3B2-R-2-5

SVM を用いたインシリコ創薬における新薬の有効性判別

Binary classification of compounds using SVM for in silico drug design

岡田 正人 Masato Okada
大和田 勇人 Hayato Ohwada

東京理科大学大学院理工学研究科
Tokyo University of Science, Faculty of Science and Technology

This paper proposes a new approach to docking scoring of a chemical compound for drug design virtual screening. There are a number of useful docking softwares that take a real number as a docking score, but it is still hard to say that biologists could use such scores in realistic experiments due to its predictive inaccuracy. In contrast, our approach takes a binary scoring that indicates whether a candidate compound is docked to a target protein or not. This leads to an automatic screening of compounds without the load of biologists in drug design processes. In this paper, we show several implementations of the method based on Support Vector Classification. The experiment shows the classification accuracy is 97%, the sensitivity is 38%, and the AUC value is 0.88.

1. はじめに

創薬研究をコンピュータ上で行うインシリコ創薬では，化合物がタンパク質と結合するかどうかを予測するためにドッキングソフトが用いられている．ドッキングソフトはタンパク質に対する化合物の結合様式を予測し，その時の結合の強さを実数の結合スコアとして出力する．しかし，予測精度が十分ではなく，結合する化合物と結合しない化合物を明確に判別できないことから，創薬研究者がそれらのスコアを現実の結合実験と同様に扱うことは難しい．

Charifson などは計算方式の異なる複数のドッキングソフトのスコアを利用し，スコア精度の向上を行っている[Charifson 1999][Okada 2011]．また，Jorissen は，化合物の情報(構造や力場等)を学習し，回帰計算によって新たなスコアを生成している[Jorissen 2005]．以上のような研究により，ドッキングソフトのシミュレーションや結合判別の精度が向上している．しかし，ドッキングソフトが出力する結合スコアから結合の判別を行うことのできる研究は少ない．

本研究では以上を踏まえ，化合物がタンパク質に結合するかどうかを示す二項スコア付けを行う．本研究では従来判別に用いることが困難であった結合スコアから，化合物の判別を行うためのモデルを作成する．その際，スコア精度の改善のために用いられている考え方に基づいて結合スコアを拡張することで，モデル作成を可能にする．また，結合スコア以外の情報として，関連研究で用いられている化合物の情報を取り入れることで，関連研究との比較を行うとともに，精度の向上を図る．さらに，作成したモデルによっての判別が困難な事例について，本研究の手法やデータを利用した確率推定を行うことで，化合物の結合の強さを新たに出力する．これにより，機械学習を利用したスコア改善が可能となる．以上の手法を実装し，実在するタンパク質，およびそれに対する結合情報をもつ化合物を利用した精度評価実験を行う．

2. 結合スコアに基づく特徴量

ここでは，判別を行うモデルを作成する際に学習器に与える，化合物の特徴量について述べる．まず，結合スコアの定義について述べる．次に，結合スコアの拡張について述べる．最後に，関連研究に基づいた特徴量について述べる．

2.1 結合スコアの定義

ドッキングソフトが出力するスコアを以下のように定義する．計算対象のタンパク質を p ，化合物を c とするとき，ターゲットタンパクと化合物の関係を表すスコア S は次式，

$$S = f(p, c) \tag{1}$$

で示される．関数 $f(p, c)$ は各ドッキングソフト固有の計算関数である．このとき，スコア s は実数であり，値が大きいほどタンパク質 p と化合物 c の結合力が高いことを示す．前述のようにスコアの精度に問題があり，結合スコアから結合の可否を判別することは困難である．

2.2 結合スコアの拡張

本研究では前述の結合スコアを拡張し，化合物とタンパク質の関係を示す複数の特徴量を得る．以下の手法によって特徴量次元数を増加させることで，従来困難であった結合スコアからのモデル作成が可能となる．

第一にドッキングソフトに基づく結合スコアの拡張を行う．本研究では独立した複数のドッキングソフトによって結合スコアを計算し，それらの値を特徴量として用いる．これらの値は多少の相関は存在するものの独立している．この手法により，タンパク質と化合物との間の結合力を多面的にとらえ，特徴量次元数を増大させることができる．また，Charifson などの手法と同様に，ドッキングソフト間でのスコア傾向を利用することができる．

第二にタンパク質に基づく結合スコアの拡張を行う．本研究では研究対象のタンパク質のみならず，その他のタンパク質について化合物との結合スコアを計算し，それらの値を特徴量として用いる．各タンパク質は基本的に独立しているものの，似た構造をもつタンパク質や，化合物が結合する位置が異なるだけで全体のタンパク質は同一のものも存在する．この手法により研究対象に特に高いスコアを出す性質や，構造の似たタンパク質間でのスコアの差などの性質，傾向を利用できる[Omagari 2008]．

以上の拡張を行うことで，化合物が持つ結合スコアを特徴量として用いる際に，特徴量次元数を増加させることができる．このとき，一つ一つの特徴量は次のようにあらわされる．研究対象を含めた q 個のたんぱく質の集合を $P = \{P_1, P_2, \cdots P_q\}$ とするとき，i 番のドッキングソフトによって得られる，j 番の化合物 C_j と k 番

連絡先：岡田正人，東京理科大学，大和田研究室，okada@ohwada-lab.net

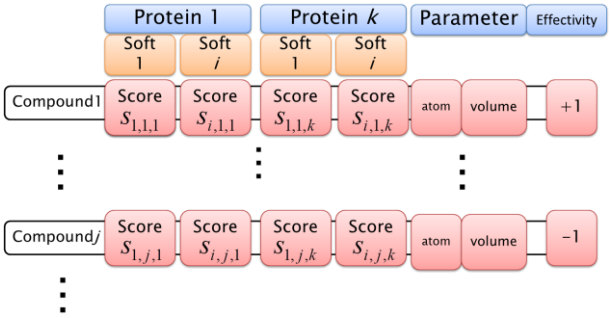

図 1. 特徴量行列

のタンパク質 P_k の関係を示す結合スコアは $S_{i,j,k}$ となる. そのため, 特徴量の集合は, ドッキングソフト数 n 個, 化合物数 m 個, たんぱく質数 q 個のとき, m 行×(n×q) 列の特徴量行列となる. このように, 化合物と複数のタンパク質について, 複数のドッキングソフトで得られる結合スコアを用いる手法を, 本研究ではマルチターゲットスコアリング(MTS)と呼称する.

2.3 化合物パラメータ

化合物パラメータは各化合物そのものの特徴を示すもので, 化合物内の各原子の数や体積, 疎水性などで構成される. これらは化合物に対して各種演算を行うことで得られるため, 特徴抽出時にタンパク質は関係しない. 化合物パラメータは化合物とタンパク質の結合判別にも用いられており, 本研究と同様に機械学習における特徴量として利用している研究が存在する[Hanna 2008][Sarah 2011]. しかし, 本研究のようにタンパク質と化合物の関係を示す数値を用いて機械学習を行う研究は少ない.

本研究ではこのような化合物パラメータを化合物の特徴量として用いる. これは, MTS という独立した手法と統合することによる性能向上を行うとともに, MTS で得られる特徴量のみで判別を行った時の判別精度と化合物パラメータによる特徴量のみでの判別精度を比較し, MTS と従来手法との性能比較を行うためである.

化合物パラメータを特徴量として用い, MTS で得られる特徴量に追加したときの特徴量集合は以下のようになる. ドッキングソフト数 n 個, 化合物数 m 個, タンパク質数 q 個, 加えて化合物パラメータを r 種類用いるとき, m 行×(n×q +r)列の特徴量行列となる.

これらの特徴量に, 化合物が研究対象のタンパク質に対して結合するかどうかの既知情報を加えることで, 機械学習のトレーニングデータとして適用可能な特徴量行列を得ることができる. この特徴量行列を図 1 に示す. 本研究では図 1 に示した特徴量行列を学習器にかけることで, 化合物がタンパク質に結合するかどうかを判別する判別モデルを構築する. そして, タンパク質に結合するかどうかがわかっていない, 未判別化合物に対しても同様の特徴量を生成し, 判別モデルを用いて結合の可否を判別する.

3. 機械学習

本研究ではタンパク質と化合物の結合を判別するため, 機械学習の中でもパターン認識能力に優れる SVM(Support Vector Machine)を用いる[Vladimir 1995]. SVM は教師あり機械学習手法であり, 正答を含むデータを用いて学習することで, 判別を行うための判別モデルが構築できる. SVM は, 特徴量の次元数が増加するとともに判別性能が落ちる"次元の呪い"に強く, 本研究のようにより多くのパラメータを用いて判別を行うという場合に適している.

本研究では図 1 の特徴量データを SVM に入力し, 判別モデルを構築する. このとき, 正事例をタンパク質に結合する化合物, 負事例をタンパク質に結合しない化合物とする. そして, 判別モデルによって未判別化合物の判別を行い, その結果正事例に属する化合物はタンパク質に結合し, 負事例に属する化合物はタンパク質に結合しないと推定する.

以上は SVM を用いる際の基本的な手法であり, 判別性能を向上させるための各種学習手法が利用可能である. 本研究を進めるに当たり, バギング[Leo 1996]やブースティング[Yoav 1996]といった性能向上手法を試行しているが, 後述の実験を行った際に結合判別性能の向上がほとんど見られなかった. そのため, 本研究では計算時間を考慮し, SVM の通常使用によって判別を行うこととした.

ここまで述べた手法により, 機械学習による未判別化合物の判別が可能となるが, 事例(タンパク質)によってはほとんどの化合物が結合しないと判別されるような場合が存在する. 本研究では判別不可能なタンパク質においても機械学習による創薬支援が行えるよう, 確率推定を利用したスコア改善を行う. 確率推定は SVM による判別時に, 判別対象のデータが各クラスに属する確率を推定したものであり, 属する確率が最も高いクラスに判別される. 確率推定の値は 0～1 であらわされる. また, この値を利用することで, 受信者操作特性(Receiver Operating Characteristic, ROC)曲線を作成できる.

本研究では確率推定によって得られる値を化合物の新しいスコア(確率推定スコア)ととらえる. これにより, MTS により得られた多数の特徴量を, 機械学習によって統合することができる. 従来の結合スコア改善手法では各改善手法を統合するためには個別の研究と最適化が必要であったが, 本手法により容易な統合が可能である.

確率推定スコアには 0.5 という判別基準が存在するものの, 未判別データの確率推定スコアのみでは結合スコアとしての値の位置づけを見出すことができない. そのため, 結合情報が存在する学習データについても leave-one-out 法で確率推定スコアを付けることで, 未判別データのスコアの位置づけを見出す. 以上の手法により, 判別が困難な事例に対して, データの追加変換等をせずに, 機械学習を用いたスコア改善として本研究で得られる特徴量行列を利用できる.

4. 実験

ここでは, タンパク質と化合物の結合スコアを基に未判別化合物の結合可否を判別する本研究の手法を, 実在するタンパク質および化合物のデータに対して適用し, 判別性能や判別精度, スコア改善性能の評価を行う. 特に, 各タンパク質に対応して存在する, タンパク質に結合する化合物(薬等)の検出能力により, 判別性能を示す.

4.1 実験環境

実験において使用したコンピュータ環境を表 1 に示す. 本研究では SVM として LibSVM を用い, その Java コードを用いて実装する. 使用する SVM タイプは C-SVC である. SVM パラメータではコストパラメータとして-c 100 を使用する. この値は各特徴量行列(MTS やパラメータ)を用いて判別を行う際に SVM パラメータの値を変化させて精度を検証し, F 値が最も高くなるときのパラメータ値であった. また, 確率推定計算を実行させるため, 学習と判別の両方において –b 1 を用いている.

表 1. 実験環境

CPU	Xeon 5520 2.26GHz 4コア×2CPU
メモリ	48GB
OS	Windows7 professional 64bit
SVM	LibSVM3.11 (JAVA)
JAVA	JRE6

表 2. 実験結果 判別性能

特徴量方式	正事例を正と判別	負事例を正と判別	適合率 (precision)	感度 (sensitivity)	特異度 (specificity)	正確度 (accuracy)
MTS	9.37	8.73	0.546	0.331	0.988	0.967
Parameter	9.03	11.57	0.452	0.330	0.984	0.963
MTS+Param	10.83	8.17	0.586	0.379	0.989	0.970

また，たんぱく質と化合物の結合スコアを求めるため，以下の 3 種類のドッキングソフトを用いた．

・CDOCKER (Discovery Studio 2.5)
・LibDock (Discovery Studio 2.5)
・AutoDock Vina (1.1.1)

化合物パラメータ抽出では，3D モデリングソフトである Discovery Studio のプロトコルである Calculate Moleculer Properties を用いた．本研究では得られる各特徴量のうち，使用時に F 値が高くなるものを選択した．その結果，不正エネルギー，原子数，ALogP，極表面積の 4 種と各原子の数 11 種の計 15 種の特徴量を用いた．これらのパラメータを使用した時の F 値をいくつかのタンパク質で調べたところ，関連研究に記載されているパラメータを使用した時と同等であった．ただし，関連研究では多数(数百)のパラメータを使用していたため，計算時間低減のためこちらを用いている．

4.2 実験データ

本実験では実際の創薬環境に近づけるため，タンパク質と化合物の結合情報を有する BindingDatabase のデータを用いた．実験においてはデータベースに存在するタンパク質の中から，結合する化合物(正事例化合物)が多いものから順に選び，30 個のタンパク質を使用した．そして，各タンパク質に結合する化合物を収集し，計 785 個の正事例化合物を得た．ただし，785 個の中で構造が近似している化合物は取り除いている．以上のタンパク質と化合物の集合を実験に用いる．

本実験では，本研究の手法の判別性能を 30 個のタンパク質それぞれについて調べる．このとき，各タンパク質における負事例化合物として，785 個の化合物から正事例化合物を取り除いたものを使用する．そのため，各タンパク質は平均で約 26 個の正事例を有し，約 759 個の負事例を有する．以上の化合物はすべていずれかのタンパク質に結合するため，薬物としての有効性(薬効性)が高い．そのため，この環境はインシリコスクリーニングにおける候補絞り込みの最終段階に似ている．

4.3 実験方法

本実験では本研究の判別性能を Leave One Out 法によって調べる．そのため，各化合物について判別を行い，結果が結合情報と一致するかによって判別が正しいかを調べる．また，客観的な評価指標の作成と，化合物に新しいスコア(確率推定スコア)を与えるため，判別時に各化合物の確率推定値を計算する．そして，受信者操作特性(ROC)曲線を作成し，グラフや曲線下部面積を用いて評価を行う．これらの手法を用いて実験を行った時の所要時間は，特徴量が MTS とパラメータすべてを用いた時にタンパク質一つに対して平均で 50.4 秒であった．処理は Java 上で並列化させている．

表 3. 実験結果 AUC 値

Method	AUC_μ	AUC_σ
MTS	0.894	0.071
Parameter	0.749	0.168
MTS+Param	0.882	0.103
LibDock	0.536	0.188
AutoDock Vina	0.596	0.167
CDOCEKR	0.544	0.200

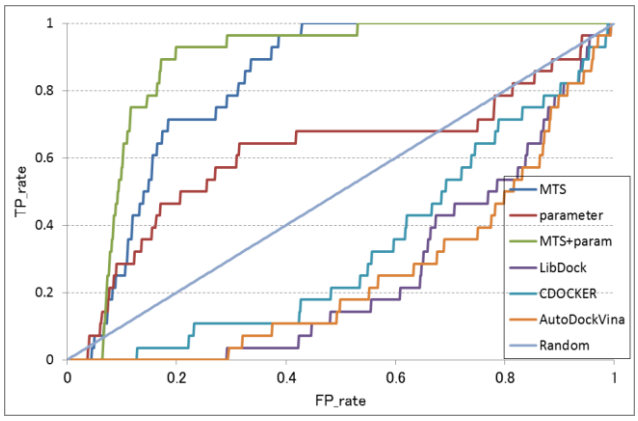

図 2. 判別不能時の ROC 曲線

4.4 実験結果

表 2 に SVM による結合判別の結果を示す．表の数値は 30 個のタンパク質で得られた数値の平均をとったものである．表中の MTS 行は特徴量に結合スコアを拡張し得られた値を用いたもの，Parameter は関連研究と同様に化合物パラメータを用いたもの，MTS＋Param は両方の特徴量を用いたものである．

表より，本研究の手法(MTS および MTS＋Param)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 Parameter 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＋Param 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．

本実験では正事例と比較して負事例の数が多く，結果として特異度や正確度は非常に高くなっている．そのため，負事例の排除には有効であるものの，多くの正事例を検出し感度を高めるためには他の手法が必要である．

表 3 は確率推定で得られた確率推定スコアを利用し，ROC 曲線を作成，曲線の下部面積である AUC 値を計算したものである．AUC 値においては本研究の手法である MTS や MTS+Param の値が高く，関連研究の Parameter 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確率推定スコアが得られていることがわかる．

図 2 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，正事例化合物の確率推定スコアも 0.5 より低く，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，本研究の手法で確率推定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確率推定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．

5. おわりに

本稿ではドッキングソフトによって得られたタンパク質と化合物結合スコアを SVM の入力として与え，化合物がタンパク質に結合するかどうかを判別する手法について述べた．そして，機械学習の手法を用いることで，結合スコアの改善手法や化合物パラメータなどの情報を特別なアルゴリズムなしに統合し，判別モデル作成に利用することが可能であることを示した．そして，実験を通じて本研究の手法による判別が有効であり，関連研究やドッキングソフトと比較して高い性能を有することがわかった．また，確率推定スコアの計算を通じて，本研究の手法が AUC 値においても他の手法より優れていること，判別が不可能な事例についてもスコア改善手法として利用できることを示した．

以上のように本研究の手法が有効であることを示したが，本研究では多数の結合スコアが必要であり，特徴量抽出に多くの計算時間が必要となる．そのため，本研究の手法は，薬の候補化合物を薬効性などを用いて選別した後の，インシリコスクリーニングにおける最終段階に用いることが有効である．これにより，生物学者の創薬手順を経ることなしに，高い精度で化合物の自動スクリーニングを行うことが可能となる．

参考文献

[Charifson 1999] 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.

[Vladimir 1995] Vladimir N. Vapnik: The Nature of Statistical Learning Theory, Springer-Verlag, 1995.

[Jorissen 2005] Jorissen R.N., Gilson M.K.: Virtual screening of molecular databases using a support vector machine, J. Chem. Inf. Model. 2005, 45, 549–561.

[Okada 2011] Okada M., Tsukamoto M., Ohwada H., Aoki S.: Consensus Scoring to Improve the Predictive Power of in-silico Screening for Drug Design, Proc. of the 2nd International Conference on Engineering and Meta-Engineering, 2011, 3, 94-98.

[Omagari 2008] 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.

[Hanna 2008] Hanna Geppert, Tamás Horváth, Thomas Gärtner, Stefan Wrobel, and Jürgen Bajorath: Support-Vector-Machine-Based Ranking Significantly Improves the Effectiveness of Similarity Searching Using 2D Fingerprints and Multiple Reference Compounds, J Chem Inf Model. 2008 Apr;48(4):742-6.

[Sarah 2011] 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.

[Leo 1996] Leo Breiman: Bagging Predictors, Machine Learning, vol.24, pp.123-140, 1996.

[Yoav 1996] 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.