7ZG-02

構造ベーススクリーニングに適したAlphaFold予測構造生成手法の提案

内河 慶輔　　古井 海里　　大上 雅史
東京工業大学 情報理工学院 情報工学系

1　序論

計算機を用いて薬剤候補化合物を選抜するバーチャルスクリーニングは，創薬のコスト削減のアプローチとして注目されている．タンパク質の立体構造情報を用いる構造ベーススクリーニングは，標的タンパク質に関する既知の実験情報を用いないことから，新規性の高い薬剤候補化合物を発見しやすいとされている．しかし，標的タンパク質の構造によってスクリーニングの精度が大きく変化し，一般に薬剤が結合していない状態の構造 (apo 体) よりも薬剤が結合した構造 (holo 体) を用いると精度が良いことがわかっている [1]．本研究では，タンパク質立体構造予測手法である AlphaFold2 [2] を用いて様々な予測構造を出力し，予測構造のスクリーニング精度を検証することを目的とした．そして，スクリーニングに適した予測構造を出力し選択できる方法について検討と評価を行った．

2　手法

2.1　実験の概要

AlphaFold2 による予測構造を様々なパラメータで大量に生成し，生成した予測構造のクラスタリングによって代表構造を選定し，AutoDock Vina によるリガンドドッキングを行った．その上で，(1) 予測構造の生成パラメータ，(2) 予測構造のクラスタリング方法，(3) 代表構造の選択方法，を変化させ，どのパターンの組合せによる予測構造がスクリーニングに適していたかを評価した．

2.2　データセット

スクリーニング性能の検証には，構造ベーススクリーニングのベンチマークデータセットである DUD-E [4] から 3 標的 (CDK2，KIF11, CXCR4) を扱った．それぞれの標的毎に active/inactive 化合物が複数紐づいており，これらの 2 値分類性能を測ることでスクリーニング性能を評価した．各標的のアミノ酸配列情報を UniProt [6] から取得して AlphaFold2 の入力とした．

2.3　AlphaFold2 予測構造の生成

UniProt から得られたアミノ酸配列を AlphaFold2 (LocalColabFold [3]) に入力し，予測構造を生成した．予測構造を変化させるために，多重配列アラインメント (MSA) の本数を制御するパラメータを 7 通り，seed 値を 10 通り変化させて予測を実行し，各実行において 5 個の予測構造を得た．このパラメータを基本とし，AlphaFold2 の生成パターンとして以下の 4 パターンを検討した．

1) vanilla　MSA サンプリングと seed 値変更のみを行う基本の設定 (350 個の構造を生成)
2) templates_apo　apo 体の実験構造をテンプレート構造として与える (140 個*の構造を生成)
3) templates_holo　holo 体の実験構造をテンプレート構造として与える (140 個*の構造を生成)
4) recycle_1　AlphaFold2 の recycling 処理の回数を 1 回に減らす（通常は 3 回）(350 個の構造を生成)

* AlphaFold2 はテンプレート構造を与えた場合，5 個の予測構造のうち 2 個はテンプレート構造を参照し，3 個はテンプレート構造を参照しない．そのため，前者の 2 個を用いている．

2.4　予測構造のクラスタリング

各生成パターンでの予測構造間の重原子 RMSD を計算し，その RMSD を距離行列として，予測構造を生成パターン毎に Ward 法で 10 クラスタにクラスタリングした．検証する RMSD の計算パターンとして，以下の 2 つを検討した．

1) all_res　全残基を対象とした重原子 RMSD
2) centroid　ドッキングの中心座標から 5 Å 以内に位置する残基の重原子 RMSD

Explore AlphaFold2 model structure generation method for structure-based virtual screening
Keisuke Uchikawa, Kairi Furui & Masahito Ohue, Department of Computer Science, School of Computing, Tokyo Institute of Technology

2.5 代表構造の選定

クラスタリング結果より，各クラスタから代表構造を 1 つずつ選定する．検証する代表構造の選定パターンとして，以下の 2 つを検討した．

1) rmsd_max　各クラスタ内での RMSD の合計値が一番大きい構造を代表構造として選定する
2) rmsd_min　各クラスタ内での RMSD の合計値が一番小さい構造を代表構造として選定する

2.6 リガンドドッキングによるスクリーニング

AutoDock Vina [5] は代表的なリガンドドッキングソフトウェアの 1 つである．選定した代表構造と DUD-E から得たラベル付き化合物間のドッキングを AutoDock Vina によって行い，得られたドッキングスコアをもとに化合物を選択（順位付け）したときの結果に対して，もとの active/inactive のラベルから ROC-AUC 値を計算する．ROC-AUC 値が高いほど，既知の化合物の選別を正しく行えた予測構造であったと言える．AutoDock Vina は version 1.2 を用い，MGLtools による前処理を行った．ドッキング部位の中心座標は各標的の複合体構造をもとに指定し，グリッドサイズを 20 Å とした．

3 結果と考察

表 1, 2, 3 に，3 標的の各検討パターンで選定された代表構造 10 個それぞれでスクリーニングを行って得られた ROC-AUC 値の最大値を示す．

3 標的全体を通して，holo 体のテンプレート構造を用いて生成した構造 (templates_holo) が，総合的にスクリーニングの性能が良かった．クラスタリングや代表構造の選び方については，標的ごとに良い結果となった組合せが異なり，普遍的に最適なパターンは得られなかった．このことから，今回検討したクラスタリングや代表構造の選定方法はスクリーニング性能にあまり寄与せず，予測構造の生成手法が最もスクリーニング性能に影響を与えやすいと考えられる．テンプレートとなる holo 体が存在すればテンプレートとして指定し，holo 体が無い場合は標準的な生成方法 (vanilla) を用いると良いと示唆される．

4 結論

本研究では，AlphaFold2 を用いたバーチャルスクリーニングの手法を確立するため，予測構造の生成手法や予測構造の選定手法の様々な組合せでのスクリーニング精度を検証し，holo 体テンプレートの利用が有用であることを示した．実用上は，ROC-AUC 値が高い予測構造を既知化合物のドッキング無しに選び出すことができると尚良く，今後の課題である．

表 1　CDK2 の予測構造での ROC-AUC の最大値

	all_res rmsd_max	all_res rmsd_min	centroid rmsd_max	centroid rmsd_min
vanilla	0.684	0.674	**0.690**	0.684
templates_apo	0.667	0.667	0.670	0.670
templates_holo	0.681	0.676	0.686	0.674
recycle_1	0.681	0.682	0.682	0.672

表 2　KIF11 の予測構造での ROC-AUC の最大値

	all_res rmsd_max	all_res rmsd_min	centroid rmsd_max	centroid rmsd_min
vanilla	0.718	0.708	0.700	0.705
templates_apo	0.720	0.734	0.734	0.751
templates_holo	0.740	0.743	**0.760**	0.742
recycle_1	0.740	0.735	0.750	0.753

表 3　CXCR4 の予測構造での ROC-AUC の最大値

	all_res rmsd_max	all_res rmsd_min	centroid rmsd_max	centroid rmsd_min
vanilla	**0.771**	0.715	0.747	0.730
templates_apo	0.718	0.710	0.722	0.709
templates_holo	0.740	0.761	0.754	0.747
recycle_1	0.736	0.714	0.701	0.710

謝辞 本研究は，JST 創発的研究支援事業 (JPMJFR216J)，JST ACT-X (JPMJAX20A3)，JSPS 科研費 基盤研究 (B) (No. 20H04280) の支援を受けて行われた．

参考文献

[1] 柳澤渓甫. タンパク質立体構造情報を用いた薬剤バーチャルスクリーニング. *JSBi Bioinform Rev*, 2: 76–86, 2021.
[2] Jumper J, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*, 596: 583–589, 2021.
[3] Mirdita M, *et al.* ColabFold: making protein folding accessible to all. *Nat Methods*, 19: 679-682, 2022.
[4] Mysinger MM, *et al.* Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem*, 55: 6582–6594, 2012.
[5] Eberhardt J, *et al.* AutoDock Vina 1.2.0: New docking methods, expanded force field, and python bindings. *J Chem Inf Model*, 61: 3891–3898, 2021.
[6] UniProt Consortium. UniProt: a worldwide hub of protein knowledge, *Nucl Acids Res*, 47: 506–515, 2018.