1 日～2026 年 3 月 31 日

4. 研究の目的

本研究の目的は、オミクスデータ（タンパク質配列・構造等）を活用し、分子機能解析の基盤となる高精度なタンパク質間相互作用（PPI）部位予測システムを開発することである。具体的には、最新の言語モデル ESM-C と ProstT5 を利用し、特に予測が困難とされるヘテロタンパク質間の PPI 部位を、配列情報から特定する高精度な基盤技術の確立を目指す。

5. 研究報告

【研究の概要】

生体内における分子間相互作用の解明は、タンパク質の分子機能の理解やオミクスデータに基づく生命現象のシステムの把握に不可欠である。特に、異なる種類のタンパク質が結合するヘテロタンパク質間相互作用は、シグナル伝達や代謝制御等の根幹をなすが、その部位予測は高い多様性と非対称性ゆえに極めて難しい。本研究は、これらを高精度に特定する分子間相互作用予測システムの基盤技術を提案するものである。近年の大規模言語モデル（PLM）の発展と AlphaFold2 による構造予測の普及を背景に、実験構造データが不足している未知のタンパク質群に対しても適用可能なアプローチを構築した。

【提案する基盤技術の4つの重要工程】

本基盤技術は、次の4つの重要な工程、データセット作成、特徴抽出、機械学習、構造フィルタリングから成る。

- ① **データセット作成**: オミクスデータの信頼性を担保するため、PiSite および PDB からヒト由来ヘテロタンパク質複合体データを収集し、配列類似性 25%以下、E-value $1e-6$ 、残基欠損率 30%未満といった厳格な基準で精製した高品質な学習データセット (8,742 本) を構築した。
- ② **特徴抽出**: タンパク質配列を高度に表現するため、進化的・構造的潜在情報を捉える ESM-C (1152 次元) と、Foldseek 由来の構造トークン情報を学習した ProstT5 (1024 次元) を併用した。
- ③ **予測モデルの構築**: 連結した 2176 次元のベクトルを特徴量とし、勾配ブースティング決定木の一つである LightGBM を用いて残基単位での PPI 部位判定を行う。クラス不均衡に対応するため、Positive 残基に重みを付加する処理を適用している。
- ④ **構造的フィルタリング**: AlphaFold2 で予測した立体構造の $C\alpha$ 原子座標を利用し、予測された残基が空間的に孤立している場合に偽陽性と判断して排除する仕組みを導入した。

【予測性能の優位性とモデルの相補性】

ベンチマークデータセット (Dset448 および Dset55) を用いた評価の結果、提案する基盤技術は既存の予測手法よりも高い性能を示した。また、ESM-C と ProstT5 を連結したモデルは、単独モデルよりも全ての指標で高い性能を発揮した。特に、ヘテロタンパク質限定の Dset55 において、MCC 0.519、F1-Score 0.614、Precision 0.674 という高い予測精度を達成した。

先行研究「Seq-InSite」をローカル環境に実装し、同一のヘテロタンパク質データで比較した際も、本技術が上回る結果となった。また、多種多様なタンパク質が混在する Dset448 においても、ROC-AUC 0.86 を記録し、先行研究と同等以上の汎用性を確認した。

解析の結果、この高い性能は二つの言語モデルの相補性に起因することが判明した。ESM-C は親疎水性といった物理化学的特性を強く捉える一方、ProstT5 は α ヘリックスや β シートといった二次構造の識別能力に優れていることが示唆された。すなわち、これら異なる生物学的側面を統合することで、モデルは単一の特徴に依存せず、複雑な相互作用様式を多角的に学習していると考えられる。

【AlphaFold2 の予測構造情報の活用によるノイズの抑制】

PPI (タンパク質間相互作用) 部位が空間的に隣接して存在する性質に基づき、予測結果に対して空間的フィルタリングを適用した。具体的には、陽性と予測された残基の半径 12.3Å 以内に他の陽性残基が存在しない場合、それを孤立した偽陽性とみなして陰性に修

正する処理を行った。この結果、MCC は 0.51 から 0.522 へ、Precision は 0.64 から 0.66 へと向上した。これは、実験構造が未解明なタンパク質であっても、AlphaFold2 による予測構造情報を事後処理として統合することで、配列情報のみでは排除しきれなかった計算上のノイズを効果的に抑制できることを実証している。

【結論と分子機能解析への貢献】

本研究は、大規模オミクスデータから高精度に分子間相互作用を特定する予測システムの基盤技術を確立した。特に、難易度の高いヘテロタンパク質の PPI 部位において、世界トップレベルの予測精度を達成した意義は大きい。本研究の主要な成果は以下の 3 点に集約される。

1. 二つの最新 PLM (ESM-C, ProstT5) の統合が、ヘテロタンパク質の複雑な特徴表現に有効である。
2. 構築されたデータセットと学習モデルは、従来のシステムよりも高い汎用性と精度を両立している。
3. 予測立体構造を用いた空間的フィルタリングは、予測の精度向上と物理的妥当性の確保に寄与する。

今後の展望として、本システムを Web サーバとして実装・公開し、世界中の研究者がモデル系動植物の分子機能解析に活用できる環境を整備する。さらに、天然変性タンパク質 (IDP) への応用や、アンサンブル学習の導入による更なる高性能化を通じ、未知の生命情報の解明や効率的な創薬ターゲット選定に資するシステムとしての完成度を高めていきたい。

6. 成果の公表

- [1] Ri, K., Tanaka, K., Murakami, Y., (2024) "Structural comparison of protein-protein interaction sites on the AlphaFold2 model structures and the natural structures", The 47th Annual Meeting of the Molecular Biology Society of Japan.
- [2] 李嘉一、田中啓介、村上洋一、(2025) "AlphaFold2 におけるタンパク質間相互作用部位周辺の構造予測精度の検証"、第 81 回バイオ情報学研究会