様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　　４　日現在

機関番号：　３２６６０
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００４０８

研究課題名（和文）リガンドデータベースからの機械学習によるハイブリッドスクリーニング法の開発

研究課題名（英文）Developing a hybrid screening method based on machine learning with Ligand database

研究代表者

大和田　勇人（Hayato, Ohwada）

東京理科大学・理工学部経営工学科・教授

研究者番号：３０２０３９５４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：リガンドデータベースを活用した機械学習によるたんぱく質と化合物の結合予測を行った。インシリコ創薬は薬として有望な化合物（リガンド）をコンピュータで選別する手法であるが、ここでは化合物の化学的性質をからSupport Vector Machine(SVM)による機械学習に加えて、化合物の構造を学習するInductive Logic Programming(ILP)を組み合わせ、予測精度の向上を図った。次に、がん放射線治療の副作用低減のためにp53標的放射線防護剤を候補化合物を予測することをターゲットにした。その成果はジャーナルや国際会議で発表した。

研究成果の概要（英文）：This research focuses on a hybrid machine learning method to predict chemical properties of drug candidates using ligand databases. In-silico screening is a promising selection method for drug discovery, we have combined support vector machines with inductive logic programming, yieding a new method for improving the predictive accuracy for drug candidate selection. Moreover, p53 targeting radio protective compounds are predicted to decrease the side effect of radio based therapy. The outcomes are presented at a journal and international conference proceedings.

研究分野：　知能情報学

キーワード：　機械学習　化合物スクリーニング　リガンドデータベース

１．研究開始当初の背景

薬剤候補化合物とタンパク質が結合するか否かを判定することは創薬の重要課題の一つとして研究されてきたが、従来は定量的な力学計算を伴うドッキングシミュレーションによる方法が中心であった。しかしながら、そのためのモデル化はかなりの手間であると同時に計算量も大きい。加えて、その方法論は特定のタンパク質に依存するものが多い。本研究は、こうした問題に対し、(1) 結合するか否かがすでにわかっているデータベースを活用し、そこから機械学習を使って、簡易で汎用な結合判定機を開発できるのではないかという観点に立って、研究を開始した。また、薬剤との結合の仕方が明らかになっておらず、ドッキングシミュレーションを行うことができないタンパク質も多く存在する。そうしたタンパク質に対して、(2) 薬剤候補化合物の物理化学的特性を元に、機械学習を用いて結合判定機を作成することができるのではないかという観点に立って、研究を開始した。

２．研究の目的

最新のリガンドデータとデコイデータから機械学習し、リガンド判別モデルを自動生成することで、薬剤候補化合物とタンパク質が結合するか否かを高速かつ高精度に判定するバーチャルスクリーニング法を開発することが研究目的である。(1) 帰納論理プログラミング（Inductive Logic Programming, ILP）による明示的なパターン判別規則の生成と市販ソフトの計算による物理化学特性からの機械学習を併せ持つハイブリッド型としての特徴をもち、前者は創薬研究者に対する対話的な支援、後者は判別性能の向上を実現する。本方法は従来の定量的な力学計算を伴うドッキングシミュレーションを越えて、特定のタンパク質に依存せず、最新の研究成果や実験等で得られた知見を有効活用した薬剤候補を絞りこむための画期的な方法である。(2) 市販ソフトの計算による物理化学的特性を用いて、機械学習による p53 タンパク質阻害剤（放射線防護剤）の候補化合物の判定を行う。本手法はドッキングシミュレーションを行うことのできないタンパク質に対して有効なスクリーニング手法である。

３．研究の方法

(1) 方法の概要を図 1 に示す。まず、公開データベースを出発点にリガンド・デコイデータを収集する。特に、DUD-E はリガンドとタンパク質のドッキング性能を測るためのベンチマークデータベースとして非常によく整理されており、これを活用してこれ以降の機械学習システムの開発を並行して進められるようにする。青木の創薬研究に関する知見とこれまでの実験成果をベースに、どのタンパク質を標的にするか検討し、重要度の高いものから計算実験を行う。計算実験は独自開発した GKS を使用し、予備的な結果が確実に得られる工夫を施す。

次に、帰納論理プログラミングのサブクラスである SIP(Simple Inductive Programming)を定義し、すでにプロトタイプとして開発してある Prolog 版の SIP を Java で再実装する。SIP は関数がなく、変数間の入出力が規定された論理プログラミングであり、関係学習が可能なように十分な表現力を維持している。Java による実装で並列・分散化が容易になり、大幅な計算効率の改善が期待される。一方、Discovery Studio による化合物の物理化学特性の計算も同時並行して実施し、市販のソフトである Discovery Studio を用い、機械学習への入力となる物理化学特性（特徴量）を数種類の標的タンパク質に登録されている化合物全体に対して計算する。

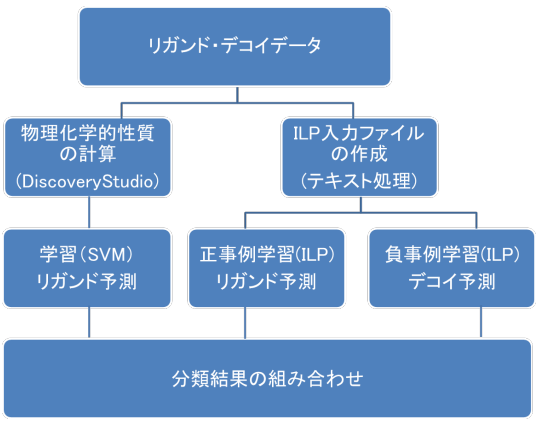

図 1．SVM と ILP を組み合わせた学習の手順

(2) 青木らによって合成、実験が行われた p53 阻害剤候補化合物の 3 次元構造から、Discovery Studio を用いてこれらの物理化学的構造を計算する。このデータを用いて、候補化合物の放射線防護機能と細胞毒性をそれぞれ予測するモデルを、機械学習を使って作成する。方法の概要を図 2 に示す。

放射線防護機能や細胞毒性の有無の予測については、Support Vector Machine (SVM)、Random Forest、k 近傍法といった一般的な機械学習手法に加え、勾配ブースティングの一種である XGBoost を用いる。

またこれら 2 つの指標を合わせて考慮し、より効率的なスクリーニングを可能にするために、候補化合物のランキングを行う。ランキングを行うにあたり、細胞毒性や放射線防護機能の有無といった二値分類ではなく、それぞれの程度を予測する回帰分析を行う。回

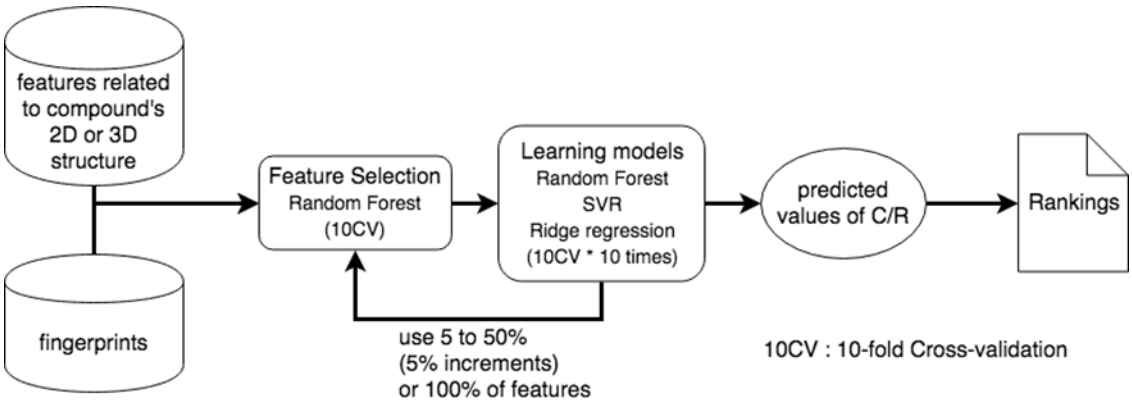

図2．p53阻害剤候補化合物ランキング作成の流れ

帰の手法として、SVM の一種である Support Vector Regression (SVR)、Random Forest、Ridge 回帰を用いる。ランキング手法として３つの手法を提案する。細胞毒性を索引とした辞書式順序、細胞毒性が低く、放射線防護機能が高いほど長くなるような距離による順序に加えて、多目的最適化手法で用いられるパレートランキング法を取り入れた順序の３手法である。以上の手法を用いて、創薬研究者の意見をより反映したランキングを作成する。

４．研究成果

（1）リガンドデータベースを活用した機械学習によるタンパク質と化合物の結合予測を行った．インシリコ創薬は薬として有望な化合物（リガンド）をコンピュータで選別する手法であるが，ここでは化合物の科学的性質を用い Support Vector Machine（SVM）などの単一の機械学習が提案されてきた．一方，本研究では SVM に加え，構造を学習する Inductive Logic Programming（ILP）を取り上げ，両者を組み合わせた学習手法を提案した．これは従来のアンサンブル学習とは違い，異なるタイプのデータからの学習が可能になり，予測精度の向上が期待できる。

まず SVM では学習結果から得られる各化合物と分類平面までの距離から信頼度を求め，ILP では，得られたルールの中から最高の評価値のルールを各化合物に適用し，被覆するかしないかで信頼度を求めた．さらに，正事例と負事例を反転させて ILP を実行し，負事例と予測した化合物に対する信頼度も算出した．最後に，これら３つの機械学習による信頼度を統合させ，その結果に基づいて予測を行った．DUD-E（リガンドデコイデータベース）に登録されている７つの創薬標的タンパク質で実験を行った結果，F 値に関して SVM 単体に比べ最大 0.06 向上させることができ，さらに他の組み合わせ方法に比べ本手法の F 値が高いことが示された。

（2）XGBoost に基づく新しい機械学習方法を提案し、84 個の化合物データを用いて放射線防護機能と毒性の有無に関してそれぞれ予測した．それぞれの予測精度の比較を図 3、4 に示す。毒性と放射線機能についてのラベル付けは専門家の意見を参考に設定し，学習に用いた特徴は，３D モデリングソフトウェアの Discovery Studio によって算出された 217 の特徴量を用いた． k 近傍法，SVM，Random Forest と精度を比較した結果、毒性の予測については本方法が最もよく、予測精度 83.8%を達成した。

また、81 個の化合物データを用いて放射線防護機能と細胞毒性の程度をそれぞれ予測する回帰モデルを作成した。学習に用いた特徴量は上記の 217 に加えて、化合物の構造をバイナリで表現したフィンガープリント 960bit のうち分散が 0 ではなかった 795bit である。それぞれの予測精度の比較を図 5、6

表1.それぞれの標的タンパク質に対する結合予測の精度(F値)

	F-measure						
分類器	AMPC	CP2C9	CP3A4	FABP4	FGFR1	MK10	NOS1
SVM	0.851	0.838	0.865	0.925	0.535	0.878	0.922
RandomForest	0.814	0.810	0.818	0.909	0.541	0.871	0.904
ILP(正事例)	0.774	0.744	0.785	0.830	0.591	0.827	0.842
ILP(負事例)	0.797	0.733	0.737	0.901	0.467	0.799	0.712
SVM + RF	0.844	0.836	0.865	0.928	0.532	0.883	0.927
SVILP	0.778	0.732	0.772	0.803	0.474	0.830	0.867
単純多数決	0.868	0.836	0.873	0.934	0.597	0.883	0.933
提案手法	0.876	0.840	0.876	0.938	0.599	0.891	0.933

に示す。その結果、どちらの予測においても SVR が最も高精度に予測することができた。また Random Forest により特徴量の重要度を計算したところ、フィンガープリントが表現する化合物の特性が、化合物の放射線防護機能に寄与するところが大きいことを示した。

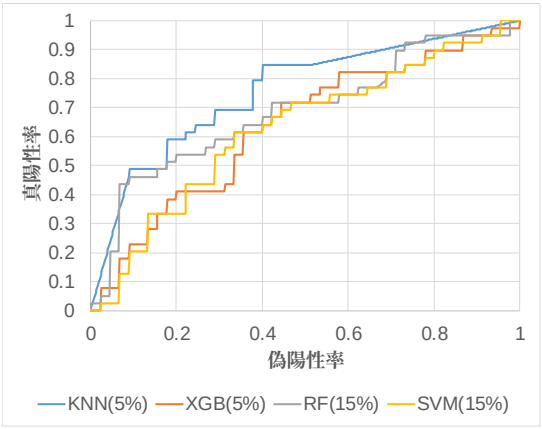

図 3．放射線防護機能の有無の予測における、各手法の ROC 曲線

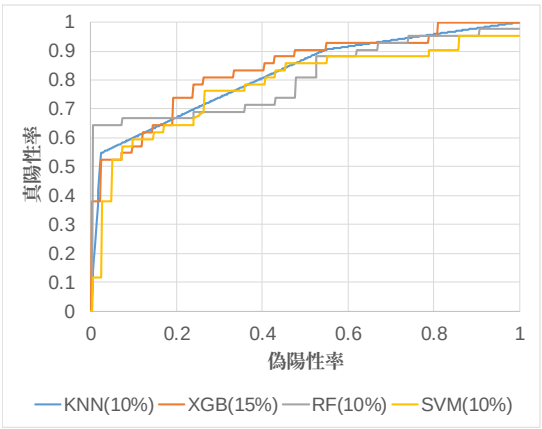

図 4．細胞毒性の有無の予測における、各手法の ROC 曲線

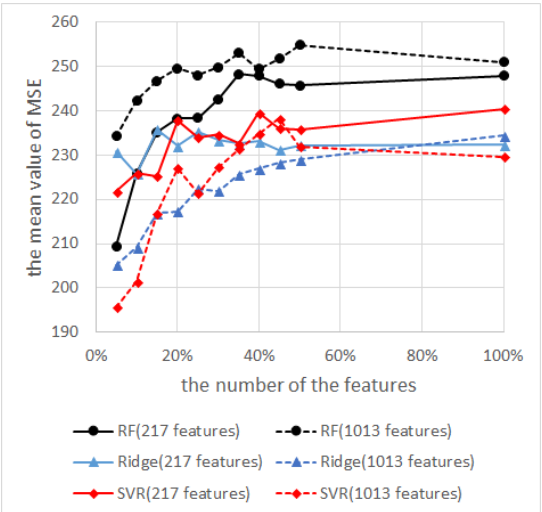

図 5．放射線防護機能の程度を予測する回帰モデルの精度比較(平均二乗誤差による)

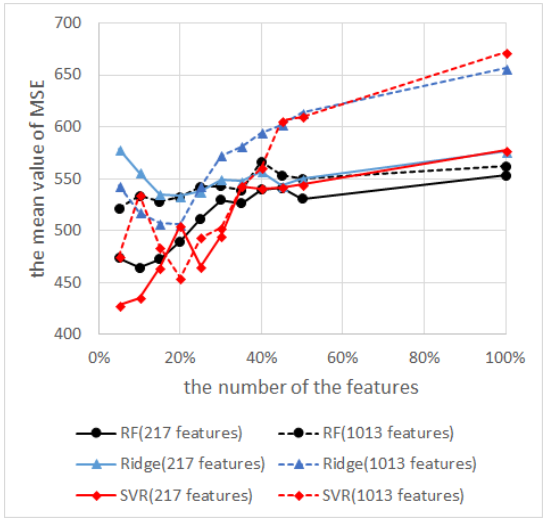

図 6．細胞毒性の程度を予測する回帰モデルの精度比較(平均二乗誤差による)

表 2．予測値を用いて作成した候補化合物ランキング

dictionary order		distance order		combination method	
rank	compound	rank	compound	rank	compound
1	MH-15	1	MH-15	1	AS-16
2	Vitamin_C	2	Vitamin_C	2	YN-9
3	YM-13	3	YM-13	3	YN-1
4	KT-2	4	YN-9	4	YM-13
5	YN-1	5	YN-1	5	MH-15
6	SAr-3	6	KT-2	6	AS-9
7	YN-9	7	SAr-3	7	KH-13
8	Vitamin_E	8	KH-13	8	KT-2
9	SAr-2	9	AS-16	9	Vitamin_C
10	KH-13	10	Vitamin_E	10	SAr-3
11	KH-25	11	SAr-2	11	YT-1
12	KH-24	12	KH-25	12	KH-20
13	naphthol	13	YN-7	13	YN-7
14	YN-5	14	KH-24	13	KH-21
15	AS-16	15	naphthol	15	KH-25
16	YN-7	16	YN-5	16	Vitamin_E
17	KT-1	17	KH-21	17	SAr-1
18	SAr-1	18	KH-18	17	YN-5
19	AS-4	19	AS-9	19	KH-18
20	KH-21	20	SAr-1	20	naphthol

さらに、これらの回帰モデルによる予測値を用いて表 2 のような p53 阻害剤の候補化合物のランキングを作成した。それぞれの手法を比較した結果、専門家の意見をよく反映したランキング手法を提案することができた。それぞれの手法によるランキングがどの程度正しく予測できているかをスピアマンの順位相関係数を用いて比較したところ、細胞毒性の低さと放射線防護機能の高さを距離

で表現し、大きな順に並べたランキングが専門家の意見を良く反映し、かつ予測が比較的容易であることが示された。

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　4 件）

① Tadasuke Ito, Masato Okada, Shotaro Togami, Shin Aoki and Hayato Ohwada, In Silico Screening of Zinc(II) Enzyme Inhibitors by SVM, Proc of the 6th International Conference on Computational Systems-Biology and Bioinformatics, 2015, pp. 22-26.

② Masato Okada, Tadasuke Ito, Hayato Ohwada and Shin Aoki, Docking Score Calculation Using Machine Learning with an Enhanced Inhibitor Database, Journal of Medical Imaging and Health Informatics, Vol.5, No.5, 2015, pp.1104-1107, http://dx.doi.org/10.1166/jmihi.2015.1503

③ Masataka Kimura, Shin Aoki, Hayato Ohwada, Predicting radiation protection and toxicity of p53 targeting radioprotectors using machine learning, 2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, 2017, pp.106, DOI: 10.1109/CIBCB.2017.8058540

④ Kuswanto H., Melasasi J.N., Ohwada H, Enzyme Classification on DUD-E Database Using Logistic Regression Ensemble (Lorens), Innovative Computing, Optimization and Its Applications. Studies in Computational Intelligence, Vol.714, 2018, pp.93-109, https://doi.org/10.1007/978-3-319-66984-7_6

〔学会発表〕（計　0 件）

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計　0 件）

○取得状況（計　0 件）

〔その他〕
特になし。

６．研究組織
(1)研究代表者

大和田　勇人（OHWADA, Hayato）
東京理科大学・理工学部経営工学科・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・薬学部生命創薬科学科・教授
研究者番号：00222474

(3)研究分担者

西山　裕之（NISHIYAMA, Hiroyuki）
東京理科大学・理工学部経営工学科・教授
研究者番号：80328567