7ZB-04

Random forestを用いたドッキング構造の学習による バーチャルスクリーニングのポスト処理

安尾　信明[†]　　関嶋　政和[†,‡]

東京工業大学 情報理工学院 情報工学系[†]
東京工業大学 科学技術創成研究院 スマート創薬研究ユニット[‡]

1. 序論

近年、創薬におけるコストダウンを目指し，情報技術を用いた創薬支援研究が盛んに行われている．情報技術を用いた創薬支援手法の一つにバーチャルスクリーニングが存在する．これは，薬剤標的となるタンパク質に対し効果のある化合物を計算機上で探索する問題である．

バーチャルスクリーニングにおける手法としては，タンパク質の立体構造を用いる SBDD (Structure-based Drug Discovery) と，既存の活性情報を用いる LBDD (Ligand-based Drug Discovery) がある．SBDD は LBDD に比べ新規の構造をもつ化合物が得やすいと言われる一方，ヒット率は LBDD に劣ると言われ [1]，その精度の改善が求められている．

SBDD でよく用いられる手法にタンパク質-リガンド間ドッキングが存在する．様々なドッキングのアプリケーションがこれまでに開発されているが，それらの精度は未だ十分ではない [2]．この問題を解決するため，既知の活性情報を SBDD に適用してリランキングを行い，精度を向上させる試みが存在する．例えば SIFt[3] や Pharm-IF[4] は，タンパク質-化合物間相互作用の有無や距離を符号化した Interaction Fingerprint を作成し，化合物を評価する．しかし，これらの手法は相互作用の有無や距離といった情報は用いているが，相互作用の強さの情報が失われているという問題がある．

本研究では，タンパク質-リガンド間ドッキングによるバーチャルスクリーニングの精度向上を目的として，ドッキングから得られる相互作用エネルギーの情報を用い，これを機械学習に入力し評価する SIEVE-Score を提案した．また，ドッキングのベンチマークセットを用いてバーチャルスクリーニングの精度評価を行った．

Postprocess of virtual screening by learning docked structure using random forest
Nobuaki Yasuo†, Masakazu Sekijima†, ‡
†Department of computer science, Tokyo institute of technology
‡Advanced Computational Drug Discovery Unit, Tokyo institute of technology

2. 研究手法

2.1 概要

図 1 に本研究におけるリランキング手法の概要を示す。まず標的タンパク質に対して活性が既知の化合物をドッキングし，その結果からタンパク質とリガンド間の相互作用エネルギーを抽出したものである相互作用エネルギーベクトルを得る．次に Random forest によりこれらの相互作用エネルギーベクトルを学習し，活性の有無を予測するモデルを作成する．活性の有無が不明な化合物は，同様にドッキングと相互作用エネルギーベクトルの抽出を行い，予測モデルによる評価を行う．最終的なスコアは，活性があるクラスに分類される確率として表される．

2.2 機械学習

各化合物を表す特徴量は，ドッキング結果における各アミノ酸残基とリガンド間の相互作用エネルギーであり，各残基について，ファンデルワールス相互作用，静電相互作用，水素結合の 3 次元をもち，活性が既知の化合物のラベルは活性の有無を表す二値を用いた．すべてのタンパク質において，Random forest に用いられる木の数は 1000，各木が使用する特徴数は 6 とした．Random forest の実装は Scikit-learn version 0.18.1 を用いた．

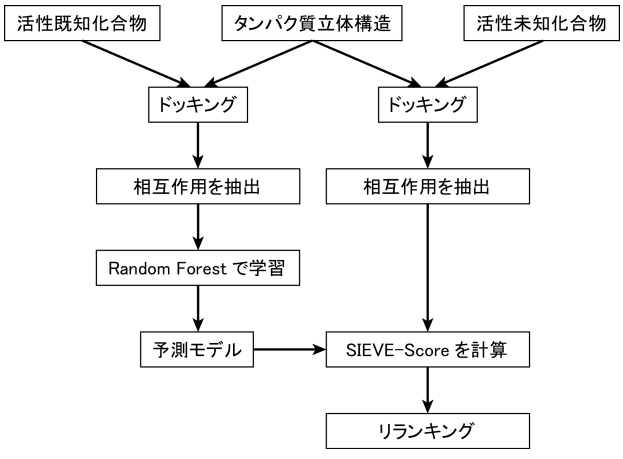

図 1. SIEVE-Score によるリランキングの概要

2.3 比較実験

比較実験では，ドッキングに Glide SP mode version 65013, データセットに DUD-E[5] を用いた．DUD-E には 102 タンパク質が登録されており，各タンパク質について共結晶構造，活性のあるアクティブ化合物，活性がないと考えられるデコイ化合物が登録されている．
本研究での実験では，各タンパク質において，共結晶構造のリガンド位置に全化合物をドッキングした．ドッキング結果が得られた化合物について 5-fold クロスバリデーションによりリランキング結果の ROC 曲線を求め，ROC 曲線下の面積である AUC の平均値を用いてリランキング前の Glide SP モードとの比較を行った．

3. 結果

DUD-E 102 タンパク質において，SIEVE-Score, Glide SP モードに対して 97 標的で AUC が向上する結果となった．図 2 に Diverse サブセットから 2 タンパク質の ROC 曲線を例として示す．青・緑・赤・水色・紫の線は SIEVE-Score の 5-fold における各 fold，黒点線は 5-fold の平均，黄線は Glide SP，灰点線はランダム予測の結果を表す．いずれも False Positive Rate < 0.05 の部分が特に改善しており，上位を正しく順位付けすることが求められるバーチャルスクリーニングにおいて有用であることがわかる．

4. 考察

SIEVE-Score が Glide SP モードに比べて精度が向上した理由は主に二つ考えられ，一つは既存の活性化合物の情報を取り入れた点である．SIEVE-Score では既存の活性化合物と類似したドッキング構造である場合に良いスコアとなるため，相互作用ベースで化合物比較を行う Interaction Fingerprint に近い効果を得ることができ，精度の向上に繋がったと考えられる．

もう一つは，使用した特徴量がそれぞれの相互作用の強さを直接表すということである．SIFt では相互作用の有無，Pharm-IF では原子ペア間の距離を特徴量としているが，より直接的な値である相互作用エネルギーの強さの方が学習に適した特徴量となっている可能性が考えられる．

本研究では，タンパク質-リガンド間ドッキングを用いたバーチャルスクリーニングの精度を向上させるための手法として，ドッキング構造から相互作用エネルギーを抽出して機械学習を行う SIEVE-Score を提案した．本手法は DUD-E データセットに含まれる 102 タンパク質中 97 タンパク質で Glide SP モードに対して精度向上を達成することに成功した．今後は，Random forest の利点である特徴量の寄与の活用などを行い，より有用な創薬支援手法としていきたい．

参考文献

[1] Chiba, S., *et al.*: Identification of potential inhibitors based on compound proposal contest: Tyrosine-protein kinase Yes as a target, *Scientific Reports* **5**, 17209 (2015).
[2] Lionta, E., Spyrou, G., Vassilatis, D. K. and Cournia, Z.: Structure-based virtual screening for drug discovery: Principles, applications and recent advances, *Current topics in medicinal chemistry*, **14**:16, 1923 (2014).
[3] Deng, Z., Chuaqui, C. and Singh, J.: Structural in- teraction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions, *Journal of medicinal chemistry*, **47**:2, 337–344 (2004).
[4] Sato, T., Honma, T. and Yokoyama, S.: Combining machine learning and pharmacophore-based interaction fingerprint for *in silico* screening, *Journal of chemical information and modeling*, **50**:1, 170–185 (2009).
[5] Mysinger, M. M., Carchia, M., Irwin, J. J. and Shoichet, B. K.: Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking, *Journal of medicinal chemistry*, **55**:14, 6582– 6594 (2012).

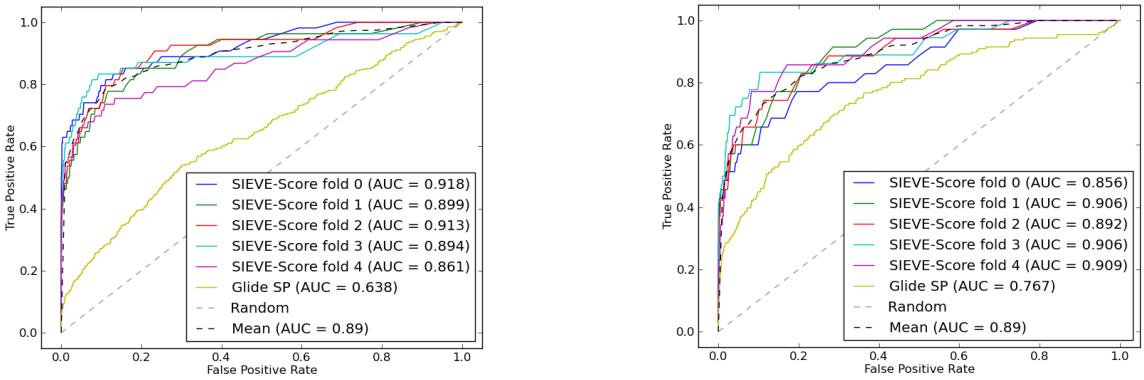

図 2: DUD-E における ROC 曲線の比較
青・緑・赤・水色・紫線: SIEVE-Score 5-fold の各 fold, 黒点線:5-fold の平均, 黄線: Glide SP, 灰点線: ランダム予測.
標的タンパク質：左：Serine/threonine-protein kinase AKT (akt1), 右：Glucocorticoid receptor (gcr).