様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　　８　日現在

機関番号：　３２６６０
研究種目：　基盤研究(C)
研究期間：　2012～2014
課題番号：　２４５００３６４

研究課題名（和文）がん創薬を支援するバーチャルランダムスクリーニング法の開発

研究課題名（英文）Developing a virtual random screening method for cancer drug discovery

研究代表者
大和田　勇人（Ohwada, Hayato）

東京理科大学・理工学部・教授

研究者番号：３０２０３９５４

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：本研究では機械学習を用いた、創薬支援のための高性能なバーチャルスクリーニング手法を提案する。コンピュータを用いた創薬研究では研究者はドッキングソフトを用いており、ドッキングシミュレーション結果や化合物の構造、その他化合物の情報を総合的に考えて、化合物がタンパク質に結合するかを判定している。現在、ドッキングソフトの性能は高いとはいえない。本研究では、創薬の専門家の経験や知識を利用し、ドッキングソフトの結果や化学的な情報を基に高性能な化合物スクリーニングをおこなう手法を提案した。実験より、本研究の手法は98.4%と高い精度と良好なROC曲線を得られることが確かめられている。

研究成果の概要（英文）：This study presents a high performance virtual screening method for drug design based on machine learning. In drug discovery with computers, drug designer often use docking software and decide the docking between the compound and the protein with the result of docking software, structure of the compound, and any information of compound. Currently, the performance of docking software is not high. The present method exploits the experiential knowledge of pharmaceutical researchers and allows to screen compounds with high performance based on the result of the docking software and chemical information of compounds. The experiment shows our method have high-accuracy as 98.4 % and excellent ROC curve.

研究分野：　知能情報学

キーワード：　機械学習　創薬

１．研究開始当初の背景

インシリコスクリーニングは非常に強力な創薬技術である．従来はドッキングソフトを使用して，阻害候補化合物と標的となるタンパク質との結合を，それぞれの構造に基づいて予測する．しかし，膨大な計算時間とタンパク質の三次元構造を必要とする制約があるため，近年は機械学習を用いた結合予測の研究が盛んに行なわれている[1][2]. 機械学習手法はタンパク質の強阻害剤(リガンド)と弱阻害剤(デコイ)をトレーニングセットとして用意して教師有り学習を行い，得られたモデルを使用して阻害候補化合物を予測する．しかし，公開されているデータベースには弱阻害剤は多く収録されていないという問題に直面している．弱阻害剤を用意するために，二つの手法が用いられている．一つは，標的タンパク質とは違うタンパク質の強阻害剤を収集し，標的タンパク質の弱阻害剤として扱う方法である．そのような弱阻害剤は薬効性が高く，標的タンパク質の強阻害剤となりうる．他の手法として，ZINC などの公開された化合物データベースからランダムに化合物を選択し，弱阻害剤として扱う方法がある．しかし，これらの化合物の構造は強阻害剤と似てしまう可能性がある．

２．研究の目的

本研究では，リガンドとデコイの両方を含むデータベースである DUD-E[3]に焦点を当てる．DUD-E に収録されているデコイのいくつかは，タンパク質に強く結合しないことがわかっている．他のデコイは，登録されているリガンドやユーザの用意したリガンドに対し，トポロジカルに異なる構造をもつ化合物を機械的に生成している．DUD-E を用いた機械学習と，既存のドッキングソフトウェアの予測精度を比較することで，スクリーニングにおける DUD-E の有用性を検証する．

３．研究の方法

まず，機械学習に用いるデータを DUD-E から抽出する．DUD-E には標的タンパク質に結合する化合物（リガンド）が actives_final.sdf として，結合しない化合物が decoys_final.sdf（デコイ）として用意されている．DiscoveryStudio を使いこれらの化合物の物理化学的性質を計算し，この性質を機械学習の入力データとする．また，リガンドに対してデコイの数が多いためオーバーフィッティングを起こす可能性がある．これを避けるため実験には，デコイの数はリガンドの３倍に制限した．

機械学習は求めた性質を特徴ベクトルとし，リガンドのラベルを１，デコイのラベルを０とする．特徴ベクトルは独立変数，ラベルは従属変数として扱う．これらの変数を使用して，SVM は，回帰式を算出する．この計算は，Support Vector Regression（SVR）として知られている． SVR は，SVM における超平面の計算に基づいて算出される．我々は，この計算値をドッキングスコアとして定義した．閾値を 0.5 として，出力が閾値以上の場合リガンド，閾値未満の場合デコイと判定する．

本研究では SVR の計算に LIBSVM を使用する．カーネル関数を使用して，非線形回帰式を算出することができる．我々は，スクリーニングにおいて最高のパフォーマンスを発揮する RBF カーネルを選択した．RBF ではペナルティパラメータ“コスト”とカーネルパラメータ”ガンマ”を設定する必要がある．SVR の予測精度はこれらのパラメータに大きく依存する．最良のパラメータを選択すするためにグリッドサーチを使用する．コスト{1, 10, 100, 1000, 10000}，ガンマ{0.1, 0.01, 0.001, 0.0001, 0.00001, 0}から学習データ内で予測精度が最高となる組み合わせの探索を行った．

分類性能の評価指標には正確度，適合率，再現率，F 値がある．スクリーニングを行う際，ほとんどの場合デコイの数がリガンドの数を上回る．たくさんのデコイの中からリガンドを見つけることが重要なので，適合率と再現率の両方が高い時にスクリーニング性能が高いといえる．従って，この二つの調和平均である F 値が高くなるようなパラメータをグリッドサーチで選択する．

機械学習手法の評価には Leave-One-Out Cross Validation を用いる．またドッキングソフトウェアの LibDock と CDOCKER と判別性能を比較し，本手法の有効性を検証する．機械学習とドッキングソフトウェアの性能を比較する際，ドッキングソフトウェアは分類の閾値がないため，分類性能の評価値で比較することができない．そこで，ROC 曲線を使用して比較を行う． ROC 曲線は，様々に閾値を変えた時の，分類性能の変化を示す．ROC 曲線の曲線下の面積は AUC と呼ばれ，評価の指標に用いられる．AUC が 0.5 のときランダムな分類であり，1 に近いほど分類性能が高い．

４．研究成果

Leave-One-Out Cross Validation とドッキングシュミレーションの比較から，提案手法は優れたスクリーニング性能を示した．表 1，表 2 は、ドッキングソフトウェアの分類性能を示している。ligand と decoy の列はドッキングソフトウェアによって分類することができた化合物の数を示す。ROC の列は、図 1 に示す ROC 曲線の AUC 値を示している．表 3 は本手法である SVR による分類精度を示している．正確度と F1 スコアは閾値が 0.5 の時の分類性能を示している。

表 1，2，3 から，提案手法の AUC はすべての標的タンパク質で，既存手法のドッキングソフトよりも高い値となった．さらに，5 つのタンパク質で本手法の AUC は 1 となった．pur2 における LibDock と CDOCKER の AUC は高

表 1. LibDock の分類性能

target name	LibDock		
	ligand	decoy	AUC
ada17	790/959	2600/2877	0.775
aofb	118/168	397/504	0.710
cah2	602/835	2103/2505	0.543
fabp4	55/57	156/171	0.292
hs90a	116/125	362/375	0.239
jak2	134/153	400/459	0.653
pur2	180/201	573/603	0.993
sahh	52/190	422/570	0.771
thb	155/168	408/504	0.626
xiap	109/129	373/387	0.944

表 2. CDOCKER の分類性能

target name	CDOCKER		
	ligand	decoy	AUC
ada17	949/959	2847/2877	0.719
aofb	167/168	502/504	0.586
cah2	408/835	1147/2505	0.312
fabp4	56/57	169/171	0.793
hs90a	120/125	373/375	0.420
jak2	153/153	458/459	0.613
pur2	201/201	600/600	0.994
sahh	189/190	511/570	0.788
thb	167/168	501/504	0.839
xiap	129/129	386/387	0.974

表 3. SVR の分類性能

target name	SVR		
	accuracy	f1score	AUC
ada17	0.994	0.987	0.999
aofb	0.938	0.863	0.951
cah2	0.985	0.969	0.997
fabp4	1.000	1.000	1.000
hs90a	0.992	0.984	1.000
jak2	0.974	0.946	0.983
pur2	1.000	1.000	1.000
sahh	0.999	0.997	1.000
thb	0.993	0.985	1.000
xiap	0.990	0.980	0.991

い値を示したが，提案手法はさらに高い 1 を示した．また，cah2，FABP4，hs90a における LibDock と CDOCKER の AUC はとても低い値となったが，提案手法も AUC はとても高い値を示した．従って，LibDock と CDOCKER の分類性能に関係なく，提案手法は優れたスクリーニング性能を示す

図 1 は，各タンパク質での ROC 曲線である．LibDock において異なる分類性能を示した 4 つのタンパク質を選択した．これらのグラフから本手法の ROC 曲線が LibDock と CDOCKER よりも優れていることを確認できる．従って，提案手法が LibDock と CDOCKER よりも良いスクリーニング手法であると言える．しかし，aofb の ROC 曲線は他の ROC 曲線に比べ低いものとなった．これを改善するためには，デコイの検出を防ぐように閾値を高く設定する（例えば 0.7）必要がある．

表 3 の分類結果から，提案手法が優れた精度でリガンドとデコイを分類することができることを確認した．各タンパク質でデコイ

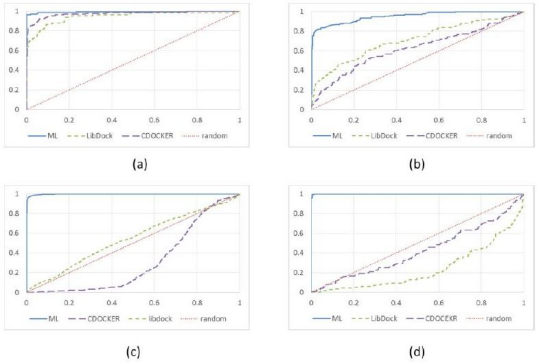

図 1. ROC 曲線の比較

(a) xiap，(b) aofb,，(c) cah2，(d) hs90a

はリガンドの 3 倍であるため，すべてデコイと分類した時の正確度は 75%である．提案手法の最低の正確度は 93.8%だった．しかし，各タンパク質で F 値が高かったため，適合率と再現率の両方が高い．つまり，高精度にリガンドを検出できているといえる．

この手法はタンパク質の構造を必要としないため，結晶構造なしに本手法を適用することができる．DUD-E 内にあるタンパク質であればどのタンパク質にも適用可能であり，高いスクリーニング性能を持つため未知の化合物から新しい阻害剤を発見することができるだろう．

＜引用文献＞

[1] Jorissen R.N., Gilson M.K., Virtual screening of molecular databases using a support vector machine, Journal of Chemical Information and Modeling.Vol. 45, pp. 549-561, 2005.
[2] C. Springer, PostDOCK: A structural, empirical approach to scoring protein ligand complexes, J. Med. Chem., Vol. 48, pp. 6821-6831, 2005.
[3] Mysinger MM., Carchia M., Irwin JJ., Shoichet BK., Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking, Journal of Medicinal Chemistry, Vol. 55, No. 14, pp 6582-6594, 2012.

５．主な発表論文等

（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

①岡田 正人、金盛 克俊、青木 伸、大和田 勇人、機械学習による創薬支援のための高精度バーチャルスクリーニング法の開発、人工知能学会誌、査読有、29 巻、2014、194--200，10.1527/tjsai.29.194.

〔学会発表〕（計 9 件）

①岡田正人、SVM を用いたインシリコ創薬に

おける新薬の有効性判別、人工知能学会全国大会 2012、2012.6.14、山口県教育会館（山口県山口市）
②岡田正人、ドッキングシミュレーション結果と化合物情報の学習による化合物の結合判別、第 11 回情報科学技術フォーラム、2012.9.4、法政大学 小金井キャンパス（東京都小金井市）
③Masato Okada, Binary Classication of Compounds by Learning from Docking Software Results and Chemical Information, 5th International Conference on Bioinformatics and Computational Biology,2013.3.5, Honolulu, Hawaii, USA.
④岡田正人、がん創薬を支援するバーチャルランダムスクリーニング法の開発、人工知能学会全国大会 2013、2013.6.4、富山国際会議場（富山県富山市）
⑤Masato Okada, Docking Score Calculation by Machine Learning and an enhanced inhibitors database, International Symposium on Technologies against Cancer 2014, 2014.3.8, Tokyo, Japan.
⑥Masato Okada, Binary Classification of Compounds by Learning from Docking Software Result and Chemical Information, International Symposium on Technologies against Cancer 2014, 2014.3.8, Tokyo, Japan.
⑦Takaya Yoshida, Discovering Compounds That Activate Plant Immunity Using Machine Learning, 6th International Conference on Bioinformatics and Computational Biology, 2014.3.24, Las Vegas, Nevada, USA.
⑧Tadasuke Ito, Combining two machine learning methods for predicting protein-ligand docking using structure and physiochemical properties, 7th International Conference on Bioinformatics and Computational Biology, 2015.3.9, Los Angeles, USA.
⑨伊東 忠佑、化合物の構造情報と非構造情報を用いたタンパク質ドッキング予測の為の機械学習手法の検討、情報処理学会第 77 回全国大会、2015.3.18、京都大学（京都府京都市）.

６．研究組織
(1)研究代表者
大和田　勇人（OHWADA Hayato）
東京理科大学理工学部　教授
研究者番号：30203954

(2)研究分担者
青木　伸（AOKI Shin）
東京理科大学薬学部　教授
研究者番号：00222472