RG-001

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
Binary Classification of Compounds by Learning from Docking Software Results and Chemical Structures

岡田 正人† 青木 伸‡ 金盛 克俊† 大和田 勇人†
Masato Okada Shin Aoki Kanamori Katsutoshi Hayato Ohwada

1. はじめに

標的タンパク質に結合する化合物は病気を治す可能性がある．そのため，ドッキング実験によって大量のタンパク質から標的タンパク質に結合する化合物を探索するスクリーニングは創薬研究のなかで重要である．ドッキングソフトはドッキング実験をシミュレーションし，化合物のポーズと，タンパク質と化合物の結合の強さを計算することができる．それらの結果から，ポーが理にかなっていて，結合の強さが強い化合物を探すことができる．ドッキングソフトの精度は様々な研究により確かめられており[1][2]，特にポーズの特定では非常に精度が高い．また，ドッキングソフトを利用し，創薬に成功した例も存在する[3] [4].

しかし，ドッキングスコアの計算方法は不完全であり，標的タンパク質によっては問題が発生する[5] [6] [7]．問題の一つはスクリーニングに用いる際のスコアの性能で，タンパク質に結合できない化合物のドッキングスコアも，しばしば高く出力してしまう[8]．また，多くのドッキングソフトは，金属原子を含むタンパク質を適切に扱うことができない[9]．これらの原因は一つの計算方法をすべてのタンパク質に適用させていることである[10].

分子動力学法(MD) [11]ならば，これらの問題を回避することができる[12]．しかし，MD はドッキングソフトの数十倍の計算時間が必要なため，スクリーニングに適さない.

ドッキングスコアの性能を向上させるため，複数のドッキングソフトが出力するドッキングスコアを利用する手法が提案されている[13]．しかし，どのドッキングソフトでもスコアの性能が悪い場合は，性能はあまり向上しない.

問題解決と計算速度を両立させるため，機械学習を用いる手法が提案されている．Jorissen[14]は化合物の性質を示す 517 個の QSAR 記述子 [15]を Dragon [16]で作成し，使用している．そして，Support Vector Machine (SVM) [17] の回帰機能を用いて各化合物の新しいスコアを計算している．また，ドッキングソフトで得られるポーズを利用して機械学習を行う手法が提案されている．Deng [18] や Ballester[19] ，Sato[20]はタンパク質の各原子と化合物の各原子の距離を用いた機械学習を行っている．Springer [21] や Sarah [22]は溶媒露出面積や水素結合エネルギーを用いた学習を行っている．これらの研究ではドッキングソフトの出力するポーズを利用するが，ドッキングスコアはほとんど利用されていない．この理由として次のことが考えられる．第一に，ドッキングスコアの性能が低いということである．ドッキングスコアの計算方法は不完全であり，結合力を正確に表現できない．そのため，機械学習におけるノイズとなる可能性がある．第二に，ドッキングソフトで得られるドッキングスコアは一つだけであり，情報が少なすぎる．そのため，化合物の性質やタンパク質と化合物を結合させた時の性質を情報として加える必要がある.

本論文では，従来用いることが困難であったドッキングソフトの出力するドッキングスコアを学習し，タンパク質に化合物が結合するかどうかの分類を行う手法を提案する．本論文ではドッキングスコアの性能改善手法と同様に多数のドッキングスコアを用い，特徴量とする．また，比較と精度改善のため，化合物の性質を示す QSAR 記述子を特徴量として用いる．以上の特徴量を用いて，本論文では機械学習を用いた分類を行う．そして、多数のドッキングスコアを特徴量として用いるため，次元の呪いに強い[23] [24] SVM を用いる．また、本論文では正例をタンパク質に結合する化合物(活性化合物)とする．そして，負例をタンパク質に結合できない化合物(非活性化合物)とする．本論文ではこれらのデータを用いて，分類モデルと回帰モデルの 2 つを作成する．そして，調査対象の化合物がタンパク質に結合するかどうかの分類と，結合の強さをモデルから計算する．提案手法は 2 つの数値のみを出力するため，ドッキングポーズを出力するドッキングソフトと異なり情報量が少ない．そのため，大量の化合物から活性化合物を探索するスクリーニングでは有効であるが，化合物の特性をより詳しく分析するためには，ドッキングソフトの出力結果を利用する必要がある．提案手法の性能は実際の結合情報を用いて評価する．実験では交差検定を用いて活性化合物の検出性能を調べる．そして，実験結果からドッキングスコアを用いた機械学習がスクリーニングにおいて有効であることを示す.

2. 提案手法

ここでは我々が提案するドッキングソフトの出力を利用した機械学習手法について述べる.

2.1 特徴量

本論文では機械学習においてドッキングソフトの出力するドッキングスコアを利用する．通常，一つの化合物は一つのドッキングスコアを得ることができる．このとき，ドッキングソフト d によって得られる化合物 p とタンパク質 p のドッキングスコア $S(c,d,p)$は次式,

$$S(c,d,p) = f_d(c,p) \quad \cdots(1)$$

で示される．関数 $f_d(c,p)$ はドッキングスコアを計算するためのドッキングソフト d の計算関数である．この関数はドッキングソフトによって大きく異なる．このとき，スコアは実数であり，値が大きいほどタンパク質 p と化合物 c の結合の強さが強いことを示す．ただし，いくつかのドッキ

† 東京理科大学大学院理工学研究科
‡ 東京理科大学薬学部

ングソフトでは値が小さいほど力が強い．本論文ではドッキングスコアの意味を統一するため，そのようなドッキングソフトでは次式，

$$S(c,d,p) = -f_d(c,p) \quad \cdots(2)$$

を適用する．

ドッキングソフトでは無作為に抽出した化合物を使用した場合，ドッキングスコアの分布は正規分布となる．本研究では標準化を行い，ドッキングスコアの分布を標準正規分布に変換する．このとき，標準ドッキングスコア $S'(c,d,p)$は次式，

$$S'(c,d,p) = \frac{S(c,d,p) - \mu_{dp}}{\sigma_{dp}} \quad \cdots(3)$$

で表される．μ_{dp}はドッキングソフト d で各化合物をタンパク質 p にドッキングさせた時のドッキングスコアの平均，σ はドッキングソフト d で各化合物をタンパク質 p にドッキングさせた時のドッキングスコアの標準偏差である．また，ソフトウェアのエラーによってドッキングスコアを得られなかった化合物では標準ドッキングスコアとして 0 を割り当てる．これは，エラーが機械学習に与える影響を低減するためである．

前述のように，ドッキングスコアは性能が低く，情報が少ないため，そのままでは機械学習に用いることはできない．そのため，多数のドッキングスコアを作成し，特徴量として用いる．第一に，複数のドッキングソフトを用いることで，複数のドッキングスコアを得る．これは複数のドッキングソフトを用いてドッキングスコアを計算し，ドッキングスコアを統合することで検出性能を高めるコンセンサススコアリングに基づく[13]．コンセンサススコアリングでは一般的に複数の値の中の最低値をとることで性能が高まる．しかし，どのタンパク質に対しても当てはまるわけではない．本論文では複数のドッキングスコアを用い，標的タンパク質と化合物との関係を示す複数の情報を得る．そのため，n 個のドッキングソフトを用いたとき，一つの化合物は n 個の標準ドッキングスコアを得る．

次に，複数のタンパク質を用いることで，複数のドッキングスコアを得る．これは標的タンパク質以外のタンパク質に対してもドッキングスコアを計算する手法に基づく[25]．この手法は他のタンパク質よりも標的タンパク質に対するドッキングスコアが高い化合物を優先して順位付することで，ドッキングソフトの性能を高める．本論文では複数のタンパク質に対するドッキングスコアを用い，複数のタンパク質と化合物との関係を示す情報を得る．これは，各タンパク質に対する活性化合物の共通性質を見つけることにつながる．その結果，n 個のドッキングソフトと l 個のタンパク質を用いた時，一つの化合物において $n \times l$ 個の標準ドッキングスコアを得る．本論文では $n \times l$ 個の標準ドッキングスコアを特徴量として使用する手法を Multi Target Scoring(MTS)と呼ぶ．

本論文では上記の MTS の比較対象として QSAR 記述子を特徴量とする手法を用いる．この手法は機械学習を用いるいくつかの研究で用いられている[14]．特に，原子数や質量といった，化合物の性質を示す記述子は，タンパク質に関わらず使用できるため，実装が容易である．本論文では化合物の記述子を特徴量とし，機械学習を行った際の比較対象とする．また，この特徴量と MTS を併用し，精度の向上が可能であるかどうかを調べる．化合物の記述子の数を r とすると，化合物は $n \times l + r$ 個の特徴量を持つ．

図 1 は本論文で用いる特徴量の行列を示している．各化合物は複数のドッキングソフトおよび複数のタンパク質を用いることで複数の標準ドッキングスコアを持つ．また，各化合物はタンパク質やドッキングソフトに関係しない，固有の記述子を持つ．また，化合物が標的タンパク質に結合するかどうかの情報がある場合，化合物に結合情報を示すラベルを付加する．化合物がタンパク質に結合する場合には+1 のラベルを付加し，結合しない場合には-1 のラベルを付加する．また，結合情報が無い場合には 0 のラベルを付加する．このような化合物を未ラベル化合物と呼ぶ．.本論文では複数のタンパク質を用いているため，化合物は複数のラベルを持つ．ただし，化合物が複数のタンパク質に対して+1 のラベルを持つことは非常に少ない．このラベルは機械学習において学習やテストに使用する．このとき，スクリーニングを行いたい標的タンパク質に対応するラベルのみを使用し，他のラベルは使用しない．

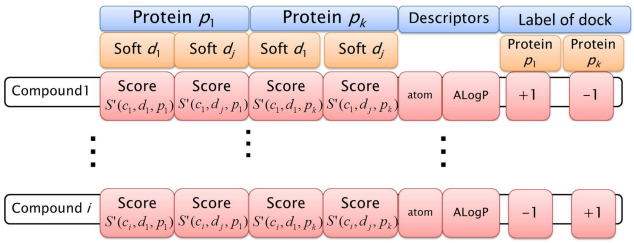

図 1. 特徴量行列

2.2 機械学習

ここでは，前述の特徴量を用いた機械学習について述べる．図 2 に機械学習の流れを示す．本論文では分類と回帰計算のため，SVM を用いる．SVM は次元の呪いに強く，多数の特徴量を使用できる．また，同一のデータで分類と回帰を行うことができる．第一に結合情報を含む特徴量行列を用いて学習を行う．特徴量行列には複数のタンパク質に対する結合情報が含まれる．そのため，各タンパク質の分類モデルと回帰モデルを作成することができる．しかし，それらのモデルは独立している．そのため，ここでは一つのタンパク質 p_k のモデル作成について述べる．

第一に分類モデルを作成する．一つの化合物の特徴量は標準ドッキングスコアと記述子で構成される．また，一つの化合物は p_k に対応する結合ラベルを持つ．ある化合物 c_i のラベルが+1 ならば，その化合物は p_k に結合できる．タンパク質 p_k に対応する化合物のベクトルデータ $V(c_i,p_k)$は次式，

$$V(c_i,p_k) = (S'(c_i,d_1,p_1),\cdots,S'(c_i,d_m,p_n), e_1,\cdots,e_r,L(c_i,p_k)) \quad \cdots(4)$$

で表される．e は c_i の記述子を示す．MTS のみを用いる場合には記述子は含まれない．また，$L(c_i,p_k)$は c_i の p_k に対応した結合情報を示す．このベクトルデータを学習することで，ラベルのないデータを分類するための分類モデルを作

成する．この分類モデルは未ラベル(unlabeled)化合物に対して+1 か-1 のラベルを付加する．本論文においては，もしラベルが+1 ならば，化合物は p_k に結合すると考える．また，もしラベルが-1 ならば，化合物は p_k に結合しないと考える．

第二に回帰モデルを作成する．学習で用いるデータは式(3)と同様である．ただし，回帰では $L(c_i,p_k)$ を従属変数とみなす．ベクトルデータを学習することで，従属変数を求めるための回帰モデルを作成する．この回帰モデルは未ラベル化合物に対して（おおよそ，ほとんど）-1~+1 のラベルを付加する．このラベルはドッキングスコアと同様に結合の強さを示す．本論文ではこの値を回帰スコアとよぶ．この回帰スコアを化合物の新たなスコアとすることで，スコアの性能向上を行う．ただし，分類モデルと異なり，回帰スコアは相対的な値である．そのため，未ラベル化合物の回帰スコアだけでは，化合物の評価が困難である．そこで，テストデータを用いて，比較用のスコアを作成する．本論文では各テストデータに対して，Leave one out cross validation を用いて回帰スコアを得る．そして，ドッキングソフトと同様に，活性化合物のスコアと比較することで，未ラベル化合物を評価する．

以上の手法は SVM の基本的な使用方法に基づいている．また，精度を向上させるための手法が使用できる．しかし，本論文のシステムにバギング[26]やブースティング[27]を組み込んだところ，ほとんど効果が見られなかった．そのため，今後はより問題に適した手法が必要である．

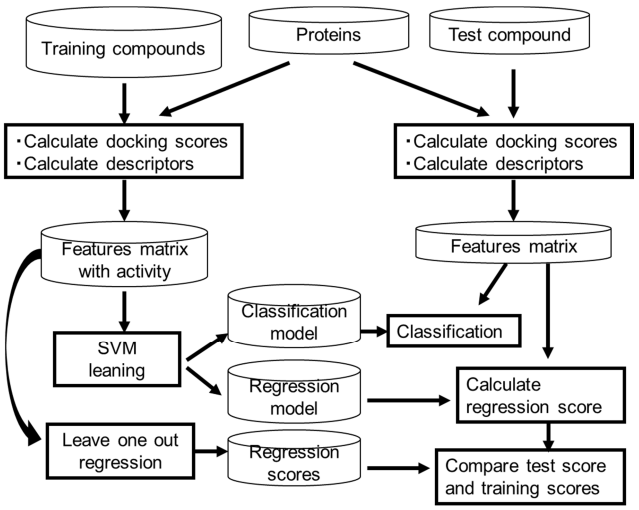

図 2. 機械学習の流れ

3. 実験環境

ここでは提案手法を評価するための実験環境について述べる．用いた化合物やタンパク質について述べ，その後システムの実行環境，そして実験方法について述べる．

3.1 データセット

本実験では結合情報を含む化合物を得るため，Binding Database のデータを用いた．Binding Database は化合物とタンパク質の結合情報を多く保有している．本論文では Binding Database 内の list of het molecules in PDB binding to Target を用いた．ここには，PDB 中に結合する化合物の情報が SD フォーマットで存在する．このリスト lists の中で，タンパク質との結合構造が PDB に登録されている化合物が多い標的タンパク質を選択する．本論文では上位 30 個を選択した．表 1 に選択した標的タンパク質とその情報，標的タンパク質に結合する化合物の数を示す．また，化合物と結合するときにとりうるタンパク質の各ポーズの中で，結合サイトが広くどの化合物でも入り込むことのできるポーズを選び，そのポーズを持つ複合体を PDB から取得した．表の PDBID は複合体の PDB を示す．また，それぞれの標的タンパク質に結合する化合物を BindingDatabase から取得し，すべての化合物の中から重複するものを Dsicovery Studio のプロトコルを用いて排除した．その結果，いずれかの標的タンパク質に結合する化合物を 785 個得た．そのため，標的タンパク質は平均で 26 個の結合する化合物を持っていた．本実験では各タンパク質において 26 個の結合する化合物を正事例にした．そして，それ以外の約 759 個の化合物を負事例とした．以上が，各タンパク質について処理を行う際のデータセットであった．表 1 に実験で使用したタンパク質と，それに結合する化合物数を示す．これらの化合物の多くは薬としての要件を満たしていた．そのため，ランダムに収集した化合物群と比較して，結合スコアが高くなりやすい．これは，スクリーニングにおける後々の段階に似ている．同じようなデータセットとして，a Directory of Useful Decoys(DUD)[28]がある．このデータには化合物と各タンパク質の結合エネルギーが含まれている．そのため，ドッキングソフトを用いずに本論文の機械学習手法を実行することが可能である．

表 1. 使用する標的タンパク質情報

ID	PDBID	ターゲット情報	リガンド数
1	1Q84	Acetylcholinesterase_(AChE)	16
2	1NDZ	Adenosine_Deaminase	17
3	1XS7	beta-Secretase_(BACE-1)	18
4	2UZT	cAMP-Dependent_Protein_Kinase_(PKA)	23
5	3B4F	Carbonic_anhydrase_I	28
6	2H15	Carbonic_anhydrase_II	41
7	1FVV	CDK2	20
8	2AYP	Checkpoint_Kinase_(Chk1)	32
9	1PKD	Cyclin-Dependent_Kinase_2_(CDK2)	83
10	3GHC	Dihydrofolate_Reductase_(DHFR)	14
11	2OQI	Dipeptidyl_Peptidase_IV_(DPP-IV)	33
12	2IOG	Estrogen_Receptor_(ER-alpha)	32
13	2GIU	Estrogen_Receptor_(ER-beta)	22
14	1LPG	Factor_Xa_(fXa)	62
15	3G2L	Glycogen_Phosphorylase_(PYGM)	13
16	2H55	Heat_Shock_Protein_90_(Hsp90)	39
17	1YT9	HIV-1_Protease	7
18	1C6Y	HIV-1_Protease_chain_A	4
19	2B6A	HIV-1_Reverse_Transcriptase	20
20	1KV2	MAP_Kinase_p38_alpha	37
21	2GG9	Methionine_Aminopeptidase_(MAP)	17
22	3JWS	Nitric_Oxide_Synthase,_brain	12
23	3E7T	Nitric_Oxide_Synthase,_inducible	19
24	1YHS	PIM-1_Kinase	13
25	2CMA	Protein-Tyrosine_Phosphatase_1B_(PTP1B)	34
26	1RNE	Renin	6
27	1O41	Src	18
28	2C8X	Thrombin	52
29	1O2Q	Trypsin	33
30	1VJA	Urokinase-type_plasminogen_activator	20

3.2 ソフトウェア

ここでは，本実験で用いたソフトウェアついて述べる．機械学習では LIBSVM[29]を用いた．本論文では LIBSVM に同梱されている JAVA コードを用いて，検証プログラムを作成した．このプログラムにより，各化合物がタンパク質に結合するかを分類する．また，SVR によって化合物の結合の強さを計算する．本論文では SVM Type として，分類のために C-SVC を，回帰のために epsilon-SVR を用いた．また，カーネルは RBF を使用した．実験から，他のカーネルでは有効な結果を得られなかった．SVM パラメータはいくつかのタンパク質を用いて調整し，C-SVC では-c 100 の時に多くのタンパク質および手法において良好な結果を得た．また，epsilon-SVR では-c 10 の時に良好な結果を得た．各手法の特徴量数が大きく異なるため，-g のパラメータはデフォルト値を用いた．

たんぱく質と化合物のドッキングスコアを計算するため，本論文では以下のソフトウェアを用いた．

・CDOCKER (Discovery Studio 2.5)[30]
・LibDock (Discovery Studio 2.5)[31]
・AutoDock Vina (1.1.1)[32]

CDOCKER と LibDock は 3D モデリングソフトである Discovery Studio[33]上で実行している．

これらのドッキングソフトは，タンパク質とリガンドの結合シミュレーションが可能であること，無料で使用できること，精度が高いこと，エラーの発生が低いことを考慮して選択した．ただし，Discovery Studio は我々の大学で購入しているものであり，学内以外での使用は有料である．

記述子の計算では Discovery Studio のプロトコルである Calculate Molecular Properties を用いた．これは，関連研究で用いられていた E-Dragon では有効な結果が得られなかったためである．そのため，関連研究の手法を本論文のデータに対して最適化した時と比較して性能が落ちる可能性がある．本論文では Calculate Molecular Properties で得られる記述子から，化合物間に違いの存在するものを選択した．また，前述のデータセットからいくつかのタンパク質を選び，特徴量選択に使用した．本論文では hbackward stepwise selection[34]によって不要な記述子を排除した．その結果，improper energy，原子数，ALogP，極表面積の 4 種を得た．また，785 個の化合物に含まれる 11 種類の原子について，化合物内の原子数を特徴量とした．本論文では以上の 15 種の特徴量を用いた．

以上のソフトウェアによって得られたベクトルデータを用いて提案手法を実行し，その性能を調べる．その際，特徴量としてドッキングスコアのみを使ったとき(MTS)，記述子のみを使用した時(Descriptor)，そしてすべての特徴量を使用した時(MTS+Desc)について調べる．

3.3 実験方法

本実験では，Leave one out cross validation に基づく性能評価を行う．そのため，あるタンパク質において，一つの化合物を取り出し，それ以外の化合物を用いて学習を行う．そして，学習によって得られたモデルによって一つの化合物を分類および回帰によるスコア付けを行う．この手法によって得られたラベルを用いて性能を評価する．評価付けでは情報検索で用いられている手法を用いる．

分類では正しく分類できているかによって評価する．本論文では正しく分類された場合を True，誤って分類された場合を False とする．また，分類の結果標的タンパク質に結合する化合物を Positive，結合しない化合物を Negative とする．例えば，タンパク質に結合する化合物を正しく分類した場合は TP である．逆にタンパク質に結合する化合物を誤って分類した場合は FN である．このとき，適合率(Precision)は TP/(TP+FP)，再現度(Recall)は TP/(TP+FN)，特異度 (specificity) は TN/(TN+FP)，正確度 (accuracy) は (TP+TN)/(TP+TN+FP+FN)とする．本実験では正事例と比べて負事例が非常に多いため，specificity や accuracy は高くなる．そのため，高い accuracy だけでなく，高い適合率や感度をもつ手法は，性能が高い．これらの値はドッキングソフトと比較することはできないが，序章で述べたように，ドッキングソフトでは適合率が非常に低くなることがわかっている．

回帰ではスクリーニング性能を評価する．本論文では Receiver Operating Characteristic (ROC)曲線に基づく評価を行う．ROC 曲線は医学や生物学で用いられるが，2 値分類や検索においても使用できる．ROC 曲線は検出性能を示す．ROC 曲線では検出するかどうかの閾値を変化させ，それぞれの閾値における TP の数を y 座標，FP の数を x 座標とする．そして，それらの座標を持つ点をグラフ上に書くことで，曲線を作ることができる．本論文では座標は数ではなく，最大数を 1 としたときの比率で表す．ある閾値で分類した時，TP が多く，FP が少ないほうが，性能が良いといえる．そのため，グラフが左上に寄るほど，性能が良いとい．逆に，グラフが左下と右上を結ぶ対角線に近いときは，データをランダムに並べた時と変わらないため，性能が低い．この対角線をランダムラインと呼ぶ．また，このグラフの下の面積は area under curve(AUC) という．この面積は 0~1 で表される．最も性能が良いときは 1 であり，ランダムラインでは 0.5 となる．本論文では各手法の AUC 値を比較することで，性能を評価する．

4. 実験結果

表 2 は分類の実験の結果を示す．各行は各手法の判別性能を示す．表の値はすべて各タンパク質について数値を出し，30 個のタンパク質の平均をとった値である．TP Average は，各タンパク質について SVM による判別を行った際に，検出した正事例及び負事例の化合物数を示す．FP Average は負事例を誤って検出した数を示す．適合率は検出した化合物の中の正事例数の割合，再現率は全正事例の中で検出した正事例の割合，特異度は全負事例の中で正しく負と判別した割合，正確度は全化合物の中で正しく判別した割合を示す．

本研究の手法(MTS および MTS ＋ Desc)では約 1/3 の正事例を検出でき，正と判別した化合物のうち半数は正事例であった．これらの値はドッキングソフトで得られる結合スコアに，任意の判別基準を与えて判別を行った際と比べて非常に高い．また，MTS は関連研究の Descriptor と比較して同等以上の性能を持つことがわかる．これらを組み合わせた MTS ＋ Desc は最も性能がよく，手法の統合による性能向上ができていることがわかる．ただし，ドッキングソフトよりも性能は向上しているが，機械学習による判別という観点では性能は低い．原因の一つは，正事例と負事

例の割合が大きく異なることである．本実験では正事例と負事例の各平均数の比率は約 29.2 倍である．この比率を改善する必要があるが，正事例を増加させるのは困難であり，負事例の減少は特異度の減少を引き起こし適合率の低下につながる．そのため，今後は各タンパク質に合わせた負事例の選択が必要となる

表 3 は回帰の実験の結果を示す．各行は各手法の検出性能を示す．下 3 行は，分類で用いた 3 手法のほかに，各ドッキングソフトの結果を示す．AUC_μ は 30 タンパク質での AUC の平均値を示す．AUC_σ は 30 タンパク質での AUC の標準偏差を示す．AUC 値においては本研究の手法である MTS や MTS + Desc の値が高く，関連研究の Desc よりも高いことがわかる．また，これらの値は各ドッキングソフトを用いた時と比べても非常に高い．そのため，本研究の手法により結合スコアを基に精度を向上させた回帰スコアが得られていることがわかる．

図 3 は 30 個のタンパク質の中で，得られる AUC 値が表 3 の平均 AUC 値に最も近かった事例(PDBID:2C8X)における ROC 曲線である．平均値との距離は，表 3 における 6 種類の特徴量・ドッキングソフトそれぞれにおける平均 AUC 値と，タンパク質から得られる 6 種類それぞれの AUC 値との差をとり，値の絶対値を合計することで得た．以上より，図 3 の各 ROC 曲線は，本実験で用いた 30 個のタンパク質の中で最も平均的な曲線と言える．図 3 では SVM による回帰計算でスコアを得た曲線は各ドッキングソフトと比べて左上に寄っており，精度が大きく改善されていることがわかる．これはドッキングソフトの精度がよく，AUC 値が高い場合でも同様である．また，本研究の提案する MTS による特徴量抽出は関連研究の Descriptor と全く異なる数値を用いているものの，精度が優れている．以上より本研究の SVM の回帰計算による結合スコアの改善が有効であることが確認できる．

図 4 は実験において SVM による判別が機能せず，正事例が一つも検出できなかった事例(PDBID: 3B4F)における ROC 曲線である．この事例では各ドッキングソフトの出力する結合スコアの精度は非常に低い．また，SVM による判別では負事例がいくつか正として検出されるのみであった．これに対し，本研究の手法で回帰スコアを各化合物に与えた時では ROC 曲線は左上に寄っており，スコア付として有効であるといえる．そのため，SVM による判別ができない事例については回帰スコアを利用してスコア精度を高めることで，研究者によるグラフからの判別基準作成に貢献することができる．

表 2. 判別実験結果

	TP	FP	適合率	再現率	特異度	正確度
特徴量方式	average	average	Precision	Recall	Specificity	Accuracy
MTS	9.37	8.73	0.546	0.331	0.988	0.967
Descriptor	9.03	11.57	0.452	0.330	0.984	0.963
MTS+Desc	10.83	8.17	0.586	0.379	0.989	0.970

表 3. 回帰計算結果

Method	AUC_μ	AUC_σ
MTS	0.894	0.071
Descriptor	0.749	0.168
MTS+Desc	0.882	0.103
LibDock	0.536	0.188
AutoDock Vina	0.596	0.167
CDOCEKR	0.544	0.2

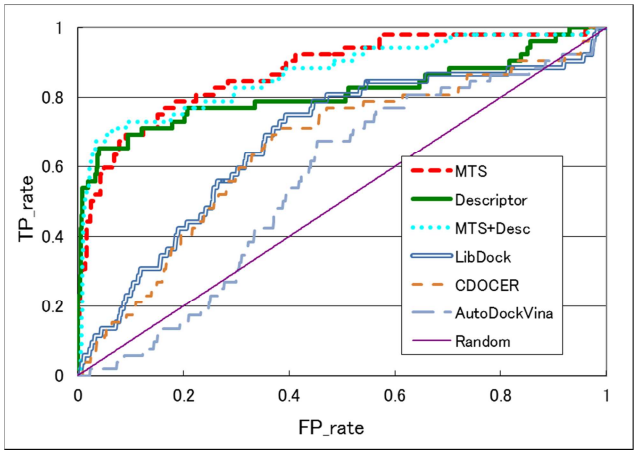

図 3.代表的な ROC 曲線

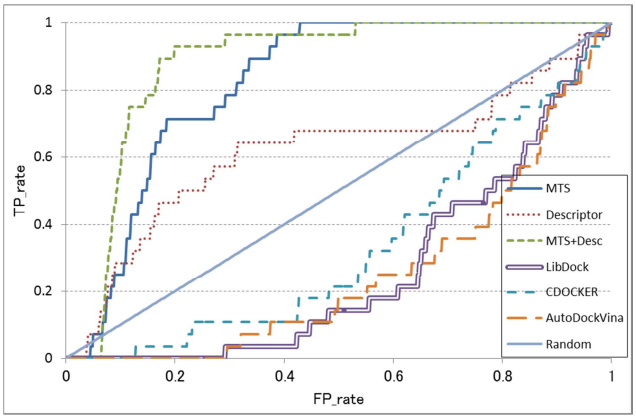

図 4 判別不能時の ROC 曲線

5. おわりに

本論文ではドッキングソフトの出力するスコアを用いて機械学習を行い，化合物が標的タンパク質に結合するかどうかを分類する手法を提案した．本論文では従来性能や情報の量から機械学習に用いられていなかった結合スコアに焦点を当て，多数の結合スコアを利用することにより機械学習に用いることを可能にした．そして，結合スコアを特徴量として SVM を用いた学習により，結合するかどうかの分類と，結合の強さのスコア付けを可能にした．実験から，本論文の手法によって結合スコアの性能を向上できることを示した．また，従来手法と比較して同等の性能があるこ

とを示した．以上より，結合スコアは機械学習における特徴量として有用であることが確かめられた．また，本論文の手法と従来手法を併用することで，より性能が高まることが確かめられた．

参考文献

[1] Cheng,T., “Comparative assessment of scoring functions on a diverse test set” J. Chem. Inf. Model., 49, 1079-1093. (2009).

[2] Wang,R., “Comparative evaluation of 11 scoring functions for molecular docking” J. Med. Chem., 46, 2287-2303. (2003).

[3] Borman S., “Drugs by design” Chem Eng News 83: 28-30. (2005).

[4] Okamoto, M., Takayama, K., Shimizu, T., Muroya, A., Furuya, T., “Structure-activity relationship of novel DAPK inhibitors identified by structure-based virtual screening”, Bioorg. Med. Chem. 2010, 18, 2728-2734. (2010).

[5] Leach,A.R., “Prediction of protein-ligand interactions. docking and scoring: successes and gaps” J. Med. Chem., 49, 5851-5855. (2006).

[6] Moitessier,N., “Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go” Br. J. Pharmacol., 153, S7-S26. (2008).

[7] Guvench,O. and MacKerell,A.D., “Computational evaluation of protein-small molecule binding. Curr. Opin. Struct. Biol., 19, 56-61. (2009).

[8] Waszkowycz B., Clark DE., Gancia E., “Outstanding challenges in protein-ligand docking and structure-based virtual screening” Wiley Interdiscip Rev Comput Mol Sci 1: 229-259. (2011).

[9] Kengo Hanaya, Miho Suetsugu, Shinya Saijo, Ichiro Yamato, and Shin Aoki, “Potent Inhibition of Dinuclear Zinc(II) Peptidase, an Aminopeptidase from Aeromonas proteolytica, by 8-Quinolinol Derivatives: Inhibitor Design Based on Zn2+ Fluorophores, Kinetic, and X-ray Crystallographic Study” Journal of Biological Inorganic Chemistry, April 2012, Volume 17, Issue 4, pp 517-529. (2012).

[10] Kitchen D.B., “Docking and scoring in virtual screening for drug discovery: methods and applications” Nat. Rev. Drug Discov., 3, 935-949. (2004).

[11] D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B.P. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvary, K.F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman: "AMBER 11" University of California, San Francisco. (2010).

[12] Brown SP, Muchmore SW., “Rapid estimation of relative protein-ligand binding affinities using a highthroughput version of MM-PBSA” J Chem Inf Model 2007, 47:1493-1503. (2007).

[13] Charifson P. S., Corkery J. J., Murcko M. A., Walters P., “Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins”, J. Med. Chem, 1999, 42, 5100-5109. (1999).

[14] Jorissen R.N., Gilson M.K., “Virtual screening of molecular databases using a support vector machine”, J. Chem. Inf. Model. 2005, 45, 549-561. (2005).

[15] Corwin Hansch, Albert Leo, "Exploring QSAR, Fundamentals and Applications in Chemistry and Biology" American Chemical Society, (1995).

[16] Todeschini, R. "Consonni, V.; Pavan, M. DRAGON 2.1. Milano Chemometrics and QSAR" Research Group: Milan, Italy. (2002).

[17] Vladimir N. Vapnik, “The Nature of Statistical Learning Theory”, Springer-Verlag. (1995).

[18] Deng,W., "Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods" J. Chem. Inf. Comput. Sci., 44, 699-703. (2004).

[19] Ballester, P. J., "Mitchell, J. B. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking" Bioinformatics 2010, 26, 1169-1175. (2010).

[20] Sato, T., "Honma, T.; Yokoyama, S. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening" J. Chem. Inf. Model. 2010, 50, 170-185. (2010).

[21] Springer, C., Adalsteinsson, H., Young, M. M., Kegelmeyer, P. W., Roe, D. C. Postdock, "A structural, empirical approach to scoring protein ligand complexes" J. Med. Chem. 2005, 48, 6821-6831. (2005).

[22] Sarah L. Kinnings, Nina Liu, Peter J. Tonge, Richard M Jackson, Lei Xie, and Philip E. Bourne, “A Machine Learning-Based Method To Improve Docking Scoring Functions and Its Application to Drug Repurposing”, Journal of Chemical Information and Modeling, Volume: 51, Issue: 2, Pages: 408-419. (2011).

[23] Yiming Yang and Xin Liu., "A Re-examination of Text Categorization Methods" In Proceedings of the Twenty-Second International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 42-49. (1999).

[24] Thorsten Joachims. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features" In Proceedings of the Tenth European Conference on Machine Learning (ECML), pages 137-142, Berlin, Germany. (1998).

[25] Omagari K., Mitomo D., Kubota S., Nakamura H., Fukunishi Y., "A method to enhance the hit ratio by a combination of structure-based drug screening and ligand-based screening" AdV. Appl. Bioinf. Chem, 2008, 1, 19-28. (2009).

[26] Leo Breiman, "Bagging Predictors" Machine Learning, vol.24, pp.123-140. (1996).

[27] Yoav Freund, and Robert E. Schapire, "Experiments with a New Boosting Algorithm" Proc. of The 13th Int'l Conf. on Machine Learning, pp.148-156. (1996).

[28] Huang, "Shoichet and Irwin,Benchmarking Sets for Molecular Docking" J. Med. Chem., 2006, 49(23), 6789-6801. doi 10.1021/jm0608356. (2006).

[29] Chih-Chung Chang, Chih-Jen Lin, "LIBSVM : a library for support vector machines." ACM Transactions on Intelligent Systems and Technology, 2011, 2:27:1--27:27. (2011).

[30] Wu G, Robertson DH, Brooks III CL, Vieth M., "Detailed analysis of grid-based molecular docking: a case study of CDOCKER-a CHARMm-based MD docking algorithm" J Comput Chem 24: 1549?1562. (2003).

[31] Diller DJ, Merz Jr KM, "High throughput docking for library design and library prioritization" Proteins Struct Funct Genet 43: 113-124. (2001).

[32] Trott O., Olson A. J., "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading." J. Comput. Chem, 31, 455-461. (2010).

[33] Discovery Studio, Accelrys Inc., San Diego, CA 92121, U.S.A. (2008).

[34] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” Journal of Machine Learning Research, vol. 3, pp. 1289-1305. (2003).