1D5-4in

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

The Virtual Random Screening Method for Drug Discovery

岡田 正人[*1] 金盛 克俊[*2] 青木 伸[*3] 大和田 勇人[*2]
Masato Okada Katsutoshi Kanamori Shin Aoki Hayato Ohwada

[*1] 東京理科大学理工学研究科 [*2] 東京理科大学理工学部 [*3] 東京理科大学薬学部
[*1] [*2]Tokyo University of Science, Faculty of Science and Technology [*3]Tokyo University of Science, Faculty of Pharmaceutical Sciences

This paper proposes a new approach to classification of docking between compounds and proteins for drug design virtual screening. Currently, docking software programs often use real numbers as docking scores; but due to their predictive accuracy, it is difficult for biologists to use such scores in realistic experiments. In contrast, our approach utilizes binary classification that indicates whether a candidate compound is docked to a target protein. This leads to automatic screening of compounds without the input of biologists in drug design. The present method provides consensus use of the scores of existing docking software, yielding higher accuracy of binary classification. In this paper, we discuss several implementations of the proposed method based on Support Vector Classification and Regression. We have created a classification model from docking scores and chemical information. The experiment demonstrates that our method outperforms the existing docking softwares in classification.

1. はじめに

創薬研究ではドッキングソフトと呼ばれるたんぱく質と化合物の結合シミュレーションソフトが良く使われている[Trott 2010] [Rao 2007] 。このソフトは、たんぱく質と化合物に含まれる多数の原子が関わる複雑な力学計算を行うことができ、実際の結合実験時に非常に近い結合ポーズを出力することができる[Diller 2001] 。また、結合ポーズにおける力場からたんぱく質と化合物の相互作用を計算し、全体の結合の強さをドッキングスコアとして出力することができる[Cheng 2009] [Moitessier 2008][Wang 2003]。これらのソフトや、その出力結果を用いて創薬研究が進められており、いくつかの薬が発明されている[Borman 2005] [Okamoto 2010]。しかし、ドッキングスコアが高くとも結合しない化合物は多く存在する[Leach 2006]。また、結合の可否を判定することができず、こちらで基準を用意する必要がある。

本研究ではこれに対し、機械学習を用いた化合物の結合判定手法を提案する。本研究ではたんぱく質との結合の強さが知られている化合物を用い、結合の強い化合物と弱い化合物の特徴を学習する。そして、新しい化合物の結合を判定するための結合判定モデルを作成する。このモデルにより、新しい化合物がたんぱく質に結合する可能性を示す結合可能性スコアを求めることができる。

従来の研究では、機械学習を行う際に、表面積や原子間距離などの、化合物そのものの情報を用いて機械学習を行っていた[Deng 2004] [Jorissen 2005]。これに対し、本研究では化合物とたんぱく質の関係を機械学習に用いる。本研究では上記のドッキングスコアを学習に用いることによって、化合物とたんぱく質の力学的関係や構造的関係を学習することができる。

また、本研究では機械学習によって結合可能性スコアを出力する。このスコアには結合を判定するための基準があり、かつその値の中で結合可能性の高低を比較することができる。本研究の手法により結合可能性スコアを計算することで、化合物の結合判定を自動的に行うことができ、創薬研究において新規化合物のスクリーニングが可能となる。

以上の手法を用い、実際の創薬研究に近い環境での実験を行うことにより、本研究の創薬支援における有用性を示す。

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

2. 機械学習

本研究では図 1 に示すような機械学習を行う。入力として、たんぱく質に対して結合力のわかっている化合物を用いる。結合力は結合実験等によって得られる実際のデータであり、Chembl [Gaulton 2012] 等の活性データベースから得ることができる。結合力の強弱の境界は定まっていないが、リガンドデータベース DUDE [Mysinger 2012]では、Ki 値で 1 μ M を境界にしている。本研究ではこれらの化合物と、その化合物の各種情報を用いて機械学習を行う。このとき、高結合力化合物と低結合力化合物の特徴から、結合するかどうかを判定する関数を回帰計算によって求める。この関数を結合判定モデルと呼ぶ。結合判定モデルを用いることにより、新しい化合物について、その化合物がたんぱく質に結合する可能性を示す結合可能性スコアを計算することができる。結合可能性スコアは 0 から 1 の実数値で表され、0.5 よりも大きいならば、化合物は結合すると判定される。以上の手法により、既存の化合物の情報から機械学習を行い、新しい化合物がたんぱく質に結合するかどうかを判定できる。

本研究では、機械学習として SVM [Vapnik 1995]を用いる。SVM は教師あり学習を用いる判別手法の一つで、未学習データに対する判別制度が高い。また、次元の呪いに強く、多数の属性を用いて機械学習を行う本研究の手法に適している。さらに、SVM は回帰計算を行うことが可能であり。本研究で行う結合可能性スコア計算を実行可能である。

3. 機械学習に用いるデータ

本研究で行う機械学習では、以下のデータを用いて学習と判定を行う。

3.1 ドッキングスコア

前述のように、ドッキングスコアはたんぱく質と化合物の結合の強さを、複雑な力学計算によって数値化したものであり、たんぱく質と化合物との関係を示す情報として非常に重要である。そのため、ドッキングスコアだけでは高精度な結合判定はできないものの、他の属性と併用して学習することで、ドッキングスコアよりも高精度な結合判定が行える。

本研究ではたんぱく質と化合物間の情報量を増やすため、複数のドッキングソフトを用いて化合物とたんぱく質のドッキングスコアを計算する。特に、実際にたんぱく質に結合する化合物

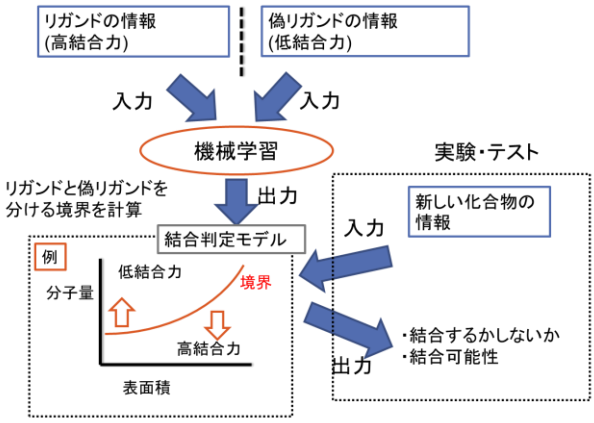

図 1. 機械学習

は、どのドッキングソフトを用いた場合でもスコアが高く出やすいため[Okada 2011]、単一のドッキングスコアよりも判定精度の向上が望める。

3.2 化合物情報

化合物情報とは化合物の化学的な性質を示す、化合物固有の情報である。例として、分子量や表面積、疎水性などがある。また、本研究で用いる化合物情報は、化合物情報を計算するための計算ソフト(Discovery Studio[Accelrys 2008] の Molecular Properties 等)で得られる。ただし、どの化合物情報が結合判定に有用であるかどうかはわからない。そのため、利用できる範囲でできる限り多くの化合物情報を用意すべきである。特に、本研究で用いている SVM は属性数の増加による判定性能低下が少ないため、十分な量の属性を用いることができる。これら多数の化合物情報から結合する化合物特有の性質を機械学習によって分析する。

以上のデータから機械学習により結合判定モデルを作成することで、化合物の結合判定を高精度で行うことが可能である。しかし、たんぱく質によってはドッキングスコアの計算が正しく行われない場合がある [Hanaya 2012]。たとえば、たんぱく質と化合物の結合部位に金属原子が存在する場合、ドッキングソフトによっては正しい結合ポーズが得られない。

3.3 複数たんぱく質でのドッキングスコア

本研究では上記の問題を解決するため、複数たんぱく質でのドッキングスコアを計算することで、より多くの化合物に関する情報を入手する。これは、他のたんぱく質とのドッキングスコアから、高結合力化合物と低結合力化合物とを分ける特徴を得るということである。特に、結合部位の似ているたんぱく質では同じようなスコア傾向がみられることから[Omagari 2009]、金属原子を含まず、研究対象のたんぱく質と構造の似たたんぱく質でのドッキングスコアを用いることによって、判定が可能になることが期待できる。また、あるたんぱく質に結合する化合物は、他のたんぱく質に対する結合力は低くなり、ドッキングスコアが低くなる傾向がある[Omagari 2009]。そのため、金属原子を含まないようなたんぱく質においても、情報量の増大につながる。

本研究では各化合物について以上のデータを収集し、機械学習に用いる。表1に機械学習に用いるデータを示す。各化合物は複数のドッキングソフトおよび複数のたんぱく質から得られるスコアと、化合物固有の化合物情報を持つ。また、すでに結合するかどうかがわかっているデータでは、クラスラベルに 1 または 0 のデータを付加する。本研究では学習を行う際も、実験

表 1. 機械学習に用いるデータ

	Protein 1, Scores		Protein 2, Scores		…	Chemical information		
Name	Libdock	Cdocker	Libdock	…		AlogP	…	Class
Ligand 001	160.2	…	63.2			0.909		1
Decoy 001	120.3		159.3			1.012		0
…								

を行う際もこの形式を用いるが、実験時にはクラスラベルの値がないデータとなる。

4. 実験

4.1 実験手法

ここでは炭酸脱水酵素(Carbonic Anhydrase 2)を用いた化合物の結合判定実験の結果を示す。炭酸脱水酵素は結合部位に亜鉛が含まれており、ドッキングソフトによる結合シミュレーションでは正しく計算できない場合が多い[Hanaya 2012]。そのため、バーチャルスクリーニングを行うことが困難になっている。

本実験では次の手順により、本研究の手法の有用性を示す。第一に学習データを用意し、その学習データを用いて適切な機械学習パラメータを求める。第二に学習データから結合判定モデルを作成し、炭酸脱水酵素に対する活性情報が存在するものの、結晶構造が得られていない化合物ついて結合可能性を計算する。この際の判定性能が高ければ、新しい化合物の結合判定が可能であるといえる。

本実験では学習データとして A Database of Useful Decoys: Enhanced (DUDE) [Mysinger 2012] のデータを用いた。DUDE では活性データベース Chembl [Gaulton 2012] からたんぱく質に対する活性が強い化合物を集めている。また、収集したリガンドについて、その化合物と構造や性質が似ているものの、たんぱく質と結合しないであろう化合物をランダムに収集している。以上の手法により、DUDE では多数のたんぱく質について、たんぱく質に結合するリガンドと、たんぱく質に結合しないデコイを収集し、提供している。DUDE から用いたデータは、CA2 に結合するリガンドを 835 個、CA2 におけるデコイ化合物を 2000 個入手した。ただし、835 個のリガンドには Chembl 上の名称が同一である化合物が存在し、Chembl 名称基準ではリガンド数は 492 個であった。

本実験ではテストデータとして、活性情報がわかっているものの、結晶構造が得られていない 19 種の化合物を使用する。化合物を図 2 および図 3 に示す。これらの化合物には活性の強い化合物と活性の弱い化合物の両方が含まれている。また、Ki 値を活性の強さとしたとき、1 μ M 以上で活性の弱い化合物が 5 個、1 μ M 以下で活性の強い化合物が 14 個であった。

学習とテストに共通して、本実験では化合物から以下の手法で化合物の特徴を計算し、属性として機械学習に用いた。第一にドッキングソフトは DiscoveryStudio で実行可能な CDOCKER と LibDock を用いた。これらのドッキングソフトはバーチャルスクリーニングに用いられており、結合ポーズの予測精度は非常に高い。第二に、化合物情報は DiscoveryStudio で実行可能な Molecular Properties を用いて計算した。そして、可能な限りの化合物情報を求めた上で、全化合物で値が得られた属性、ひとつの値のみを持つ属性、全化合物で偏差が 0 でない属性を選択した。結果、67 の属性を使用した。第三に他のたんぱく質として 30 個のたんぱく質を利用し、計 31 個のたんぱく質について、Libdock と CDOCKER の二つのドッキングソフトによってドッキ

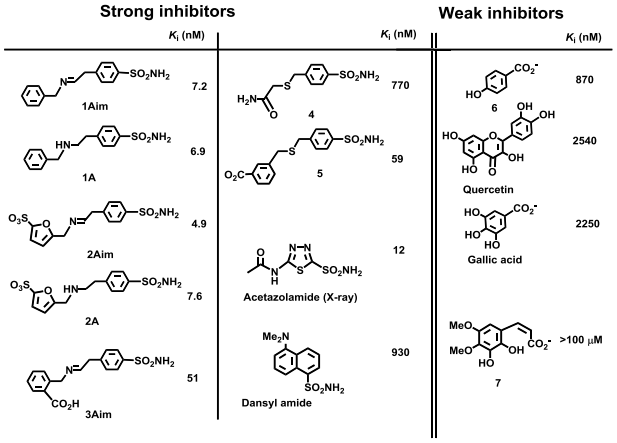

図 2. テスト化合物 1

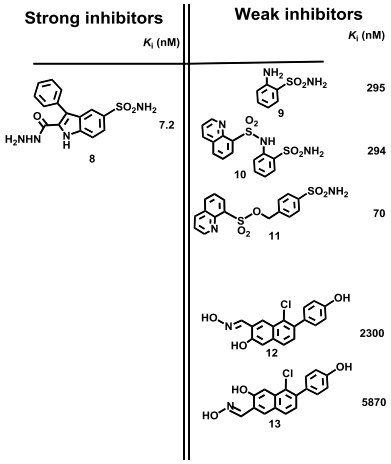

図 3. テスト化合物 2

ングスコアを計算した。そのため属性の数は、67 化合物情報＋2 ソフト×31 たんぱく質で 129 個となった。

また、本実験では SVM のソフトウェアとして、LIBSVM[Chang 2011] の Java を用いた。カーネルは RBF を、SVM タイプは epsilon-SVR を用いている。epsilon-SVR は SVM 回帰計算(Support Vector Regression) を行うための SVM タイプである。

4.2 学習用データにおける判定性能

まず、学習用データ 2835 個を用いて、判定性能の高い SVM のパラメータを求めた。このとき、Leave-One-Out 法によって各化合物の判定を行った。ただし、Chembl 名称が同一である化合物については学習に用いないように設定した。また、良好な SVM パラメータは、グリッドサーチによって求めた[Gestel 2004]。

グリッドサーチによってパラメータを変化させ、結合可能性スコアの結合可否の閾値を 0.5 として精度の変化を見た。結果、SVM パラメータが-c 100、-g 0.001 のとき、適合率(Recall) 98.2%、再現率(Precision) 96.5%、精度(Accuracy) 98.4%の良好な判定精度を得た。

また、図 4 に、結合判定の閾値を変えていったときに、結合すると判定された化合物の割合がどう変化するかを表す ROC 曲線を示す。ROC 曲線では、曲線の左寄り部分は判定閾値を厳しくしたときの正負事例の検出割合を示し、右寄り部分は判定閾値をゆるくしたときの検出割合を示す。そのため、グラフが左上に寄るほど、結合する化合物をよく検出できているといえる。図において、ドッキングスコアから作成した ROC 曲線は、対角線付近にあり、判定性能が非常に低い。対して、本研究の手法で得られる曲線は左上に寄っており、判定性能が非常に高いことがわかる。曲線の下部面積（AUC 値）は 0.997 であった。

4.3 テストデータにおける判定性能

4.2 で得られたパラメータおよび学習データを用いて、テストデータの結合判定を行った。19 個の化合物における結合可能性スコアと、実際の活性値の関係を図 5 に示す。x 軸は結合可能性スコアを、y 軸は活性値 Ki について、Log(1/Ki)をとった値である。そのため、1μM は y 軸では 6 となる。また、図中の斜線は、回帰直線を示す。

図において、結合の閾値である 0.5 を境に、活性の強い化合物の多くが右側に、活性の弱い化合物が左側に集まっている。これは、4.2 の実験と同様に、テストデータについて結合判定が行えていることを示す。ただし、4.2 の実験と比較して、判定精度は低くなっている。

また、回帰直線およびその R^2 値を見たとき、結合可能性スコアと Log(1/Ki)に高い相関が見られる。特に、結合可能性スコアが 0.5 を超える化合物の中でも、活性の弱い化合物は比較的結合可能性スコアが低くなっている。結合可能性スコアを元に結合の判定を行うことで、活性の強い化合物を得ることが可能である。

5. おわりに

本研究では機械学習による化合物の結合判定手法を提案した。本研究の特徴として、機械学習により、0~1 の値を持ち、0.5 の閾値を持つ結合可能性スコアを出すこと、そして複数たんぱく質のドッキングスコアや化合物情報を機械学習に用いることがあげられる。本研究の手法により、新しい化合物がたんぱく質に結合するかどうか、活性が強いかどうかを自動的に判定することができ、創薬研究におけるバーチャルスクリーニングを行うことが可能である。

参考文献

[Trott 2010] 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.

[Rao 2007] Rao SN., Head MS., Kulkarni A., LaLonde JM.: Validation studies of the site-directed docking program LibDock, J Chem Inf Model. 2007 Nov-Dec; 47(6):2159-71, 2007.

[Wu 2003] 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.

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

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

[Moitessier 2008] 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.

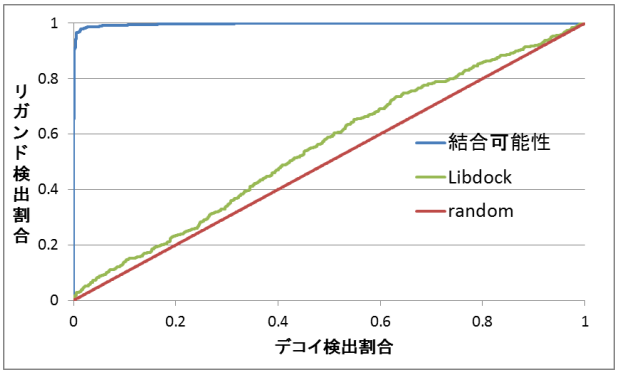

図 4. 機械学習時の ROC 曲線

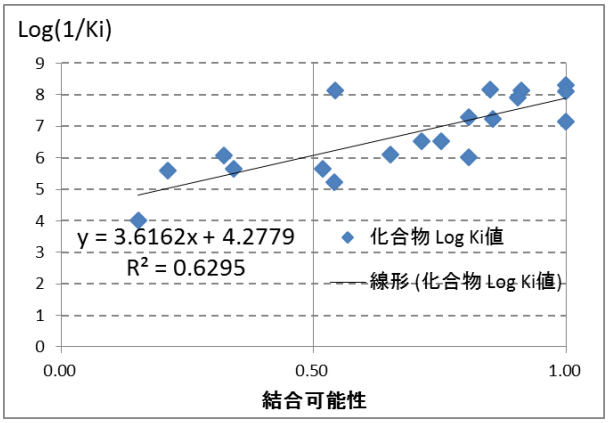

図 5. 活性値と結合可能性

[Wang 2003] Wang,R.: Comparative evaluation of 11 scoring functions for molecular docking. J. Med. Chem., 46, 2287-2303. 2003.

[Borman 2005] Borman S.: Drugs by design. Chem Eng News, 83: 28-30. 2005.

[Okamoto 2010] 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.

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

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

[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. 2005.

Gaulton 2012] Gaulton A., Bellis L. J., Bento A. P., Chambers J., Davies M., Hersey A., Light Y., McGlinchey S., Michalovich D., Al-Lazikani B., Overington J. P.,: ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 2012, 40, D1100.1107, 2012.

[Okada 2011] Masato Okada, Mmasato Tsukamoto, Hayato Ohwada, Shin Aoki: 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, 94-98,2011.3, 2011.

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

[Hanaya 2012] Kengo Hanaya, Miho Suetsugu, Shinya Saijo, Ichiro Yamato, 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.

[Omagari 2009] 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.

[Mysinger 2012] Mysinger MM., Carchia M., Irwin JJ., Shoichet BK.: Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking, J. Med. Chem., 2012, 55 (14), pp 6582—6594, 2012

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

[Chang 2011] C.-C. Chang, C.J. Lin: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27, 2011.

[Gestel 2004] T. van Gestel, J.A.K. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. de Moor, J. Vandewalle :Benchmarking least squares support vector machine classifiers, Machine Learning, vol. 54, pp. 5-32, 2004.