インシリコ解析ユニット

補助事業代表者（所属・職名）
富井　健太郎（産総研 人工知能研究センター・上級主任研究員）

ライフサイエンス研究加速のためのバイオインフォマティクス研究

補助事業分担者（所属・職名）
池田　和由（理化学研究所計算科学研究センター・ユニットリーダー）、深沢　嘉紀（宇都宮大学バイオサイエンス教育研究センター・准教授）

高度化における研究背景・目的

多様化するニーズへの対応と支援体制の強化のため、インシリコ解析技術の更なる開発を実施。データベースでは、タンパク質と基質の相互作用について、結合部位推定のためのデータベースPoSSuMやタンパク質間相互作用(PPI)阻害化合物データベースDLiPなどの継続的開発を実施。また、AI駆動型反復スクリーニング法の開発や、Boltz-2を用いたAI仮想スクリーニングの高度化研究を実施。さらに、タンパク質表面の点群特徴（表面原子を点の集まりとして表した形の特徴）とタンパク質配列の言語表現（多数の配列からタンパク質の「らしさ」を学ぶAI表現）を統合したESpmaを開発。

高度化の成果

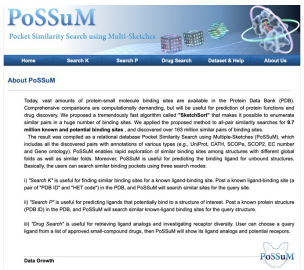
多様化するモダリティに対応するためのデータベースの拡充

■低分子化合物などリガンドとの相互作用部位と類似する局所構造を列挙したデータベースPoSSuMの開発を継続
•リガンド側、レセプター側の双方から検索が可能。(PoSSuM)
•承認済みの経口低分子化合物の結合部位と類似した部位、また類似部位に結合する分子を探索した結果を公開。(PoSSuMds)

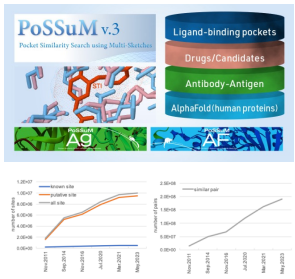
• 既知のリガンドの相互作用部位との類似性に基づく新規結合部位の検索。
• 研究対象のタンパク質の構造/構造モデルに対して、どのようなリガンドが結合する可能性があるか予測。
• 承認済みの経口低分子化合物と類似の化合物、またレセプター側の多様性をPDBスケールで解析した結果を公開しており、化合物の結合様式の比較やファーマコフォアの推定などに利用可能。

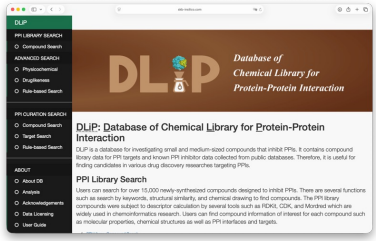
DLiPでは、大環状合成中分子ライブラリーを追加
これを活用し特定のPPI標的に対して、深層学習と組み合わせた反復スクリーニング(iterative screening)を実施

■タンパク質の立体構造データベースProtein Data Bank (PDB)の成長に伴い、現在のPoSSuMは、およそ1,000万の結合(候補)部位の網羅的な類似性比較に基づく、億単位の類似部位ペアを収載するデータベースとなっている。
■PoSSuMは、そのサブセットとして、承認済の経口薬やプロドラッグなどの結合部位類似性データセットPoSSuMds (PoSSuM drug search)を備えていル。
■さらに、抗体と抗原の相互作用類似性検索のためのPoSSuMAg
■AlphaFoldによるヒトタンパク質予測構造モデルの基質結合候補部位と既知基質結合部位の類似性検索用DB PoSSuMAF(https://possum.cbrc.pj.aist.go.jp/PoSSuMAF/)も利用可能。



https://possum.cbrc.pj.aist.go.jp/PoSSuM/



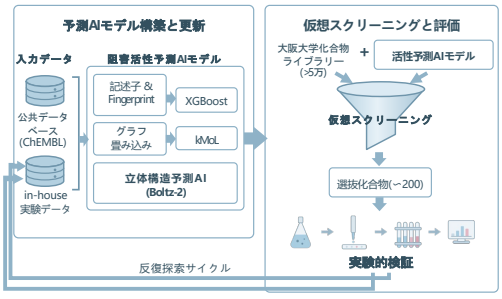


https://skb-insilico.com/dlip

J Chem Inf Model. 2023 Nov 28;63(23):7578–7587.

AI駆動型反復スクリーニング法の確立

■少数の化合物を用いた実験と、得られた活性データに基づくAI予測を反復することで、有望化合物を効率的に探索する反復スクリーニング法を開発した。これまでに、中分子合成化合物ライブラリーとリガンドベース機械学習を組み合わせ、Keap1/Nrf2をはじめとするPPI標的に対して、初回スクリーニングを上回るヒット率で新規活性化合物を取得し、ヒット化学空間の拡大に成功した（Shimizu *et al.*, *Chem. Commun.*2023）。今後は、実験データの蓄積と予測モデルの更新を繰り返すことで、本手法の汎用性を検証するとともに、立体構造情報を活用したAI予測や化合物生成技術を組み合わせ、さらなる探索効率の向上を目指す。



タンパク質複合体における特徴量としての分子表面解析と定式化

ESpma

タンパク質間相互作用解析の高度化に向け、タンパク質表面の三次元点群特徴と、タンパク質言語表現モデル（pLM）による配列由来のAI表現を統合したマルチモーダル解析手法（ESpma）を開発した。生物学的な相互作用と結晶化に伴う偶発的な接触を識別する課題を学習対象として、相互作用界面に局在する配列・構造特徴を活用することで、タンパク質複合体構造のスクリーニングを効率化する解析基盤を構築した。

本高度化では、複数のデータセットおよびAIモデルを用いた検証を通じて、相互作用界面に着目した特徴抽出の有効性を確認するとともに、点群による立体構造情報を組み合わせた統合解析や、予測に寄与する残基の可視化・解釈を可能とするモデル(図1)を開発した。
データサイズの制約を考慮し、残基ごとの情報を直接学習するのではなく、相互作用界面全体に情報を縮約した表現を用いて学習を行った。その結果、情報を縮約した簡便なモデルでありながら性能低下は認められず、複数のデータセット・AIモデルにおいて安定した分類性能を示した（表1,2, 図1）。
一方、pLMは高い予測性能を示す反面、その判断根拠の解釈が難しい。そこで、モデルの線形性を利用して予測スコアを残基単位に分解したところ、残基の埋没度に関連する特徴を暗黙的に捉えていることが示された（図2）。この結果から、pLMの内部表現には、相互作用界面を特徴づける物理的制約が内包されている可能性が示唆された。

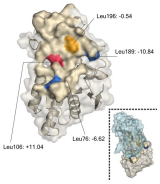


図1: 相互作用面におけるロインシと予測への寄与度

Method	ACC	MCC	AUROC
ESpma (提案手法1)	0.846	0.698	0.912
ESpma$_{bc}$ (提案手法2)	0.840	0.679	0.906
PRODIGY-crystal	0.735	0.508	0.843
EPPIC	0.827	0.664	0.899
DeepRank-GNN	0.820	0.640	0.865
DeepRank-GNN-esm	0.833	0.673	0.889

表1: ESpmaと既存の競合手法との性能比較

Method	ACC	MCC	AUROC
ESpma (提案手法1)	0.802	0.628	0.938
ESpma$_{bc}$ (提案手法2)	0.865	0.742	**0.957**
EPPIC	**0.921**	**0.845**	0.955
DeepRank-GNN-esm	0.802	0.659	0.802

表2: 別データセットにおける競合手法との性能比較

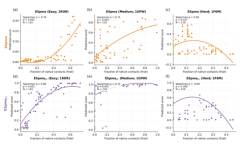


図2:異なる難易度のドッキングデコイに対するESpmaの性能

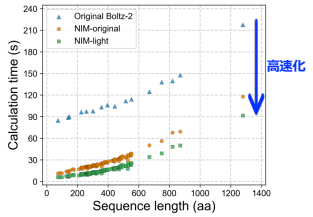
本成果は論文として受理されており [1]、タンパク質間相互作用解析の効率化に資するAI・構造情報統合技術として公表予定である。一般向けとして、以下にコードやGoogle Colabによる計算環境を提供している[2]。

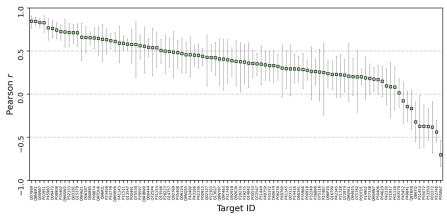
[1] Nozawa S, Tomii K, Fukasawa Y, Protein Science, *in press*.
[2] https://github.com/fukasawa-group/espma

界面レベルの縮約による性能向上と手法の限界。界面の定義に依存するため、界面構造の誤りに対しても脆弱である。

Boltz-2を用いたAI仮想スクリーニングの高度化 (Shimizu and Akabane *et al. CBI j.* 2026, 26, 11-27)

■NVIDIA NIMの利用と適切な設定によって論文実装と同等の予測で計算時間の60–90%削減に成功した。
■Boltz-2の結合親和性予測における未学習データを収集 (標的数: 91、化合物-標的のペア数: 8,350)、大規模ベンチマークを実施したところ、様々な標的クラスに対する有効性が示された一方で、化合物の新規性・物理化学的特性と予測誤差の大小に相関は確認されず、Boltz-2の広範な適用範囲が示唆された。





今後の発展

創薬スクリーニング法の高度化に向け、独自データベースの開発に加え、既存手法を活用したより強力かつ効率的なスクリーニング法の開発を継続している。今後は、これらの手法やデータベースの統合と相互活用をさらに進め、多様な課題に対応可能な汎用性と高速性を両立するインシリコ探索フローの構築を推進する。

本発表について開示すべきＣＯＩはありません