Vol.2013-BIO-34 No.15
2013/6/28

結合自由エネルギー計算を用いた薬物クリアランス経路予測の改善

齊藤有紀[†1]　石田貴士[†1]　関嶋政和[†1]　秋山泰[†1]

概要：近年，薬剤開発に必要な時間や費用が増大しており，薬剤開発の期間短縮や費用削減が求められている．そのための手法としてコンピュータ上でのシミュレーションを用いて，ターゲットとなるタンパク質を阻害する薬物の構造を設計する手法が注目を浴びている．一方で，薬物として使用する化合物は，体内で代謝・排泄されなければならないという条件がある．この条件を満たす薬物の選定のために行われるのが薬物クリアランス経路予測である．そこで我々は，コンピュータ上で行われるシミュレーションのひとつである，タンパク質と化合物のドッキング計算による結合自由エネルギー計算を，薬物クリアランス経路の予測に応用し，これまで我々が開発してきたクリアランス経路予測システムの精度改善を行った．

キーワード：薬物クリアランス経路予測，創薬支援，ドッキング計算，結合自由エネルギー計算，機械学習

Improvement of drug clearance pathway prediction using binding free energy calculation

YUKI SAITO[†1]　TAKASHI ISHIDA[†1]　MASAKAZU SEKIJIMA[†1]
YUTAKA AKIYAMA[†1]

Abstract : Recently, drug development needs longer term and larger budget than before, so it is needed to decrease of them. From such a reason, the computer simulation-based techniques to design drugs which inhibit a target-protein is in the spotlight now. On the other hand, the drugs must be metabolized and excreted. And, one of the techniques to search the chemical compound which satisfy this condition is the drug clearance pathway prediction. Then, we tried to use binding free energy calculation of protein-ligand complex with the docking calculation, one of the computer simulation, for the drug clearance pathway prediction, we improved precision of the system for drug clearance pathway prediction previously developed in our lab.

Keywords : drug clearance pathway prediction, drug discovery assistance, docking calculation, binding free energy calculation, machine learning

1. はじめに

現在，新薬の開発には多くの費用と長期にわたる開発期間が必要となっており，開発後期の段階において開発が中止となると非常に大きな損失となる [1]．そのため，開発初期の段階における薬物候補化合物の選定精度を向上させ，無駄なコストをできる限り削減し，開発期間の短縮を行うことの重要性が高まっている．

薬物候補化合物の選定には様々な手法が取り入れられているが，我々は化合物の物理学的特徴を入力として機械学習を用いることで，薬物クリアランス経路の予測を行うシステムの開発を行ってきた [2)3)4)5)6)7)]．クリアランス経路とは薬物が代謝・排出される経路であり，このクリアランス経路を正確に予測することができれば，生化学実験を必ずしも行わなくても代謝・排出されない化合物を薬物候補化合物から除外することができ，薬物候補化合物の選定精度を向上させ，無駄なコストをできる限り削減することが可能になる．我々が以前に開発した手法では，有機アニオントランスポーターによる肝臓への取り込み(Organic anion transporting polypeptide, OATP)などの経路については高い予測精度を達成した一方，グルクロン酸抱合(UDP-glucuronosyltransferase, UGT)やシトクロムP450(Cytochrome P450, CYP)の一部の経路については十分な予測精度が得られておらず，それらの経路の予測については，更なる改良が必要とされている．

CYPは，水酸化酵素ファミリーの総称であり，いくつもの種類のCYPが発見されている．それらはアミノ酸配列の相同性によって分類され，CYP1A2,CYP2C9,…といった名前が付けられている [8)]．

薬物候補化合物とタンパク質のドッキング計算は，薬物候補化合物とタンパク質の構造データから，分子間，原子間に働く力を計算することで最も安定な構造やその時のエネルギーを計算する手法である．ドッキング計算は阻害剤の設計などに用いられてきたが，薬物クリアランス経路予測の領域においても，薬物代謝を行うタンパク質は必ず薬物と結合して代謝を行うため，ドッキング計算によって代謝タンパク質と薬物候補化合物との結合自由エネルギーを計算し，代謝を行うタンパク質と結合するか判定することでクリアランス経路予測に利用できる可能性が考えられる．

そこで本研究では，このドッキング計算を用いて，我々

†1 東京工業大学大学院情報理工学研究科
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

ⓒ2013 Information Processing Society of Japan

がこれまで研究を行ってきた機械学習によるクリアランス経路予測の精度を改善することを試みた．この手法では腎排泄(Renal)のような過程の複雑な経路の予測は難しいため, 代謝タンパク質がはっきりとしている CYP を今回の対象経路とする。

2. 先行研究

計算機上での薬物クリアランス経路予測に関する研究はこれまで複数の研究グループによって行われてきており，CYP による水酸化，ドーパミン-β-水酸化酵素，UGT 等の経路について既知の薬物の構造から論理プログラミングによって予測を行う手法 [9]，CYP による薬物代謝を機械学習を用いて予測する手法 [10]や，CYP1A2 の阻害剤かどうかを機械学習で予測する手法 [11]などが提案されている．また，我々は矩形領域法 [2][3][4]，サポートベクターマシン(support vector machine, SVM)[5]，ブースティング [6]，など様々な学習アルゴリズムを用いた予測システムを開発してきた．

これらの手法は単一のクリアランス経路を持つ薬物を対象として行われた研究であったが，薬物の中には複数の代謝タンパク質の作用によって代謝されていくものも存在する．そういった薬物のクリアランス経路予測に対応するために我々は複数のクリアランス経路を持つ薬物に対する予測手法を開発し，さらに精度向上のために多段階での予測を行った [7].

我々の研究では，{CYP3A4, CYP2C9, CYP2D6, CYP1A2, CYP2C8, CYP2C19, Renal, OATP, UGT}の 9 経路を対象として，薬物クリアランス経路の予測を行った．

先行研究において示された多段階予測による薬物クリアランス経路予測の f 値による精度を示す．f 値は recall と precision の調和平均で表される．recall と precision は以下のように求められる．

- True Positive(TP):実際に正例のものを正しく予測できた時
- True Negative(TN):実際に負例のものを正しく予測できた時
- False Positive(FP):実際は負例のものを間違って予測した時
- False Negative(FN):実際は正例のものを間違って予測した時

$$\text{recall} = \frac{TP}{TP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

一般に recall の値を上げれば precision の値を上げれば recall の値が下がる傾向にあり, その調和平均である f 値は総合的な判定性能を表す．表 1 から明らかなように，CYP2D6，CYP1A2 および UGT の予測精度が他の経路に比べて低くなっており，この手法は薬物のクリアランス経路予測としてまだ十分完成しているとは言えない.

表 1　先行研究における結果
Table 1 Result of prior research

	f 値
CYP3A4	0.765
CYP2C9	0.611
CYP2D6	0.388
CYP1A2	0.489
CYP2C8	0.750
CYP2C19	0.616
Renal	0.711
OATP	0.770
UGT	0.433

3. 薬物ドッキングによる結合自由エネルギー計算

ドッキング計算は通常，ターゲットのタンパク質のポケット構造によく適合する薬物を設計するために利用されるが, その適合の良し悪しは結合自由エネルギーを近似的に計算することで判定している. ある薬物が代謝されるためには, まずその代謝に関わるタンパク質と結合することが必須であるが，その薬物が代謝に関わるタンパク質と結合しないことを知ることができれば，我々はその薬物がその経路によって代謝されないことを知ることができる．そのため，薬物と代謝に関わるタンパク質との間の結合自由エネルギーを正確に求めることができれば，その情報からクリアランス経路予測の精度の改善がされる可能性が考えられる.

3.1 薬物ドッキング

薬物ドッキングは，低分子化合物とタンパク質との間の複合体構造をコンピュータ上で計算的に推定する手法である.

ドッキング計算を行うプログラムは, 1982 年に開発された DOCK[12]をはじめ，GOLD[13]，FlexX[14]，Glide[15]等数多く存在する. 本研究では, 無償で使用することが可能であり, 長期間に渡って多くの薬剤開発研究で使用されてきた AutoDock[16]を使用した.

3.1.1 AutoDock

AutoDock はスクリプス研究所によって開発されたタンパク質-リガンド間のドッキング計算を行うプログラムである. AutoDock では複合体構造の探索アルゴリズムに遺伝的アルゴリズムを使用している.

ドッキングポジションの探索では図 1 に示すように，まずタンパク質および化合物の構造を変化させて複数種類発生させ，化合物の位置を移動させてエネルギー的に最も安定なポジションを探索する.

ⓒ2013 Information Processing Society of Japan

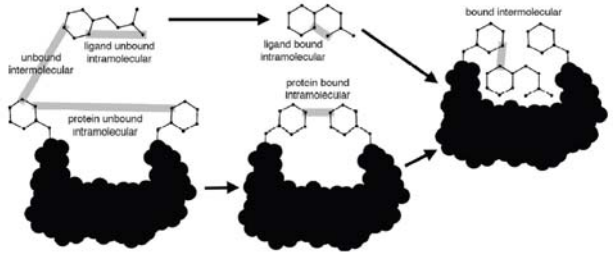

図 1　AutoDock のドッキング手順
(AutoDock4.2 User Guide[17]より引用)
Figure 1 Procedure of AutoDock docking calculation
(from AutoDock4.2 User Guide[17])

AutoDock では，iを化合物のi番目の原子，jをタンパク質のj番目の原子として以下のようなスコア関数を用いて結合自由エネルギーを kJ/mol 単位で計算し，そのエネルギーの安定な順に結果をランキングして返す.

$$\Delta G = f_{vdW}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}\right)+f_{elec}\sum_{i,j}\left(\frac{q_i q_j}{\epsilon(r_{ij})r_{ij}}\right)$$
$$+f_{hbond}\left(\sum_{i,j}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}}-\frac{D_{ij}}{r_{ij}^{10}}\right)+E_{hbond}\right)$$
$$+\Delta G_{tor}N_{tor}+f_{sol}\sum_{i,j}E_{ij}e^{\left(-\frac{r_{ij}^2}{2\sigma^2}\right)}$$

ただし，r_{ij}はiとjの距離，A_{ij}, B_{ij}はファンデルワールス・パラメータ，C_{ij}, D_{ij}は水素結合を表現するポテンシャルの係数，E_{ij}は原子間距離のガウス関数の係数である．また，E_{hbond}は水素結合の数に基づく補正項，$f_{vdW}, f_{elec}, f_{hbond}, f_{sol}$はそれぞれファンデルワールス，静電相互作用，水素結合，脱溶媒和の各項の重み係数を表す．ΔG_{tor}は化合物がタンパク質と結合することで構造が安定することによる単結合 1 つあたりのエントロピー減少によるエネルギー減少を表し，それに回転可能な結合の数をかけて自由エネルギーに考慮している．

4. 提案手法

本研究ではドッキングによって推定された結合自由エネルギーを用いて薬物クリアランス経路の予測を行う 2 つの手法を開発した．第一の手法では，結合自由エネルギーの値で従来手法の予測結果にフィルターをかけることで，不正解化合物を除外して精度を向上させることを目指した．第二の手法では，結合自由エネルギーの値を機械学習の特徴量に加え，薬物クリアランス経路予測を行った．

この手法を行うには，代謝に関わるタンパク質の構造が決定されている必要がある．表 1 にある経路のうち，CYP についてはすべて構造データが確認された．このうち，CYP1A2 および CYP2D6 についての f 値が 0.5 を下回り，予測精度が十分でないと考えられるため，対象経路をこの 2 つとする．

4.1 データセット

本研究では堀田らの先行研究 [7]で用いられたものと同じデータセットを使用した．

データセットは各化合物の電荷，血漿タンパク質非結合率，分子量，分配係数の 4 つの物理化学的特徴量と，薬物クリアランス経路をまとめたものである．このデータセットは {CYP3A4, CYP2C9, CYP2D6, CYP1A2, CYP2C8, CYP2C19, Renal, OATP, UGT}の 9 種類の代謝経路に関するデータセットであり，複数のクリアランス経路を持つ化合物も存在する計 244 化合物のマルチラベルのデータセットである．ただし，本研究では上述の通り CYP1A2, CYP2D6 の 2 経路のみを予測対象とした．

4.2 提案手法 1：結合自由エネルギーによるフィルタリング

本手法は，既存の予測手法において，あるクリアランス経路で代謝されると判定された候補化合物について結合自由エネルギーの値をもとに並び替え，ある基準を満たす候補化合物をその経路で代謝されると予測する手法である．これは誤って代謝されると予測をされている化合物を予測から取り除くことを目指すものである．以下この手法を「結合自由エネルギーによるフィルタリング」と呼ぶ．

4.3 提案手法 2：結合自由エネルギーを特徴量に加えた予測

2 つ目の提案手法は，先行研究で選んだ 4 つの特徴量に，結合自由エネルギーを新たに特徴量として加えた学習を行う手法である．提案手法 1 では結合自由エネルギーの値とその他の特徴量を線形に扱っているため，機械学習の特徴量に直接加えることにより，さらに複雑な関係を学習可能になる可能性がある．先行研究では多段階に SVM を使用する手法を提案していたが，分類器が複雑になってしまうため今回は一段階の SVM で学習，予測を行った．以下この手法を「結合自由エネルギーを特徴量に加えた予測」と呼ぶ．

4.3.1 学習パラメータと特徴量の正規化

SVM のカーネルには, 線形カーネルや多項式カーネル等が存在するが，今回はカーネルとして以下の式で表される Gaussian Kernel を用いた．

$$\mathrm{K}(x,z) = exp(-\gamma\|x-z\|^2)$$

また，γの値として，{0.005，0.01，0.05，0.1，0.2，0.5，0.7，1.0，2.0，5.0，7.0，10} の 12 通りとソフトマージンのコストパラメータ C の値として, 以下の 226 通りを探索.

- 0.01 から 0.1 まで 0.01 刻みで 10 通り
- 0.15 から 1.0 まで 0.05 刻みで 18 通り
- 1.5 から 100 まで 0.5 刻みで 198 通り

計 2712 通りの組み合わせで予測を行い，f 値が最大になる組み合わせを探索した．

また，正解化合物と不正解化合物の個数に偏りがあるため，先行研究と同様にその比率が 1:1 になるようにデータに重みづけをした．

各特徴量は値の範囲が異なるため，学習を行う前に各特徴量を正規化(normalization)した．正規化の方法としては，先行研究に従い以下の変換によって正規化を行った．

これは，Xを各データ値，μはデータの平均値，σはデータの標準偏差とし，平均 0，分散 1 となるように，以下の式を用いて変換を行ったものであり，一般に z_score と呼ばれる．

$$X_{norm} = \frac{X - \mu}{\sqrt{\sigma^2}}$$

5. 実験と考察

5.1 タンパク質立体構造データと化合物構造データ

タンパク質の立体構造データは PDB[18]からダウンロードした．CYP2D6 の構造データは PDBID:2F9Q のものを使用した. CYP1A2 については, 単体の結晶構造のデータが PDB に存在しなかったため，alpha-naphthoflavone との共結晶構造である PDBID:2HI4 のものから CYP1A2 部分を取り出して使用した．

化合物構造データは ZINC[19]からダウンロードした. この際，適当な構造データが発見できなかった化合物があり，先行研究よりデータ数が減少している．また先行研究において，その経路において代謝されるかどうかはっきりとした分類ができていなかった化合物もあり, それも除外した. その結果, 実際に本実験に用いた化合物数は表 2 のようになった．

表 2　実験に使用した化合物数
Table 2 Number of compounds

タンパク質	先行研究 [7]の化合物数		本研究の化合物数
CYP1A2	244	⇒	217
CYP2D6	244	⇒	219

5.2 予測性能評価

本研究では，結合自由エネルギー予測値を特徴量に加えた SVM による予測を使用する際に交差確認法(cross validation)を使用し評価を行った. 提案手法 2 の性能評価では, leave-one-out cross validation を使用し, f 値を算出した．

5.3 実験結果

5.3.1 実験 1 : 結合自由エネルギーによるフィルタリング(提案手法その 1) と従来手法の比較

結合自由エネルギー予測値によるフィルタリングを行った場合と従来手法単体との予測性能の比較を行った．

図 2 に CYP1A2 と CYP2D6 について結合自由エネルギー予測値による従来手法のフィルタリングと従来手法によるそれぞれの f 値を示した．CYP1A2 について，従来手法の f 値は 0.489 であったのに対し, 提案手法 1 では 0.578 と大きく改善した．CYP2D6 についても同様で，従来手法の f 値が 0.387 であったのに対し，提案手法 1 では 0.412 に改善した．

図 2 より, 薬物ドッキング計算による結合自由エネルギーの予測値は従来手法の精度を改善するためのフィルターとして利用し得ることがわかった．しかし，その有効性について検定は行っていないため，今後十分に検討する必要がある．

5.3.2 実験 2 : 結合自由エネルギーを特徴量に加えた予測(提案手法 2)

ドッキング計算による結合自由エネルギーの予測値を新たに化合物の特徴量として加え，計 5 つの特徴量を入力にした SVM を使用した．また，特徴量はすべて z_score で正規化した．この結果を，従来の 4 つの特徴量を入力とする SVM による予測の結果と比較した．

図 3 に CYP1A2 と CYP2D6 について結合自由エネルギーを特徴量に加えた SVM による予測と従来の SVM による予測のそれぞれの f 値を示した．この結果，CYP1A2 については提案手法 2 の f 値が 0.525 となり，他の手法を上回った. しかし, CYP2D6 については改善が見られなかった．

これは，一部の候補化合物の構造データを見つけられなかったために，正例，負例それぞれの代表点となるような特徴的な化合物が除外され，学習によってきれいな境界面を見つけることが困難だった可能性がある．そのため，先行研究に用いたデータと同じデータを用いることで CYP2D6 についても改善が見込まれる．

解決法としては，他の化合物データベースから構造データを収集することが考えられるが，それでも発見できない場合は構造式から分子モデルを作成するソフトウェアを使用するなどして独自で構造データを用意する必要がある．

5.4 クリアランス経路予測の考察

実験 2 の結果では期待したほどの精度の改善は見られなかった．しかし，1 段階の SVM 同士で比較すれば結果は改善しており，結合自由エネルギーを特徴量に加えるという手法自体は正しいと考えられ，先行研究と同様に多段階での予測を行えば先行研究の精度よりも高い精度での予測を行うことが可能であると考えられる．

ⓒ2013 Information Processing Society of Japan

図 2　提案手法 1 による予測および従来手法による予測の f 値
Figure 2 f-value of method 1 (filtering of previous method with docking free energy)

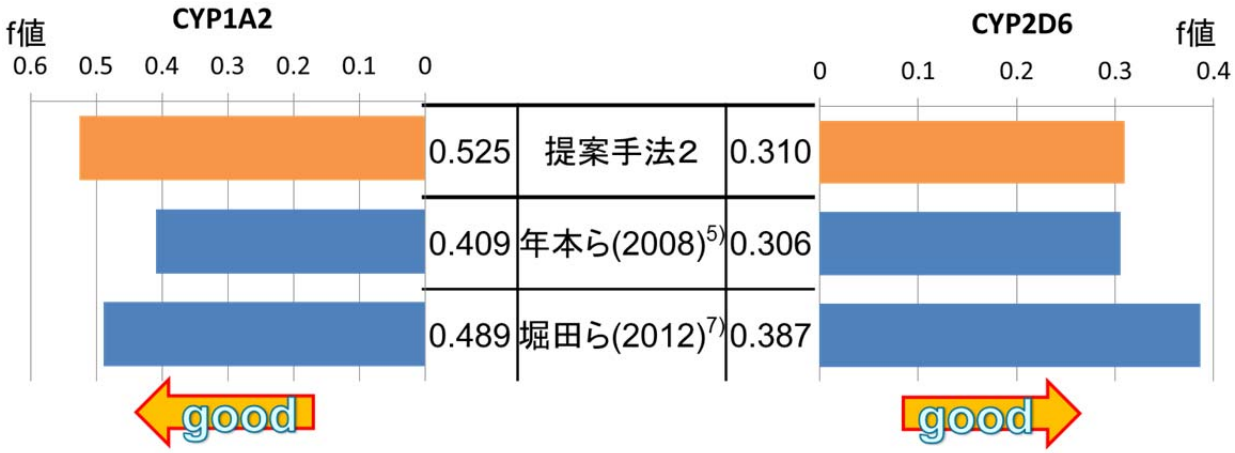

図 3 提案手法 2 による予測および従来手法による予測の f 値
Figure 3 f-value of method 2 (prediction with SVM using binding free energy)

6. 結論

6.1 本論文のまとめと結論

本研究では，ドッキング計算による結合自由エネルギー計算を利用して，薬物クリアランス経路予測の精度改善を行った．その手法として，結合自由エネルギー予測値によるフィルタリング(提案手法 1)と結合自由エネルギー予測値を特徴量とする SVM を用いて学習・予測を行う手法(提案手法 2)を提案した．

提案手法 1 では，従来の薬物クリアランス経路予測の手法の予測結果に結合自由エネルギーの予測値によるフィルタリングを行うことで，ある経路で代謝されると誤って判定されていた候補化合物を除外することにより，予測精度を向上させる効果があることが分かった．

提案手法 2 では，結合自由エネルギーの予測値を特徴量とする SVM による学習・予測を行ったが，期待した程の精度改善は見られなかった．しかし，学習データ数の増加やドッキング計算の精度改善，学習・予測の多段階化などにより，今後，従来手法の精度を改善できる可能性がある．

これらの結果により，本研究では結合自由エネルギーの予測値を他の特徴量とあわせて使うことで従来までの手法による予測結果の精度を改善することができることがわかった．

6.2 今後の課題

本研究の実験において，いくつかの化合物を除外しており，先行研究との厳密な比較ができていないと思われる．そのため，データセットを統一して再実験を行うことで SVM の学習精度も上がり，分類精度が上がる可能性がある．

また，今回の実験 1 および実験 2 では厳密な学習や予測を行わずに精度を比較したが，きちんと学習及び予測を行った上での精度はどのようになるのか確認する必要がある．

今回の実験 2 では学習，および予測が複雑になってしまうため，5 つの特徴量による多段階での予測は行わなかった．正確に先行研究の結果と比較するためには，5 つの特徴量を用いた多段階予測について今後改めて精度比較を行う必要があると考えられる．

ⓒ2013 Information Processing Society of Japan

謝辞

今回の研究データを使わせていただいている理化学研究所杉山特別研究室の杉山雄一先生に御礼申し上げます.

参考文献

1) 杉山雄一，楠原洋之編:“分子薬物動態学”，南山堂(日本), pp.2-pp.28:pp.99-pp.153，2008.
2) Kusama M，Toshimoto K，Maeda K，Hirai Y，Imai S，Chiba K，Akiyama Y，Sugiyama Y，“In Silico Classification of Major Clearance Pathways of Drugs with Their Physiochemical Parameters”，Drug Metabolism and Disposition Vol 38(8)，pp.1362-pp.1370 ，2010.
3) 草間真紀子，平井由香，前田和哉，今井覚己，千葉康司，年本広太，秋山泰，杉山雄一．:“医薬品の物理化学的特性に基づいた薬物動態プロファイリング(I).”，第 24 回日本 DDS 学会，2008
4) Kusama M.，et al.:“Classification of major clearance pathways of drugs based on physicochemical parameters.”，23rd Annual Meeting of the Japanese Society for the Study of Xenobiotics，2008.
5) 年本広太，草間真紀子，前田和哉，杉山雄一，秋山泰.“医薬品の物理化学的特性に基づいた薬物動態プロファイリング(II).”，第 24 回日本 DDS 学会，2008.
6) Ikeda K.，et al.:“Prediction of drug clearance pathway by boosting algorithm”，IPSJ SIG Technical Report 2009-BIO-17(10)，pp1-pp.8，2009.
7) 堀田　駿，他.:“ウェブアプリケーションによるクリアランス経路予測”, IPSJ SIG TechnicalReport 2010-BIO-21(20), pp.1-pp.8, 2012.
8) 大村恒雄，石村巽，藤井義明:“P450 の分子生物学”，講談社(東京)，pp.1-pp.13:pp.244-pp.248，2003.
9) Ferenc D.:“ Predicting metabolic pathways by logic programming ”，J. Mol. Graphics Vol 6(8)，pp.80-pp.86，1988
10) Dmitry K.，et al.:“ Modeling of Human Cytochrome P450-Mediated Drug Metabolism Using Unsupervised Machine Learning Approach ”，J. Med. Chem. 46，pp.3631-pp.3643，2003.
11) Poongavanam V.，et al.:“ Classification of Cytochrome P450 1A2 Inhibitors and Noninhibitors by Machine Learning Techniques ”，DRUG METABOLISM AND DISPOSITION vol 37(3)，pp.658-pp664，2008.
12) Kuntz I.D.，et al.:“A geometric approach to macromolecule-ligand interactions.”，J. Mol. Biol.161，pp.269-pp.288，1982.
13) Jones G.，et al.:“Development and Validation of a Genetic Algorithm for Flexible Docking”，J.Mol. Biol. 267，pp.727-pp.748，1997.
14) Rarey M.，Kramer B.， Lengauer T.，Klebe G.:“A fast flexible docking method using an incremental construction algorithm.”，J. Mol. Biol. 261(3)，pp.470-pp.489，1996.
15) Richard A.F.，et al.:“Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy”，J. Medicinal Chemistry 47(7)，pp.1739-pp.1749，2004.
16) Morris G.，et al.:“Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity.”，J. Computational Chemistry 16，pp.2785-pp.2791，2009.
17) “AutoDock4.2 User Guide” http://autodock.scripps.edu/faqs-help/manual/autodock-4-2-user-guide/AutoDock4.2 UserGuide.pdf
18) Berman H.M.，et al.:“The Protein Data Bank”，Nucleic Acids Research 28(1)，pp.235-pp.242，2000.
19) John J.I., Teague S., Michael M.M., Erin S.B., Ryan G.C..:“ZINC: A Free Tool to Discover Chemistry for Biology”，J. Chem. Inf. Model.，2012.

ⓒ2013 Information Processing Society of Japan