2008－BIO－13 (10)
2008／6／26

機械学習を用いた薬物のクリアランス経路予測

年本 広太† 草間真紀子†† 前田 和哉†† 杉山 雄一†† 秋山 泰†

† 東京工業大学 大学院情報理工学研究科 〒152-8552 東京都目黒区大岡山2-12-1
†† 東京大学 大学院薬学系研究科 〒113-0033 東京都文京区本郷7-3-1

あらまし 薬物のクリアランス経路の特定は薬物動態学において重要とされている．そこで各薬物の多次元の物理化学的記述子から，クリアランス経路を予測するシステムを構築した．当システムでは主要な5つの経路ごとにサポートベクターマシン (SVM) を用いてクリアランス経路を学習させ，各予測結果を総合して最も支配的なクリアランス経路を1つ与える．今回準備できた記述子は1089個あるが，その全てを学習の入力として用いると汎化誤差の問題や説明性の低下が生じる．そこで貪欲法や相関係数などを用いた実験により特徴選択を行ったところ，12個前後の少ない記述子数で85%以上の予測精度を得ることができた．

キーワード 機械学習，薬物動態，サポートベクターマシン (SVM)，特徴選択

Prediction of Drug Clearance Pathway with Machine Learning

Kouta TOSHIMOTO†, Makiko KUSAMA††, Kazuya MAEDA††, Yuichi SUGIYAMA††, and Yutaka AKIYAMA†

† Graduate School of Information Science and Engineering, Tokyo Institute of Technology
Ookayama 2-12-1 Meguro-ku, Tokyo, 152-8552 Japan
†† Graduate School of Pharmaceutical Sciences, The University of Tokyo
Hongo 7-3-1 Bunkyo-ku, Tokyo, 113-0033 Japan

Abstract The clearance pathway is one of the important factors to consider the pharmacokinetics of drugs. We have developed a machine learning system of drug clearance pathway for a given drug compound from its physicochemical descriptors. The system is composed of five support vector machines (SVMs), each corresponding to one of five major candidate clearance pathways, and prediction is given by choosing a pathway with largest SVM output. We prepared pathway data for 157 drugs, and 1089 physicochemical descriptors for each of them. However, if we use all the descriptors, we will have over-learning problem and less explainable model. Thus we performed exhaustive feature selection procedure, by a modified greedy algorithm or a correlation coefficient-based method, and our system showed more than 85% prediction accuracy when using 12 selected descriptors.

Key words Machine learning, Pharmacokinetics, Support vector machine(SVM), Feature selection

1. はじめに

現在，新薬の開発には数十億円という莫大な開発費と十数年以上の多大な時間が必要とされている．一方，*in vitro* 実験や動物実験の結果を基にした前臨床試験により選択された医薬品候補化合物の中でも，ヒト臨床試験に到達して初めて，不十分な効果・予期せぬ毒性発現のために開発が中止される薬物が後をたたない．そのひとつの原因として，薬物の血中濃度推移や臓器への分布特性といった薬物動態学的要因が注目されている [1].

薬物の体内動態を予測する上で重要な情報の1つに，クリアランス経路が挙げられる．一般に薬物は，異物解毒系と呼ばれる非常に多様な代謝酵素・トランスポーター群の協調的な機能により体外へと排泄される．クリアランス経路とは，薬物の体内からの消失手段としての，肝臓の代謝酵素による物質変換や，肝臓・腎臓に発現するトランスポーターを介した排出輸送といった，おのおのの素過程のことを意味している [2]．これまでの研究から，薬物によって多様なクリアランス経路が存在することが明らかとなっている．一方，代謝酵素・トランスポーターといったタンパク質と薬物の相互作用により代謝・輸送が実行されることを考慮すると，薬物の化学構造により決定付けられる脂溶性，極性，分子サイズなどの物理化学的特性によっ

てクリアランス経路は支配されていることが想定される．クリアランス経路を試験管内での実験から正確に予測するには，多様な情報が必要であることから，創薬初期段階のような候補化合物が多数存在するようなステージでは難しいと考えられる．

そこで本研究では薬物の物理化学的記述子からクリアランス経路を予測するシステムを構築した．当システムでは主要な5つの経路ごとにSVM(サポートベクターマシン)[3]を用いてクリアランス経路を学習させ，各予測結果を総合して最も支配的な経路を1つ与える．また今回準備できた記述子は1089個あるが，その全てを学習の入力として用いると汎化誤差の問題や説明性の低下が生じる．そこで貪欲法や相関係数などを用いた実験により特徴選択を行うことで，入力次元を減らしかつ精度の高いシステムの構築を試みた．このとき，5つの予測機械に同一の記述子の組み合わせを用いるアプローチと，個々の予測機械に異なる記述子を用いるアプローチが考えられる．今回は前者の考えに基づいて記述子を選択することにした．

2. クリアランス経路

薬物の異物解毒は，多様な代謝酵素・トランスポーター群の協調的な作用によっていることから，きわめて多くのクリアランス経路が存在する．本研究では，それらの中で多くの薬物の体内からの消失において比較的主要な経路である，3種類のcytochrome P450 (CYP3A4, CYP2C9, CYP2D6)を介した代謝，OATP(Organic Anion Transporting Polypeptide)を介した肝取り込み，腎排泄，の合計5つのクリアランス経路についてのみ予測を行うことにした．これまで薬物がどのCYPを介して代謝されるかを予測する先行研究はさまざま存在するが，本研究では，腎排泄や肝取り込みトランスポーター等，分子種を超えて代謝・輸送を含めてクリアランス経路の予測を行った点に新規性がある．また各薬物について，クリアランス経路は単一ではなく複数存在することが少なくないが，体内からの薬物消失に占める寄与が最も大きなクリアランス経路において，薬物間相互作用や遺伝子多型によって機能変動がおこった際に，薬物動態に最も多大な影響が出ること，さらに今回は結果の正しさを評価する事も優先して，最も支配的なクリアランス経路を1つ予測するシステムを構築した．以下，今回行なったクリアランス経路による分類カテゴリーについて説明する．

腎排泄とは，腎臓からの未変化体としての尿中排泄として定義しており，以下このカテゴリーをRenalと命名する．

次にCYPとは，薬物の第I相代謝反応の大部分を担うCytochrome P450と呼ばれる酵素群のことである[4]．今回は，その中でも特に多くの薬物の代謝に関与すると考えられる3つのCYP(CYP3A4,CYP2C9,CYP2D6)について考えることにした．CYP3A4は医薬品代謝の約50%に関与しており，また肝で全CYP量の約30%を占める代表的なCYPである．またCYP2C8,CYP2C9,CYP2C18,CYP2C19の4分子種は相同性が80%以上とよく似ており，またその総量は肝で約20%を占める．この中で最も含量が多いCYP2C9を代表として用いた．CYP2D6は全肝CYP量の約2%しか存在しないが，医薬品全体の代謝に占める寄与は大きいことが知られている．以下表中では“CYP”の表記は省略することにする．

経路名 (単数)	データ数
Renal	40
3A4	58
2C9	13
2D6	16
OATP	9
合計	136

経路名 (複数)	データ数
Renal&3A4	1
Renal&OATP	4
3A4&2C9	4
3A4&2D6	9
3A4&OATP	3
合計	21

表1　各データに記載されているクリアランス経路の総数

最後に，OATPカテゴリーには，有機アニオントランスポーターの一種で，肝臓の血管側に発現し，種々の薬物の肝取り込みに関与するOATP1B1, OATP1B3などにより肝臓に取り込まれ，かつその過程が肝クリアランス全体の律速過程となっていることが想定される薬物群を入れるようにした．HMG-CoA還元酵素阻害薬やアンジオテンシンII受容体拮抗薬など，アニオン性を示す多様な薬物がこのカテゴリーに属している．

3. データの統計処理と記述子

3.1 データの統計処理

本研究の実験に使用した化合物のデータセットは，市場に流通している医薬品に関するもので共著者の杉山雄一教授らのグループが集めたものである．データの総数は157個である．各データは，解であるクリアランス経路と，各薬物の分子量(MW)，n-オクタノール/水(pH 7.4)分配係数(LogD)，血漿中非結合型分率(fu)，そして電荷(Charge)の4つの特徴量が記述されている．この4つの特徴量は，薬物動態学の専門家がクリアランス経路の判別に有効だと考えたものである．各クリアランス経路におけるデータの総数を表1に示す．表1にあるとおり，本データセットの一部には主要なクリアランス経路を1つに特定することが難しいために，複数のクリアランス経路が定義されているものがある．そのようなデータは，本研究の実験では学習用のデータとしてのみ用い実際の評価には用いないことにする．

また，今回の実験において各クリアランス経路における正例と負例の個数では負例のほうが数が多く，クリアランス経路によってはかなりの偏りがある．その偏りを補正するために，(正例) : (負例) = 100個 : 100個となるようにオーバーサンプリングおよびアンダーサンプリングを行うことにする[5][6]．これらはランダムな試行であるため，本実験はこれを100回繰り返し，それぞれのデータにおいてSVMを用いて学習を行う．テストデータの出力値は，各サンプリング集合で学習させたSVMの出力値の平均値を取ることにする．

また各SVMの予測精度を評価するために，今回は交差確認法(cross validation)の一種であるLeave-one-out法を用いた．誤差の指標としてはf値(f-measure)を用いる．f値は再現率(recall)と適合率(precision)の調和平均の値で，以下のように求めることができる．

$$\text{f値} = 2 \cdot \frac{\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}}$$

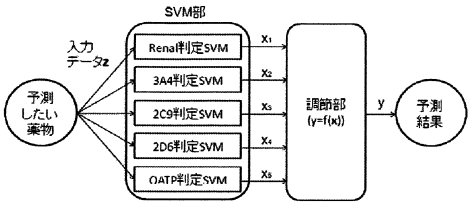

図 1　クリアランス経路予測システムの概要

f 値が高ければ高いほど判定精度が総合的に高いということを表している．またシステム全体の予測精度の評価の際には，二項分類ではないため f 値を計算することが出来ない．そこで式 (1) で表すことのできる正確度 (accuracy) を用いることにする．

$$\text{正確度 (accuracy)} = \frac{\text{正しく予測できたデータの総数}}{\text{データの総数}} \quad (1)$$

3.2 記 述 子

記述子 (Descriptor) とは，化合物の化学構造から計算することができるパラメータの総称である．例えば分子量や炭素原子数，二重結合数などの化学構造そのものを表すものから親水性，極性といった化合物の物理化学的特性を表す量まで全てが記述子に含まれる．記述子は現在まで数多く定義され，またその計算方法も確立されているので瞬時に記述子を求めるソフトウェアが存在する．

今回記述子の計算に使用したソフトウェアは韓国の Bioinformatics & Molecular Design Research Center(BMDRC) が開発した PreADMET ver2.0 である [7]．これにより 1082 個の記述子を得ることができた．しかしこの記述子の中には，我々が用意したデータセット内データでは一部数値が計算できなかったものや，半分以上の化合物のデータにおいて同じ値をとる記述子も存在した．そのような記述子は学習の入力に使用することのは適当ではないので除外する．その結果，合計で 681 個の記述子を得た．

4. 予測システムの構築

SVM は二値分類器なので，多値分類を行うために one-versus-rest(1vR) 法を用いた [8]．one-versus-rest 法は，複数のクラス $\{C_1, C_2, \cdots, C_n\}$ に対しあるカテゴリ C_i を正例とし，それ以外を負例とした学習を全ての i に対して行う．今回の場合，クラスは 5 つ存在するので 5 つの SVM を用いて予測を行う．なお各 SVM のカーネルは Gaussian Kearnel を用い，SVM のプログラムとして SVM^{light} を用いた [9]．

今回実装した予測システムは図 1 のようにシステムを SVM 部と調節部の 2 つに分かれている．SVM 部は，各クリアランス経路ごとに SVM のパラメータ (C, σ) を様々に変化させ，実際に学習を行い判別機械を作成する．調節部は，パラメータ変化によってできた判別機械をクリアランス経路ごとに組み合わせて，最も精度よく予測を行うような判別機械の組み合わせを求める．なお今回の実験では各 SVM の出力値のうち，最大値を出力した SVM の表わすクリアランス経路を予測結果とする．

本研究では SVM の予測精度を比較するため以下の実験を行った．

実験 A)　基本的な記述子のみを使用した予測

3.1 で述べたとおり，薬物動態学の専門家がクリアランス経路の判別に有効だと考える 4 つの特徴量 (MW, LogD, fu, Charge) のみを使用して予測を行う．ただし電荷には負に帯電 (Anion)，正に帯電 (Cation)，帯電なし (Neutral)，両極性 (Zwitter) の 4 つの状態が存在し，数値が定義されているものではない．そこで各カテゴリに対してビット表現を用いることにし，全部で 7 記述子による予測を行った．

実験 B)　全ての記述子を使用した予測

上記 7 個の記述子に加えて，681 個全ての記述子を使用して計 688 記述子で予測を行う．

しかし 688 個全ての記述子を学習に使用しても，予測とはあまり関係のない特徴量を含んでしまっていると考えられる．人間の理解のしやすさの面からも特徴選択を行って記述子を絞ることが有効である．そこで最初に使用した基本的な 7 記述子に，有効であると判断された記述子を 1 つずつ加えて予測精度の向上を測定した．記述子を加える順番としては，以下の 2 つの異なる実験を行った．

実験 C)　段階的貪欲アルゴリズムによる特徴選択

貪欲アルゴリズムは最も単純な近似解法の 1 つとして知られているが，今回の実験ではサンプリングによる繰り返しと Leave-one-out 法による交差確認法に相当な時間を要する．そのため現実的な時間内で近似解を求めることが難しい．そこで記述子を加えて予測を行った際に，予測システム全体としての予測精度が小さかった半分程度の記述子に関しては，以降の記述子選択の候補から除外することにした．この手法をここでは段階的貪欲アルゴリズムと呼ぶことにする．

段階的貪欲アルゴリズムを用いた予測システムの構築法は以下の通り．

（1）　$X :=$ { 基本の 7 記述子 }, $Y :=$ { 追加候補記述子 } とする．

（2）　すべての $\mathbf{y} \in Y$ について，$X \cup \mathbf{y}$ を入力記述子として予測システムを構築する．

（3）　最大の正確度を出した予測システムに入力として加えた $\mathbf{y}$ を $\mathbf{y}'$ とする．

（4）　$X := X \cup \{\mathbf{y}'\}, Y := Y/\{\mathbf{y}'\}$ とする．

（5）　集合 Y のうち，正確度が小さかった $\frac{|Y|}{2}$ 個の記述子を Y から除く．

（6）（2）～（5）を $Y = \phi$ となるまで繰り返し，最も予測精度が高かった記述子の組み合わせを求める．

このとき，（3）や（5）における正確度とは，予測システム全体の予測精度のことであり個々の判別機械の予測精度とは無関係である点に注意されたい．

実験 D)　相関係数による特徴選択

2 つのベクトル $\mathbf{x}$,$\mathbf{y}$ の類似性を測る指標の 1 つに式 (2) で表わされる相関係数がある．

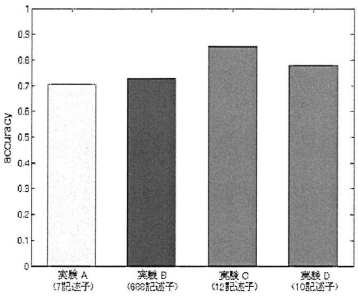

図 2　実験別の予測システムの精度

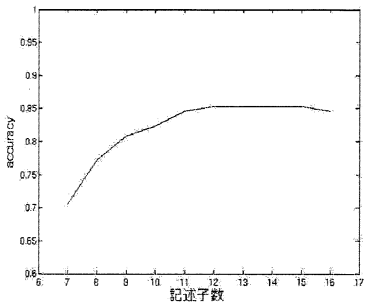

図 3　段階的貪欲法による記述子数別の予測精度

$$\text{相関係数}:\sigma_{\mathbf{xy}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \quad (2)$$

相関係数は $-1 \leqq \sigma_{\mathbf{xy}} \leqq 1$ の範囲の実数値をとり，$\sigma_{\mathbf{xy}} > 0$ のときを正の相関，$\sigma_{\mathbf{xy}} < 0$ のときを負の相関という．また $|\sigma_{\mathbf{xy}}| \approx 1$ のとき類似度が高い．

相関係数を用いて記述子を追加する順番を決定するアルゴリズムは以下の通り．

（1） $X :=$ { 基本の 7 記述子 }, $Y :=$ { 追加候補記述子 } とする．
（2） すべての $\mathbf{x} \in X, \mathbf{y} \in Y$ について $|\sigma_{\mathbf{xy}}|$ を計算する．
（3） 計算された $|\sigma_{\mathbf{xy}}|$ の最小値のときの $\mathbf{y}$ を $\mathbf{y}'$ とする．
（4） $X := X \cup \{\mathbf{y}'\}, Y := Y/\{\mathbf{y}'\}$ とする．
（5）（2）～（4）を繰り返し，出力された $\mathbf{y}'$ の順に記述子を加えて，学習を行う．

このアルゴリズムは記述子を選出するにあたり実際に学習を行っていない．そのため，先に示した段階的貪欲アルゴリズムと比べてきわめて高速である．なおこの 2 つの特徴選択では，前述のとおり全てのクリアランス経路について同じ記述子の組み合わせを使用している．

上記以外にも相関係数をタニモト係数に置き換えてた手法や，kNN(k-nearest neighbor) 法を用いて予測 [10] を行ったが，よい精度を得ることができなかったので割愛する．

5. 結果と考察

5.1 システム全体の予測精度比較

はじめにシステム全体の予測精度を比較する．前述の各実験 A～D について式 (1) で定義される正確度の値を図 2 に示す．なお実験 C と実験 D では記述子を順に追加したとき，最もシステム全体の予測精度が高かったときの結果を示すことにする．

図 2 から，基本の 7 記述子 (実験 A) や全 688 記述子を使用する (実験 B) よりも，基本の 7 記述子に適当な記述子を加えることで (実験 C,D) 予測精度が向上することが判明した．

また，段階的貪欲法 (実験 C) と相関係数を用いた特徴選択 (実験 D) とでは，段階的貪欲法のほうが予測精度の向上が大きかった．しかし，段階的貪欲法は約 128CPU を使用して 2 週間程度の時間を要した一方で，相関係数の実験では同 CPU 数あれば 1 日以内で簡単に実行することができた．各実験の実行時間の詳細な比較については割愛するが，段階的貪欲法は他の 3 つの実験と比べて実行時間にかなりの問題がある．しかしこれは更なる並列処理を施すことである程度は改善できる問題であると考える．

次に段階的貪欲法 (実験 C) において，記述子を 1 つずつ加えていったときの正確度の変化を図 3 に示す．この図から，記述子数が 12 個以下においては予測精度が単調に増加していることがわかった．しかし記述子数が 13 個以上になっても，予測精度が大きく上昇することはなくほぼ一定であった．最も予測精度が高いのは記述子数が 12 個，13 個，14 個または 15 個のときであるが，以後全ての場合において記述子数が 12 個の場合の実験結果を示すことにする．

さて，最も予測精度がよかった段階的貪欲法 (実験 C) の予測結果の詳細を表 2 に，比較資料として基本の 7 記述子 (実験 A) での予測結果の詳細を表 3 に示す．表 2 と表 3 の行成分はデータに記載されていた実際のクリアランス経路を表している．列成分は予測システムが出した予測結果を表している．たとえば表 2 の (行, 列)=(3A4(解), Renal(予)) の数値 3 は，実際の解

	Renal(予)	3A4(予)	2C9(予)	2D6(予)	OATP(予)
Renal(解)	**34**	0	2	1	3
3A4(解)	3	**51**	2	2	0
2C9(解)	1	1	**11**	0	0
2D6(解)	0	4	0	**12**	0
OATP(解)	0	1	0	0	**8**

表 2　段階的貪欲法による予測結果の詳細

	Renal(予)	3A4(予)	2C9(予)	2D6(予)	OATP(予)
Renal(解)	**27**	2	3	5	3
3A4(解)	5	**42**	1	9	1
2C9(解)	1	1	**9**	0	2
2D6(解)	3	0	0	**13**	0
OATP(解)	1	1	1	0	**6**

表 3　基本の 7 記述子による予測結果の詳細

は CYP3A4 であったが，予測システムは Renal と予測したものが全 136 個のデータ中に 3 個あったことを表している．

表の対角成分は正しい予測を行えたデータ数を表している．基本の 7 記述子のみを使用した場合 (実験 A) に比べて，段階的貪欲法による特徴選択を行った場合 (実験 C) のほうが対角成分が大きく増加していることがわかる．段階的貪欲法では対角成分の総和は 116 となり，予測精度は図 2 で示したとおり $116 \div 136 \approx 0.85$ となっている．これより，本研究では予測システムの予測精度を 85% 程度まで上昇させることができた．

また表 3 を見ると，この予測システムでは起こりやすい間違い (CYP3A4 と CYP2D6 など) と間違いが全く起こらない箇所 (CYP2C9 と CYP2D6 など) が顕著に表われている．CYP3A4 は他の CYP 群酵素と比べて基質特異的でなく，多くの薬物の代謝経路に含まれている．最も寄与の大きなクリアランス経路を 1 つ予測するという今回の実験では，誤りとして表示されてしまうが現実的な観点からみると非常に興味深い誤りであるともいえるであろう．しかしこのような間違いも表 2 ではほとんどが改善されている．

5.2 クリアランス経路ごとの精度比較

ここでは例として Renal と CYP3A4 の結果についてのみ示す．まず各実験による f 値の差を図 4 と図 5 に示す．

図 4 と図 5 では各実験 A〜D について，システム全体の予測精度が最良となったときに採用された Renal, CYP3A4 の各判別機械の f 値を示している．このため個々の判別機械のみでみ

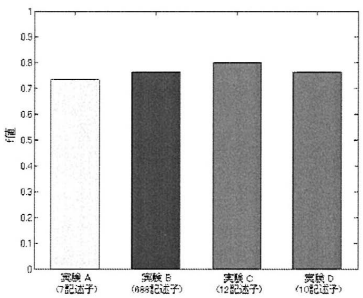

図 4　Renal における実験別の f 値

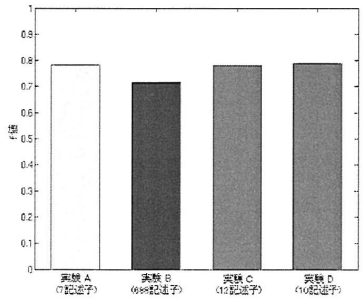

図 5　CYP3A4 における実験別の f 値

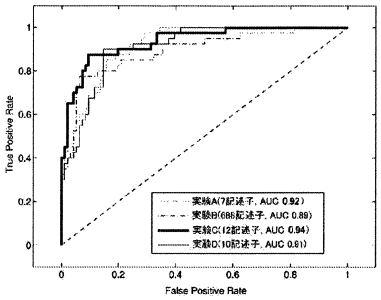

図 6　Renal における実験別の ROC 曲線

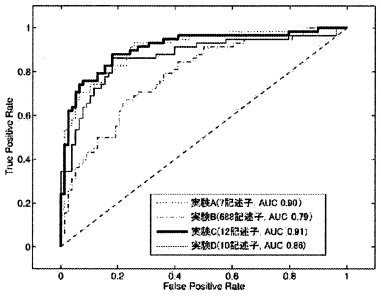

図 7　CYP3A4 における実験別の ROC 曲線

れば，図で示した値より大きな f 値を持つパラメータ (C, σ) が存在することに注意されたい．図 4 より，Renal においては段階的貪欲法 (実験 C) の判別機械が最も f 値が高い．しかし図 5 を見ると，全記述子を使用した場合 (実験 B) 以外の 3 つのどの場合についても f 値に大きな違いは見ることができなかった．よって CYP3A4 の判別機械は，記述子を加えても f 値が大きくならず，あまり変化がないということがわかる．同時に，システム全体の予測精度の良さと個々のクリアランス経路判別機械の予測精度の良さは必ずしも一致しないことがわかる．

そこで，f 値以外の比較として ROC 曲線による比較を行った．ROC 曲線は，分類器 (SVM) の閾値を変化させながら縦軸に True Positive Rate，横軸に False Positive Rate をとった曲線のことである．理想的な分類器では ROC 曲線は 3 点 (0,0),(0,1),(1,1) を通る折れ線である．また ROC 曲線は分類器の視覚的な比較を行う道具だが，数値的な比較を行うために AUC(Area Under the Curve) を用いることが多い．AUC は ROC 曲線と横軸に挟まれた面積で，最良の場合その値は 1.0 となる．

さて，Renal, CYP3A4 における ROC 曲線をそれぞれ図 6 と図 7 に示す．図 6，図 7 をみると共に段階的貪欲法 (実験 C) のときに AUC の値が最大である．しかし Renal のときは他の実験法でも AUC の値は高く，CYP3A4 に関しては段階的貪欲

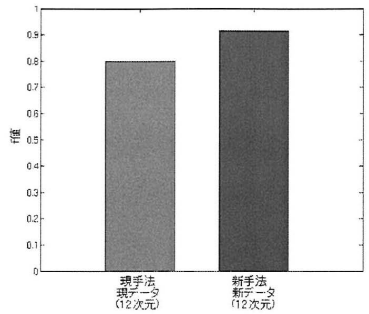

図 8　経路ごとに記述子を変える効果 (Renal)

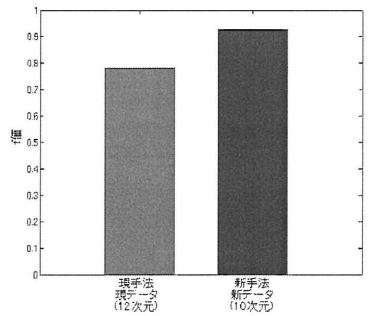

図 9　経路ごとに記述子を変える効果 (CYP3A4)

法 (実験 C) と基本の 7 記述子を使用した場合 (実験 A) と比べてあまり差が見られない．しかし本研究において，予測システムの最終的な出力は各クリアランス経路判別機械の出力の最大値である．そのため評価においては AUC の値を比較するよりも，ROC 曲線の False Positive Rate が小さい値のとき (横軸の値が 0 に近いとき) に True Positive Rate がどの程度高いかを比較することが現実的に有効な比較法であるといえる．この観点から見ると，図 6，図 7 ともに段階的貪欲法において，最も True Positive Rate の立ち上がりがよいことがわかる．

6. 今後の課題

6.1 過学習の回避

本研究において特徴選択を用いて記述子を加えていったとき，今回のデータセットに対して入力次元が高くなりすぎて過学習をしている可能性が完全には否定できない．過学習があると，2 つのデータが近傍に位置するとき，もう片方のデータの情報のみを用いて予測を行っている可能性がある．これはデータ間の相関係数やタニモト係数を算出し，あまりにも近接したデータがあるときは片方を除外するといった措置をとれば回避することができる．しかし実際にクリアランス経路の予測を行うべき化合物の母集団における記述子の分布はまだ得られていないので，この問題を現時点で完全に回避することは難しい．

6.2 クリアランス経路ごとの異なる特徴選択

本研究では特徴選択の際，全てのクリアランス経路の判別機械が同じ記述子を用いて判別を行っている．しかし個々のクリアランス経路ごとに選出する記述子を変えることにより [11]，ある特定のクリアランス経路を判別する際にのみ非常に有効な記述子を活用できるという利点が考えられる．

この考えに基づいて，段階的貪欲法を用いて記述子を追加した実験 (実験 E) を現在進めている．参考として実験 C と実験 E について Renal と CYP3A4 の各判別機械における f 値の比較をそれぞれ図 8 と図 9 に表す．f 値は Renal において実験 C では 0.80 であったが，実験 E では 0.91 に上昇しており，CYP3A4 では 0.78 から 0.92 に上昇している．ただしこの実験 E では，研究上の進展にともないデータ数が増加しており前のデータと単純比較することはできない．

7. 結　　論

本研究では，薬物の物理化学的記述子から腎排泄，肝臓の主要な 3 つの CYP 代謝，OATP による胆汁排泄の 5 つのクリアランス経路のうち最も支配的な経路を予測するシステムを構築した．薬物の記述子は数が多く，その全てを予測に使用することは学習効率の低下や過学習を招くので適当な記述子を特徴選択により抽出することにした．このとき段階的貪欲アルゴリズムとよぶ近似解法を求める特徴選択アルゴリズムを用いると，システム全体の予測精度が向上し，85% 程度の予測精度を得られることがわかった．またクリアランス経路ごとの ROC 曲線をみても，段階的貪欲アルゴリズムを使用した場合には他の手法と比べて，True Positive Rate の立ち上がりが大きくなることを示した．

文　　献

[1] 杉山 雄一, 山下 伸二, 加藤 基浩, ファーマコキネティクス－演習による理解－, 南山堂, 東京, 2003.
[2] 杉山 雄一, 楠原 洋之, 分子薬物動態学, 南山堂, 東京, 2008.
[3] Nello Cristianini, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2000
[4] 加藤 隆一, 鎌滝 哲也, 薬物代謝学－医療薬学・毒性学の基礎として－, pp.9-63, 東京化学同人, 東京, 1995.
[5] Stefan Lessmann, Solving Imbalanced Classification Problem with Support Vector Machines, International Conference on Artificial Intelligence 2004, pp.214-220, 2004.
[6] Cen Li, Classifying Imbalanced Data Using A Bagging Ensemble Variation(BEV), ACM-SE, no.45, pp203-208, 2007.
[7] PreADMET version2.0, http://www.bmdrc.org/
[8] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks, vol.13, no.2, pp.415-425, 2002.
[9] SVM^{light}, http://svmlight.joachims.org/
[10] 年本 広太, 機械学習による薬物のクリアランス経路の予測, 東京工業大学 学士研究論文, 2008.
[11] 胡 欣, Mauricio Kuguler, Anto Satriyo Nugroho, 黒柳 奨, 岩田 彰, 多クラスサポートベクターマシンによる各 SVM モジュールの独立特徴選択, 電子情報通信学会 NC 研究会技術研究報告, NC2005-86, pp.31-36, 2005.