Vol.2014-MPS-98 No.10
Vol.2014-BIO-38 No.10
2014/6/26

半教師付き学習を用いた薬物クリアランス経路予測

柳澤 渓甫[1,a)]　石田 貴士[1]　秋山 泰[1,2]

概要：近年，新薬開発にかかる時間および費用は莫大なものであり，この費用や開発期間の削減が求められている．薬剤候補化合物が新薬として認められるためには体内で代謝・排泄されるという安全性を確認する必要があるが，この観点から計算機を用いることで薬物候補化合物の早期の選定を行うのが薬物クリアランス経路予測と呼ばれるものである．この予測問題は既知薬物のクリアランス経路を利用して学習を行うが，この情報を得るには実験が必要となるため，データ数が少ないことが問題となっている．そこでこの予測問題に対する半教師付き学習の有用性を評価し，この手法を用いることで予測精度の改善を試みた．また，入力の特徴量を増加させることが有効であると考え，貪欲法により 802 個の化合物記述子より選択した特徴量の追加を行い，その効果も検証した．

キーワード：薬物クリアランス経路予測，創薬支援，機械学習，半教師付き学習

Drug clearance pathway prediction using semi-supervised learning

Yanagisawa Keisuke[1,a)]　Ishida Takashi[1]　Akiyama Yutaka[1,2]

Abstract: Nowadays, drug development requires too much time and budget, and it is necessary to reduce them. In order to accept a compound as a new drug, it must be confirmed that it is metabolized and excreted. In this respect, one of the computational methods used for selecting compounds is drug clearance pathway prediction. This prediction method uses well-known drug's clearance pathway data as a training set. However data is expensive to get, and thus there are too few data. For this reason, we evaluated the usefulness of semi-supervised learning in this prediction problem, and tried to improve accuracy of this clearance pathway prediction. We also tried to add some features of compounds which are selected from 802 features by greedy algorithm to improve accuracy and evaluated their effect.

Keywords: drug clearance pathway prediction, drug discovery assistance, machine learning, semi-supervised learning

1. はじめに

現在，1 つの新薬が開発されるまでには候補化合物の探索から臨床研究まで含め 10 年以上の年月・数百億ドルの費用が必要となっている [1]．もしこの膨大な時間や費用をかけて研究を進めている候補化合物に問題が見つかり開発中止になると，製薬会社にとっては大きな損害となってしまう．このため，創薬研究の初期段階において薬物候補化合物の選定精度を向上させることはコスト・期間両面から重要であり，コンピュータを用いた様々な研究が盛んに行われている [2]．

薬物候補化合物の選定には大きく，薬効を示すかどうかという「有効性」，また薬物が適切な速度で代謝・排泄されるかどうかという「安全性」の 2 つの問題が存在している．この中でも「安全性」に着目し，化合物の物理化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が様々なグループで行われてきている [3], [4]．クリアランス経路とは，ヒト体内

[1] 東京工業大学 大学院情報理工学研究科 計算工学専攻
Graduate School of Information Science and Engineering, Tokyo Institute of Technology
[2] 東京工業大学 情報生命博士教育院
Education Academy of Computational Life Sciences, Tokyo Institute of technology
[a)] yanagisawa@bi.cs.titech.ac.jp

において薬物がどのように代謝・排泄されるかという経路のことを指し，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ヒトが体外に排出することのできない化合物を薬物候補化合物から除外することができる．

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も複数の研究グループによって行われており，グルクロン酸抱合（UDP-glucuronosyl transferase, UGT）を用いたクリアランスを機械学習を用いて予測する手法 [3] や 3 種類のシトクロム P450（Cytochrome P450, CYP）による薬物代謝を機械学習を用いて予測する手法 [4] などが提案されているが，これらの研究は単一のクリアランス経路，ないしは単一のカテゴリに分類される複数のクリアランス経路に対する予測であった．

これに対し，我々は複数のカテゴリに分類される複数のクリアランス経路に対する予測を行ってきており，これまでに 3 種類の CYP，および腎排泄（Renal），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）の 5 種類の経路を対象として矩形領域法 [5]，サポートベクターマシン（Support Vector Machine, SVM）[6]，ブースティング [7] など様々な教師付き学習手法を用い，また予測の統合 [8] といった手法も提案してきた．しかし年本ら [6] による予測は，**表 1** に示した通り OATP のクリアランス経路については実用的な予測精度が得られておらず，f 値の平均は 0.536 となっている．

表 1　先行研究における予測精度
Table 1　Result of previous work

	Precision	Recall	予測精度 (f 値)	正例数
CYP2C9	0.600	0.500	0.545	52
CYP2D6	0.857	0.333	0.480	12
CYP3A4	0.824	0.808	0.816	18
Renal	0.732	0.732	0.732	41
OATP	1.000	0.056	0.105	18

この先行研究では，クリアランス経路についてラベル付けされた 141 個の化合物の情報から 5 クラスの予測システム作成を試みているが，一般的にこのデータセットは学習に十分な大きさを持ったものであるとはいえない．一方，化合物の公開データベースである ZINC[9] や DrugBank[10] には大量の化合物の構造情報が登録されており，特徴量の計算も数千個の化合物に対して数分のオーダーで完了する．このようにラベルありデータが少なく，ラベルなしデータを大量に入手することが容易である場合，半教師付き学習（Semi-supervised learning）が適していると考えられる．半教師付き学習とは，1960 年代に Scudder によって提案された self-training[11] にその源流を持ち，正解情報（ラベル）がある「ラベルありデータ」に加え，正解のわかっていない「ラベルなしデータ」を訓練データセットとして利用することで予測精度を向上させる学習法全般を指す．co-Training[12] や transductive フレームワーク [13] など様々な半教師付き学習手法が提案されており，バイオインフォマティクスの分野でも用いられている学習手法である [14]．そこで本研究では半教師付き学習を用いたクリアランス経路予測を行った．

2. 半教師付き学習によるクリアランス経路予測

2.1 データセット

本研究では Kusama ら [5] によって用いられたデータセットを拡張した，9 種類のクリアランス経路についてラベル付けされた 240 種類の化合物の情報を用いた．このデータセットは各化合物の電荷（正電荷および負電荷），血漿タンパク質非結合率，分子量，分配係数の 4 種類の物理化学的特徴量と，薬物クリアランス経路情報をまとめたものである．なお，物理化学的特徴量は薬剤開発の専門家の知見に基づき選定されたものであり，すべて preADMET v2.0（Bioinformatics & Molecular Design Research Center, South Korea）で計算された値を利用する．

また，本研究では予測する対象のクリアランス経路を {CYP1A2, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP3A4, Renal, OATP, UGT} の 9 種類とし，これらを予測する．1 つの薬物化合物について実験的にクリアランス経路を 9 種類確かめることはコストが非常に高く，データセットの拡張は限定的なものとなっている．

表 2　各クリアランス経路の正例数
Table 2　The number of positive data

クリアランス経路	正例数
CYP1A2	18
CYP2C8	5
CYP2C9	17
CYP2C19	10
CYP2D6	25
CYP3A4	79
Renal	69
OATP	18
UGT	27

表 2 にそれぞれのクリアランス経路に対する正例数を示す．複数のクリアランス経路を経て体外に排出される化合物も存在するため，正例数の総和は 240 より大きい．

また，ラベルなしデータとして，公開データベースの 1 つである ZINC[9] の ZINC drug database subset にある化合物セットを利用し，これらについても特徴量を計算した．この化合物セットは 2924 個の化合物からなるが，そのうち平面構造が同じもの（鏡像異性体など）やラベルありデータとの重複を除外し，本研究では 1390 個の化合物

からなるデータセットとなっている．ZINC drug database subset に存在している化合物は薬剤として認められているため，ヒト体内の何らかのクリアランス経路で代謝・排泄されるものであるが，本研究が対象とするクリアランス経路で代謝・排泄されるとは限らないという点には注意が必要である．このデータセットを以後ラベルなしデータセットと呼ぶこととする．

2.2 既存手法による予測

本研究で用いた化合物は先行研究とほぼ変わらないが，物理化学的特徴量の計算ソフトが変更されたため，特徴の値が変化している．このため，同条件下での比較を実現することを目的として，先行研究で提案されている手法のうち，最も高い精度を示したサポートベクターマシン（SVM）を用いて本研究のデータセットを学習し直した．

また，このデータセットを用いた場合の精度の測定についてはラベルありデータの数だけ交差検定法を行う Leave-one-out Cross Validation を用い，精度は f 値で評価するものとする．f 値とは，正例のうち正しく予測できた割合である再現率（recall），正例と予測されたものの中で正しく正例であったものの割合である適合率（precision）の調和平均であり，その値が高いほど判定性能が「総合的に」高いことを示している．

表 3　SVM を用いた結果
Table 3　Result using SVM

	予測精度 (f 値)
CYP1A2	0.456
CYP2C8	0.500
CYP2C9	0.455
CYP2C19	0.340
CYP2D6	0.658
CYP3A4	0.766
Renal	0.743
OATP	0.744
UGT	0.421
Average	0.565

既存手法（SVM）を再現して得た薬物クリアランス経路予測の精度を**表 3** に示す．この表からわかるように，CYP2C19 や UGT などの予測精度が他の経路に比べて低くなっている．また，予測精度の平均は先行研究と同程度であり，改善の余地が見込まれる．

2.3 transductive SVM

本研究では半教師付き学習手法として transductive SVM[15] を用いた．この半教師付き学習手法は Joachim が提案した手法で，SVM に Vapnik らによる transductive フレームワーク [13] を適用したものである．そのため，教師付き学習である SVM との予測精度の比較を行いやすい．

以下に transductive SVM の学習アルゴリズムについて示す．

(1) ラベルありデータをもとに Support Vector（以下 SV）をつくり，ラベルなしデータを全て分類する
(2) SV を作る際のペナルティパラメータを設定する
ただし，ラベルなしデータを誤識別した場合のペナルティはラベルありデータを誤識別した場合のペナルティよりも小さくする
(3) ラベルありデータとラベルなしデータ，全てのデータを利用して SV を作り，再度ラベルなしデータを全て分類する
(4) ラベルなしデータの分類結果が変化しなくなるまで (3) の作業を繰り返す
(5) ラベルなしデータの分類結果が変化しなくなったらラベルなしデータへのペナルティを強化して再度 (3), (4) の作業を繰り返す
(6) ラベルなしデータとラベルありデータのペナルティが同等になったら学習を終了し，そのときの SV を学習結果とする

この手法は反復回数が非常に多く，また SVM そのものも比較的学習に時間を要するため，数千個のデータに対して用いるのが実用限界であると言われている [16]．一方，SVM の予測精度は高いため，本研究のようにデータ数が少ない場合には適切であると考えられる．

SVM のカーネルには線形カーネルや多項式カーネルなど様々なカーネルが存在するが，今回は RBF カーネルを用いた．RBF カーネルのパラメータ γ およびソフトマージンのコストパラメータ C は

$$2^{-15} \leq \gamma \leq 2^{5} \quad (1)$$

$$2^{-5} \leq C \leq 2^{15} \quad (2)$$

で大域的にパラメータチューニングを行い，ここで得られた値を γ_0, C_0 として

$$\gamma = \{\gamma_0 \cdot 2^{-3}, \gamma_0 \cdot 2^{-2.75}, \cdots, \gamma_0 \cdot 2^{2.5}, \gamma_0 \cdot 2^{2.75}\} \quad (3)$$

$$C = \{C_0 \cdot 2^{-3}, C_0 \cdot 2^{-2.75}, \cdots, C_0 \cdot 2^{2.5}, C_0 \cdot 2^{2.75}\} \quad (4)$$

の $24 \times 24 = 576$ 通りを網羅的に探索することで詳細なパラメータチューニングを行う．なお，本研究では transductive SVM の実装として SVMlight を利用した．また，比較対象として教師付き学習を行う場合は，パラメータチューニングの手法などをそのままにして学習手法そのものの比較を行うために SVM を利用し，この実装も SVMlight を利用している．

2.4 半教師付き学習の評価

まず，半教師付き学習が今回の予測問題に対して有用であるかを評価するため，ラベル付きデータのラベルの一部

を隠すことで，以下の 4 通りの教師付き学習および半教師付き学習用のデータセットを作成し，それぞれの精度の確認を行った．なお，(1), (2), (3) についてはデータの選択によって精度が変化してしまうことが考えられるので，本研究ではランダムサンプリングを行い，10 回の精度の測定を実行，その平均値を利用した．

(1) 240 個のデータセットのうち 140 個を選択して教師付き学習で実行
(2) 240 個のデータセットのうち 140 個を選択し，ラベルなしデータセットから化合物を 100 個追加して半教師付き学習で実行
(3) 240 個のデータセットのうち 140 個はそのまま，100 個はクリアランス情報を隠し，半教師付き学習で実行
(4) 240 個のデータセットを全て利用して教師付き学習で実行

精度の測定には 2.2 節と同様に Leave-one-out Cross Validation を用い，その f 値で評価する．

表 4 半教師付き学習の評価
Table 4 Valuation of Semi-supervised learning

	予測精度 (f 値)			
	(1)	(2)	(3)	(4)
CYP1A2	0.446	0.424	0.434	0.456
CYP2C8	0.193	0.251	0.291	0.500
CYP2C9	0.483	0.501	0.472	0.455
CYP2C19	0.456	0.447	0.451	0.340
CYP2D6	0.624	0.651	0.641	0.658
CYP3A4	0.747	0.751	0.755	0.766
Renal	0.733	0.729	0.736	0.743
OATP	0.729	0.721	0.766	0.744
UGT	0.425	0.440	0.408	0.421
Average	0.537	0.546	0.550	0.565

表 4 より，140 個のデータのみを用いた教師付き学習 (1) よりも 140 個のラベルありデータと 100 個のラベルなしデータを用いた半教師付き学習 (2), (3) のほうが予測精度が高いことがわかる．この結果から，半教師付き学習はこの予測問題に対し十分に有用であると言える．

2.5 半教師付き学習を用いた予測

ラベルありデータの一部を隠すなどして行った半教師付き学習の検証より，本研究が対象としている予測問題に半教師付き学習は有用であるということが示された．そこでラベルなしデータを ZINC などの公的データベースから追加することで予測精度の改善を試みた．

ラベルありデータは 2.1 節で示された 240 個のまま，ラベルなしデータセットからランダムに 100，200，400，800 の 4 通りの数を選択し，半教師付き学習に用いることで予測精度の変化を確かめた．ラベルなしデータの選択に関しては 2.4 節と同じく 10 回のランダムサンプリングを行い，それぞれ性能測定をした平均を予測精度とする．**表 5** に結果を示す．

表 5 ラベルなしデータ数と予測精度の関係
Table 5 Relationship between Unlabeled-data and precision

	予測精度 (f 値)				
ラベルなしデータ数	0	100	200	400	800
CYP1A2	0.468	0.426	0.427	0.389	0.365
CYP2C8	0.500	0.462	0.452	0.469	0.483
CYP2C9	0.417	0.433	0.437	0.443	0.454
CYP2C19	0.333	0.351	0.374	0.359	0.367
CYP2D6	0.667	0.658	0.674	0.662	0.694
CYP3A4	0.743	0.746	0.744	0.742	0.742
Renal	0.731	0.732	0.729	0.726	0.726
OATP	0.650	0.653	0.665	0.654	0.632
UGT	0.386	0.400	0.415	0.436	0.424
Average	0.544	0.540	0.546	0.542	0.543

ラベルなしデータ数の増加に応じた予測精度の向上が期待されたが，この結果を見ると，表 4 に示されたほどの予測精度変化は見られず，ほぼ横ばいで推移していることがわかる．

3. 特徴量の追加

本研究で用いた特徴量は 5 次元と非常に少なく，ラベルありデータが 240 個存在することはその次元数にとってはすでに十分であることが考えられる．この場合，予測精度の低い原因は今回用いている特徴量の次元が少なく，データを分離するためには不十分であることが考えられる．計算可能な化合物の特徴量は大量に存在するため，この問題を検証するために特徴量の追加を行い，再度半教師付き学習を行った．

一般に，特徴量を追加することでデータの表現は複雑になり判別能力は高まるが，薬学的な観点からすると物理化学的に説明が難しい特徴量が大量に追加されることは意味の解釈の問題から決して好ましいとは言えない．したがって特徴量の追加は最小限に抑える必要がある．本研究では貪欲法を用いることで最小限の特徴量の選択/追加を行った．

3.1 特徴量の計算

今回選択の対象となる特徴量の計算には Canvas v1.7.014（Schrödinger, USA）を利用した．Canvas は化合物の平面構造から 757 種類もの特徴量を計算することができる．

また，齊藤らによってクリアランス経路予測に有用であると示されたドッキング計算を用いて [17]，Protein Data Bank（PDB）[18] から取得した 15 種類の CYP タンパク質

の立体構造に対して AutoDock[19], Glide(SP モード，XP モード)[20] の 2 種類 3 モードのドッキングツールを利用し，計 45 種類の結合自由エネルギーを算出した．このエネルギー値も特徴選択の対象とする．

3.2 特徴量の選択

最も高い予測精度を出す特徴量の組を求めるためには，すべての特徴量の組み合わせについて予測を行う必要がある．しかし本研究において候補となる特徴量は 802 種類存在し，これらの組み合わせは $2^{802} \approx 4 \cdot 10^{240}$ と網羅的に探索を行うのは不可能である．そこで本研究では貪欲法に基づく特徴選択 [21] を行い，近似的に最適な特徴量の組み合わせを探索した．

また，候補となる特徴量は 802 種類存在しているが，これだけ多くの選択肢が存在すると偶然データセットに適合して予測精度を高めるようなものが存在する可能性がある．したがって，802 種類のランダムな値で構成されたダミー特徴量を構成し，これを用いて有意に予測精度が向上しているかどうかを確かめることが必要である．本研究では候補特徴量 802 種類に対して，化合物とその値との対応をシャッフルすることでダミー特徴量を構成した．貪欲法による特徴選択の手順を以下に示す．

(1) 候補特徴量・ダミー特徴量をそれぞれ 1 つだけ特徴量に追加した $802 \times 2 = 1604$ パターンで教師付き学習によるクリアランス経路予測を行い，f 値を算出する
(2) 最も f 値を高くする特徴量が候補特徴量である場合には，その特徴量を追加し，(1) に戻る
(3) 最も f 値を高くする特徴量がダミー特徴量である場合には，特徴選択を終了する

なお，ここにおける教師付き学習は LIBSVM[22] を利用している．LIBSVM は transductive SVM を実行することはできないが，比較的高速に学習を進めることができる．

以上の手法で追加された特徴数，およびその予測精度を**表 6** に示す．

表 6 特徴選択の結果
Table 6 The result of feature selection

	追加前の予測精度（f 値）	追加特徴数	追加後の予測精度（f 値）
CYP1A2	0.462	4	0.821
CYP2C8	0.500	0	0.500
CYP2C9	0.465	7	0.800
CYP2C19	0.311	2	0.750
CYP2D6	0.658	6	0.840
CYP3A4	0.763	7	0.906
Renal	0.738	13	0.906
OATP	0.640	3	1.000
UGT	0.414	3	0.605
Average	0.550	-	0.792

3.3 追加特徴量を利用した半教師付き学習

特徴量が追加され，それに伴い予測精度が向上した．しかし本来の特徴量の追加目的は半教師付き学習を効果的に行うためである．そこで特徴量が追加された状態で半教師付き学習を行い，予測精度の向上を試みた．

本研究では，特徴量が 3 つ追加され，f 値が比較的低く半教師付き学習を用いることでさらなる精度の向上が望まれる UGT に対し，ラベルなしデータ数が 200 および 800 の場合において各々 1 回ずつの半教師付き学習を行った．結果を**表 7** に示す．

表 7 特徴量の追加を行った状態での半教師付き学習による予測精度（UGT）
Table 7 Semi-supervised learning with additional feature

	予測精度 (f 値)		
ラベルなしデータ	0 個	200 個	800 個
特徴量追加前	0.386	0.415	0.424
特徴量追加後	0.605	0.500	0.500

0, 200, 800 のどのラベルなしデータ数を用いても特徴量追加前にくらべて特徴量追加後のほうが精度が向上している一方，特徴量追加後ではラベルなしデータを加えた半教師付き学習よりもラベルありデータのみを利用する教師付き学習のほうが良い精度でクリアランス経路の予測ができており，期待された結果は得られなかった．

今回用いた手法では，特徴選択に教師付き学習のみを用い，半教師付き学習は用いなかった．このため選択された特徴量が 240 個のラベルありデータに強く依存したものになってしまっていた可能性がある．しかしこの問題を解決するには特徴選択に半教師付き学習を用いることになるが，各特徴量を追加した場合の f 値を計算する際にそれぞれランダムサンプリングが必要となる．例えば 10 回のランダムサンプリングを行う場合には約 1,600 種類の特徴量× 10 回のランダムサンプリング＝約 16,000 回のパラメータチューニング込みの半教師付き学習を 1 つの特徴量を追加するごとに行う必要がある．1 回の学習でも transductive SVM は SVM に比べて最低数倍の計算量を必要とするため，超大規模な並列計算やより高速な半教師付き学習手法を用いるなどの工夫が必要になることが考えられる．

4. 本研究のまとめ

本研究では半教師付き学習を，評価を行った上で薬物クリアランス経路予測に利用することで精度の改善を試みた．ラベルありデータの一部を隠し，140 個のラベルありデータに対してラベルなしデータを追加した場合には精度の改善が見られたが，240 個のラベルありデータに対して ZINC のような公開データベースから取得したラベルなしデータを追加した場合には精度の改善は見られなかった．

また，貪欲法を用いて特徴量を追加し，比較的多数の特徴量から半教師付き学習を用いて薬物クリアランス経路予測を行う手法では，特徴量を追加することで教師付き学習も半教師付き学習についても予測精度を改善することができたが，教師付き学習と半教師付き学習を比較するとラベルなしデータを用いない教師付き学習のほうがより精度が良くなることが示された．これは教師付き学習によって特徴選択を行ったことが一因として挙げられる．

ラベルありデータは実際に薬物クリアランス経路情報を取得することができた化合物であり，最低でもどれか 1 つのクリアランス経路を経るものである．しかし本研究が標的としている現実世界に存在する薬剤の候補として挙げられる化合物は，どのクリアランス経路も経ずに体外に排出されないものも存在している．このようなデータの偏りが選択される特徴量の偏りとなって現れ，予測精度の改善を妨げていると考えることができる．

一方，もし半教師付き学習を用いて同様の特徴選択を行うとすると，transductive SVM は比較的学習に要する計算量が多く，特徴選択を行う上で大きなネックとなる．よって，他の半教師付き学習や transductive SVM の並列化による計算高速化など，様々な工夫を考える必要があると考えられる．

謝辞 本研究は 文部科学省 博士課程教育リーディングプログラム 東京工業大学「情報生命博士教育院」の支援を受けて行われた．

参考文献

[1] “PhRMA New Drug Approvals in 2011”, http://www.phrma.org/
[2] 船津 公人, 佐藤 寛子, 増井 秀行 訳: “ケモインフォマティクス 予測と設計のための化学情報学”, 丸善, 東京, pp.593-pp.611, 2005
[3] Sorich M.J., Miners J.O., McKinnon R.A. *et al.*: “Comparison of Linear and Nonlinear Classification Algorithms for Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms”, *Journal of Chemical Information and Computer Sciences* 43(6), pp.2019-pp.2024, 2003
[4] Hammann F., Gutmann H., Baumann U. *et al.*: “Classification of Cytochrome P450 Activities Using Machine Learning Methods”, *Molecular Pharmaceutics* 6(6), pp.1920-pp.1926, 2009
[5] Kusama M., Toshimoto K., Maeda K. *et al.*: “In silico classification of major clearance pathways of drugs with their physiochemical parameters.”, *Drug Metabolism and Disposition* 38(8), pp.1362-pp.1370, 2010
[6] 年本 広太, 草間 真紀子, 池田 和史 *et al.*: “SVM を用いた薬物クリアランス経路予測システムの開発 -複数経路予測への拡張と外部データによる評価-”, **情報処理学会研究報告** 2010-BIO-20(8), pp.1-pp.8, 2009
[7] 池田 和史, 年本 広太, 草間 真紀子 *et al.*: “ブースティングによる薬物クリアランス経路予測”, **情報処理学会研究報告** 2009-BIO-17(10), pp.1-pp.8, 2009
[8] 堀田 駿, 年本 広太, 池田 和史 *et al.*: “ウェブアプリケーションによる薬物クリアランス経路予測”, **情報処理学会研究報告** 2010-BIO-21(20), pp.1-pp.8, 2010
[9] John J.I., Teague S., Michael M.M. *et al.*: “ZINC: A Free Tool to Discover Chemistry for Biology.”, *Journal of Chemical Information and Modeling* 52(7), pp.1757-pp.1768, 2012
[10] Law V., Knox C., Djoumbou Y. *et al.*: “DrugBank 4.0: shedding new light on drug metabolism”, *Nucleic acids research*, 42(D1), D1091-D1097, 2014
[11] Scudder III H.: “Probability of error of some adaptive pattern-recognition machines.”, *Information Theory, IEEE Transactions* 11(3), pp.363-pp.371, 1965
[12] Blum A., Mitchell T.: “Combining labeled and unlabeled data with co-training.”, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pp.92-pp.100, 1998
[13] Vapnik V.: “Statistical Learning Theory.”, Wiley, 1999.
[14] Röttig M., Medema M.H., Blin K. *et al.*: “NRPSpredictor2 - a web server for predicting NRPS adenylation domain specificity”, *Nucleic Acids Research* 39(suppl 2), pp.362-pp.367, 2011
[15] Joachims T.: “Transductive inference for text classification using support vector machines”, *Proceedings of the Sixteenth International Conference on Machine Learning*, pp.200-pp.209, 1999
[16] Collobert R., Sinz F., Weston J. *et al.*: “Large scale transductive SVMs”, *The Journal of Machine Learning Research* 7, pp.1687-pp.1712, 2006
[17] 齊藤 有紀, 石田 貴士, 関嶋 政和 *et al.*: “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”, **情報処理学会研究報告** 2013-BIO-34(15), pp.1-pp.6, 2013
[18] Berman H.M., Westbrook J., Feng Z. *et al.*: “The Protein Data Bank”, *Nucleic Acids Research* 28(1), pp.235-pp.242, 2000
[19] Morris G.M., Huey R., Lindstrom W. *et al.*: “AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.”, *Journal of Computational Chemistry* 30(16), pp.2785-pp.2791, 2009
[20] Friesner R.A., Banks J.L., Murphy R.B. *et al.*: “Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.”, *Journal of Medicinal Chemistry* 47(7), pp.1739-pp.1749, 2004
[21] Caruana R., Freitag D.: “Greedy attribute selection”, *Proceedings of the Eleventh International Conference on Machine Learning*, pp.28-pp.36, 1994
[22] Cheng C.c., Lin C.J.: “LIBSVM: a libbrary for support vector machines.”, *ACM Transactions on Intelligent Systems and Technology* 2(3), pp.27:1-pp27:27, 2011