

令和6年度 東京情報大学総合情報研究所プロジェクト研究
研究実績報告書

1. 研究課題名

オミクスデータを統合的に活用し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	大学院総合情報学研究科・大学院生
	松井 千里	総合情報学部 総合情報学科・学部生

3. 研究期間

2024年4月1日～2025年3月31日

4. 研究の目的

生体内のさまざまなタンパク質が関わる分子間相互作用の全容を明らかにすることで、重要な生体機能や疾患メカニズム等の解明につながることが期待される。そこで本研究ではタンパク質間相互作用に焦点を当てて、オミクスデータを統合的に活用した相互作用予測システムの基盤技術の開発を目指し、そのための要素技術として期待されるタンパク質の立体構造予測の深層学習モデルである AlphaFold2 の精度について検証を行う。

5. 研究報告

AlphaFold2 (Jumper, J. et al., Nature, 2021, vol.596, pp.583-589., 以下、AF2) は、タンパク質の構造予測において大きな進歩をもたらし、分子機能解明や創薬研究等において重要なツールになっている。また、さまざまなタンパク質が関わる分子間相互作用やその相互作用部位の予測において、これまで構造情報が未知の場合、配列情報のみを用いる方法が基本であったが、AF2 によって予測された構造を利活用することで、その予測精度の向上が期待されている。

しかしながら、未知のタンパク質に対する AF2 の予測精度、特にタンパク質間相互作用 (以下、PPI) の部位及びその周辺構造の予測精度について十分な検証がされていないため、PPI 部位等の予測への応用可能性に対して不明確なところがある。そこで本研究では、AF2 の予測構造の信頼性を検証するために、PPI 部位及びその周辺構造の AF2 の予測精度について検証を行った。具体的には AF2 の学習データに含まれていないタンパク質を用いて、

AF2 の予測構造上の PPI 部位及びその周辺構造と、天然構造上のそれらの構造を、Contact Map 及び RMSD を用いて比較を行った。

■ 解析方法

(1) 構造比較のためのデータセット

AF2 をローカル環境で実行するためのツールである LocalColabFold (Mirdita, M. et al., Nature Method, 2022, vol.19, pp.679-682., 以下、LCF) の最終更新日以降 (2024/5/24) に、タンパク質の構造データベースに登録されたヒトのタンパク質構造データのみ (天然構造) を比較対象とした。またその各々の配列データを取得し、LCF を用いて予測構造を計算した。最終的に比較される構造データのペア数 (天然構造とその予測構造) は 187,151 ペアであった。

(2) Contact Map に基づく構造類似度

Contact Map (以下、CM) は、天然構造とその予測構造の全残基間の $C\alpha$ 原子のユークリッド距離の情報を 2 次元で表現したマップである。両構造間の PPI 部位とその周辺構造の CM に基づく構造類似度は、PPI 部位上の各残基を中心とした前後 4 残基内の距離の差分平方和の平均値 ($SSDAvg_{Local}$) とした。一方、CM による全体の構造類似度は、全残基の距離の差分平方和の平均値 ($SSDAvg_{Global}$) とした。

(3) RMSD に基づく構造類似度

RMSD は、2 つの構造の原子間の距離の平均二乗偏差の平方根から得られる構造間の差異を定量的に表す指標である。この値が小さいほど、構造が類似していると評価される。RMSD に基づく構造類似度は、天然構造とその予測構造の PPI 部位のみを重ね合わせて計算された RMSD 値 ($RMSD_{Local}$) と全体を重ね合わせた RMSD 値 ($RMSD_{Global}$) とした。

■ 解析結果

(1) Contact Map による構造比較

準備したデータセットにおいて、天然構造とその予測構造のペア数全体の約 89.2% (165,060/187,151 ペア) が、 $SSDAvg_{Local} \leq 2\text{\AA}$ 、 $SSDAvg_{Global} \leq 2\text{\AA}$ の範囲に含まれていた (図 1 の下の赤枠部分)。これは予測構造全体と PPI 部位及びその周辺構造の大部分が、実験由来構造の類似することを示す。すなわち、多くの予測構造の PPI 部位及びその周辺構造また全体構造が、天然構造に近い精度で予測されることが示唆された。しかしながら、図 1 で示したユビキチン (8F1F, chain A) の例のように、予測構造全体は天然構造に近くても、PPI 部位の構造が一部正しく予測されていない例があった。

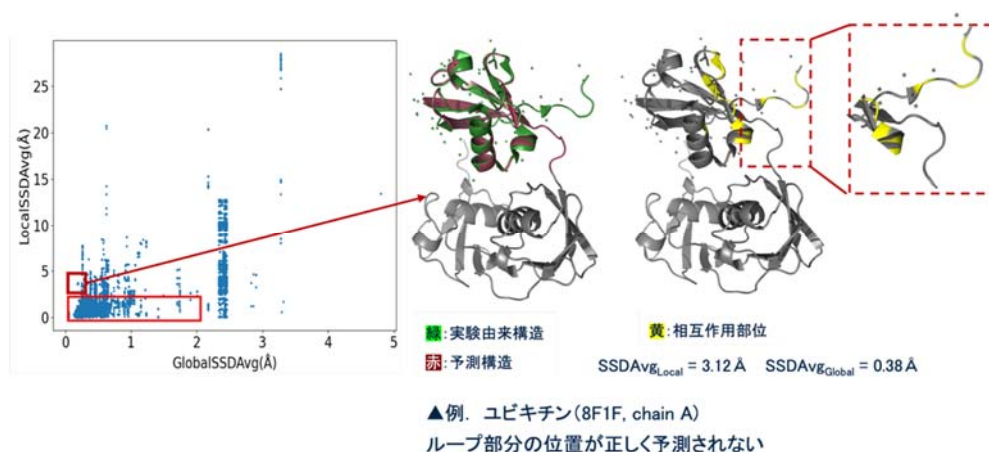


図 1. Contact Map による構造比較結果(SSDAvg_{Local} vs SSDAvg_{Global})

(2) RMSD による構造比較

天然構造とその予測構造のペア数全体の約 68.1% (106,393/156,515 ペア) が、 $RMSD_{Local} \leq 2 \text{ \AA}$, $RMSD_{Global} \leq 2 \text{ \AA}$ の範囲に含まれていた (図 2 の小さい赤枠部分)。また、ペア数全体の約 98.4% (154,005/156,515 ペア) が、 $RMSD_{Local} \leq 3 \text{ \AA}$, $RMSD_{Global} \leq 3 \text{ \AA}$ の範囲に含まれていた (図 2 の大きい赤枠部分)。CM と同様、全体構造と PPI 部位の構造が正しく予測される傾向が示された。しかしながら、複数のタンパク質で構成された複合体の一部として働き、周囲の環境に応じて変化するタンパク質の場合、構造予測が難しいことが示唆された ($RMSD_{Global} > 5.0 \text{ \AA}$)、例. ヒストン脱アセチル化酵素 (7XUZ, chain A)。

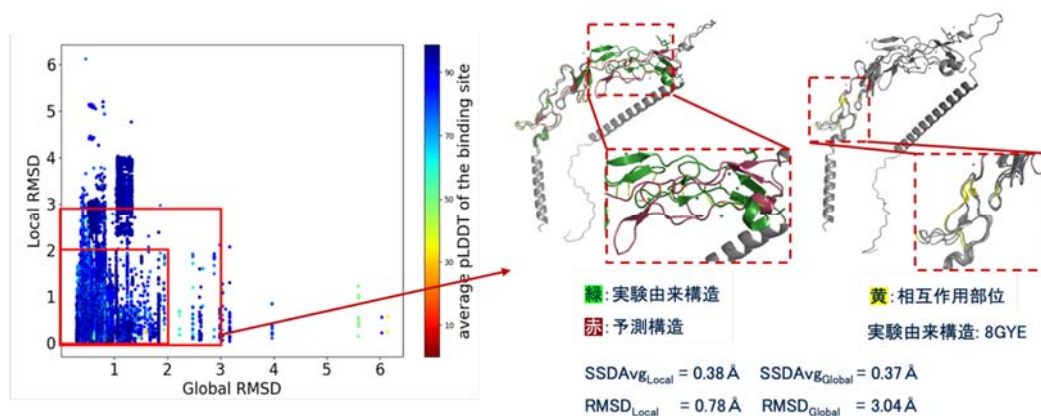


図 2. RMSD による構造比較結果($RMSD_{Local}$ vs $RMSD_{Global}$)

■ 考察と結論

本研究では、AF2 によるタンパク質の予測構造の PPI 部位及びその周辺構造の予測精度を検証するために、CM 及び RMSD を用いた解析を実施した。その結果、CM を用いた解析では、多くのタンパク質では PPI 部位及びその周辺構造は、全体構造と同様に高い精度

で予測されることが確認された。一方、PPI 部位が末端に位置するようなタンパク質では、構造予測が事例も確認できた。RMSD を用いた解析においても同様に、PPI 部位及びその周辺構造と全体構造の予測精度が高いと確認できたが、複数のタンパク質で構成された複合体の一部として働き、周囲の環境に応じて変化するような構造の場合、構造予測が難しい事例が確認できた。以上のことから、一部のタンパク質において構造予測が困難な事例もあったが、AF2 は全体構造と PPI 部位及びその周辺構造を高い精度で予測できる可能性が確認できた。今後、AF2 及び AlphaFold3 のモデル構造を利活用することで、PPI 部位予測等の精度向上が図られることが期待される。

6. 成果の公表

- Kaichi Ri, Keisuke Tanaka, Yoichi Murakami, "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, 2024. (ポスター発表)
- 李嘉一, 田中啓介, 村上洋一, "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, 2025/3/7. (口頭発表)