Vol.2010-BIO-20 No.8
2010/3/5

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—

年本広太[†1] 草間真紀子[†2] 池田和史[†1]
堀田駿[†1] 前田和哉[†2]
杉山雄一[†2] 秋山泰[†1]

薬物のクリアランス経路予測として，我々は既に，SVM, 矩形領域法，ブースティングを用いた3種の予測システムを開発した．本研究では，専門家からの評価を基に薬物データを見直し，SVM 予測システムにおいて複数の経路を予測できるよう改良した．残りの2手法と比較したところ，本予測システムが最も高い予測精度であった．また，特徴選択によって入力に使用する最適な記述子を探索した結果，Accuracy=0.87 と高い予測精度を得た．さらに，特徴選択の有意性と，外部データによる評価を行った．

In silico Prediction System of Major Drug Clearance Pathways -Expansion for Multiple Pathway Prediction and External Validation-

Kouta TOSHIMOTO,[†1] Makiko KUSAMA,[†2]
Kazushi IKEDA,[†1] Shun HOTTA,[†1]
Kazuya MAEDA,[†2] Yuichi SUGIYAMA[†2]
and Yutaka AKIYAMA[†1]

We have already developed three kinds of prediction system of drug clearance pathway using SVM, rectangular method and boosting. In this study, we have revised the drug data based on the evaluation from the specialist. We also improved SVM prediction system to be applicable to multiple clearance pathways prediction. The improved method of SVM prediction system showed higher accuracy compared with other two techniques. In addition, when we searched for the best combination of descriptors for the input using the feature selection, we obtained very high prediction with Accuracy=0.87. Finally, we evaluated significance of the feature selection and validated this system using external dataset.

1. はじめに

新薬の開発から承認までにはおよそ 20 年以上の期間が必要とされ，さらに，年々医薬品の研究開発投資額は指数関数的に増加する一方で，認可された新薬数はほぼ頭打ちである[⋆1]．一方で，分子生物学の進歩や倫理規定の整備に伴い確立された *in vitro* 評価系により，臨床試験における候補化合物の開発中止理由として，1991 年には薬物体内動態が 40% を占めていたのが，2000 年には 10% に減少している[1)]．また，コンピュータを用いた *in silico* による structure-based drug design などが注目されており，薬物体内動態の様々な予測法が提案されている[2)]．

薬物の異物解毒は，多様な代謝酵素や，トランスポーターと呼ばれる輸送タンパクの協調的な作用により，別の化合物へと代謝・消失されていく．この代謝・消失に影響を及ぼす酵素等を**クリアランス経路**という．医薬品評価における薬物動態特性のなかでも，臨床使用において特に重要であるのはクリアランス経路である．なぜなら，これを知ることにより薬物血中濃度の変動が併用薬や遺伝的背景から予測可能となるからである．しかし，クリアランス経路を医薬品解釈の初期段階で正確に特定することは難しい．

そこで我々は，先行研究として化合物の構造式から得られる物理化学的特性 (記述子) に基づき，代謝のみならず，腎排泄や肝取り込みという全体的なクリアランス経路の振り分けの予測を行うシステムを開発した[3)4)]．先行研究では SVM，矩形領域法，ブースティングを用いた 3 種類の予測システムを構築した．この中で SVM を用いた予測システムが精度が最も高く，また特徴選択を用いて，入力に使用する記述子の組み合わせを最適化することで更なる高精度化を達成した．しかし，SVM を用いた予測においては，予測結果としてひとつのクリアランス経路しか出力できなかった．現実の化合物全体をみたとき，複数のクリアランス経路を有する化合物は相当数存在する．

そこで本研究では，専門家からの評価を基に見直した新たな薬物データを用いて，SVM を用いた予測システムに複数解を予測できるよう改良し，残り 2 手法との比較を行った．ま

†1 東京工業大学 大学院情報理工学研究科
Graduate School of Information Science and Engineer, Tokyo Institute of Technology
†2 東京大学 大学院薬学系研究科
Graduate School of Pharmaceutical Sciences, The University of Tokyo
⋆1 “PhRMA annual survey, 2000”, PhRMA, http://www.phrma.org/

た，入力に使用する記述子の組み合わせを特徴選択によって決定し，更なる高精度化を図った．さらに，特徴選択の統計的有意性と，更なる新規薬物データによる評価実験を行った．

2. 解析対象

2.1 クリアランス経路

本研究では，主要なクリアランス経路である 3 種類の **cytochrome P450(CYP3A4, CYP2C9, CYP2D6)** を介した肝臓での代謝，**腎排泄 (Renal)**，**OATP(Organic Anion Transporting Polypeptide)** による肝取り込みの計 5 種類を対象に予測を行った．このように生理的に異なる枠組みの反応を統合的に予測する試みは，本研究の独創的な特徴のひとつである．

Cytochrome P450(以下 CYP) は微生物から植物，動物まで生物界に広く分布する一群のヘムタンパク質である．CYP には，触媒する反応の基質特異性が異なるきわめて多数の分子種 (ファミリー) の存在が知られている[5]．本実験では，主に肝臓での代謝を行う CYP3A4, CYP2C9, CYP2D6 の 3 種類を予測する．CYP3A4 は医薬品代謝の約 50% に関与しており，ヒトの肝全 CYP 量の約 30% を占める代表的な CYP である．CYP2C9 は遺伝的多型が存在し，CYP2C9 遺伝子の調整およびコーディング領域には 50 もの一塩基多型 (SNP) が存在している．CYP2D6 は全肝 CYP の約 2% しか存在しないが，多くの医薬品の代謝に関与することが知られている．

腎排泄は糸球体で毛細血管を通して加圧濾過され，尿中に物理的にこし取られるというプロセスを経て行われる[6]．本研究では未変化体，つまり構造変化を伴うことなくそのまま尿中に排泄されるものがクリアランス量全体の 50% 以上を占める場合を考える．

トランスポーターは肝，腎，消化管など多くの組織の上皮細胞，または血管内皮細胞に存在し，物質の生体膜透過 (取り込み・排泄) に寄与しているタンパク質である[6]．種々の研究により，多くの薬物がトランスポーターの基質になることが明らかになっている．OATP トランスポーターは主に肝臓の血管側に発現し，薬物の肝取り込みに寄与することが知られている．本研究では 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が属している．

2.2 解析に用いるデータ

専門家からの評価を基に化合物のクリアランス経路の見直しを行った結果，本研究では 141 個の化合物データを得ることができた．このデータは，成書[7)8)9)] と一般的な OATP 基質薬とを合わせて 294 薬物をリストアップし，主なクリアランス経路の情報を抽出し，その中で主要なクリアランス経路が CYP3A4, CYP2C9, CYP2D6, Renal, OATP であるものを選出したものである．このデータセットは上記 5 つのクリアランス経路を 1 つだけ有する化合物のみで構成されている．実際の化合物には複数のクリアランス経路を有するものも存在するが，そのようなデータは含まれていない．また，化合物の電荷状態によって反応するタンパク質をある程度特定することが可能である．表 1 に各クリアランス経路において全データの中から正例となるデータの数を記載する．

表 1 各クリアランス経路に属する化合物数
Table 1 The number of data for each clearance pathway

経路名	3A4	2C9	2D6	Renal	OATP	Total
Anion(-)	0	11	0	18	18	47
Cation(+) or Neutral	52	1	18	23	0	94
合計	52	12	18	41	18	141

2.3 入力記述子

本研究で構築した予測システムに与える入力は記述子である．記述子とは，化合物の物理化学的物性値の総称である．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，大量の記述子を瞬時に求めるソフトウェアが存在する．本研究で使用する記述子もまた計算値を用いることにする[⋆1]．計算できた記述子には，一部の化合物で算出できなかったものも存在したため，それらは除外することにした．最後に，計算値が全ての化合物で等しい値であった記述子を除外した結果，合計 885 種類の記述子を得ることができた．本研究ではこれら 885 記述子を，以下の 2 分類に分けて使用した．

2.3.1 基本記述子 (Default descriptors)

基本記述子は，クリアランス経路判別において有効であると専門家の経験知から導出した記述子群のことを指す．基本記述子には，血漿中タンパク非結合率 (fup), 分子量 (MW), n-オクタノール/水分配係数 (logD), 電荷 (Charge) の 4 種が属する．本研究で構築した全ての予測システムの入力には，基本記述子が必ず含まれている．

2.3.2 その他の記述子 (Other descriptors)

その他の記述子は，先述した全 885 種類の記述子から基本記述子 4 種を除いた 881 種類

⋆1 ADME Boxes v4.0(Pharma Algorithms, Canada), SciFinder Scholar 2007(Chemical Abstracts Service, USA), ADMET Predictor(Simulations Plus, USA), PreADMET Ver2.0(BMDRC, Korea) の 4 つのソフトウェアの計算値を使用

の記述子群のことを指す．その他の記述子は第 6 章以降に使用する．

3. SVM を用いた予測システム

3.1 SVM 予測システムの構成

SVM を用いた予測システム全体の構成は図 1 のようになる．本システムは 5 つの SVM の出力から予測結果を算出する．本研究では SVM プログラムとして SVM$^{light\star 1}$を使用し，カーネル関数には Gaussian kernel $K(\mathbf{x},\mathbf{z}) = \exp(-\gamma||\mathbf{x}-\mathbf{z}||^2)$ を用いた．5 つの SVM は，あるクラスとそれ以外のクラス間で識別を行う One Versus the Rest(1-v-r) を用いて学習を行った[10]．各 SVM は，対応するクリアランス経路の判別を行い，判別境界面からのマージンを判別結果として出力する．予測結果が正の値のとき，入力として与えた化合物は対応するクリアランス経路を持つと予測されることになる．そして 5 つの SVM の出力のうち，**正の値を出力した SVM に対応するクリアランス経路すべてを予測結果として返すようにする**．我々の先行研究では，5 つの SVM の出力値のうち最大値を出した SVM に対応するクリアランス経路をひとつ予測結果としていた．

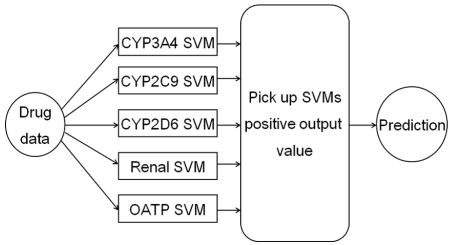

図 1 クリアランス経路予測システムの概要
Fig. 1 The outline of clearance pathway prediction system

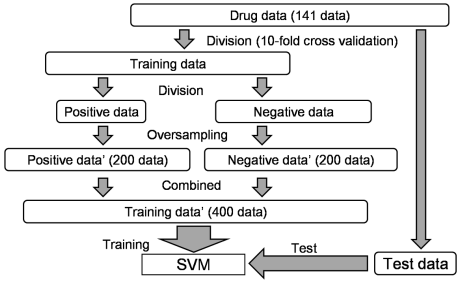

図 2 オーバーサンプリングによる学習
Fig. 2 Learning scheme by oversampling

3.2 学 習 方 法

3.2.1 パラメータチューニング

SVM の最適な学習にはパラメータのチューニングが必要で，今回の実験ではソフトマージンの大きさを示すパラメータ C と Gaussian Kernel の γ が相当する．そこで C, γ を様々な値で実際に学習を行い，予測精度を検証することで最適な値を探索する．本研究では C = {1, 5, 10}, γ= {0.001, 0.01, 0.1, 1, 10} での探索を行った．予測精度の検証には交差確認法のひとつである 10-fold cross validation を 10 回行うことにする．また，SVM の予測精度には f 値 (f-measure) を採用した．f 値は Recall, Precision の調和平均のことで，f 値が高いほど良い判別であるとみなすことができる．それは式 (1) の形になる．

$$\text{f-measure} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{1}$$

3.2.2 オーバーサンプリング

本研究で用いる 141 個の化合物データはクリアランス経路ごとにばらつきが大きい．また本研究では SVM が 1-v-r で学習を行うため，対応するクリアランス経路に属するものを正例，それ以外を負例として学習する．そのため今回の事例では CYP2C9 の場合，正例と負例の割合が (正例):(負例)=12:129 と極端に偏る場合が存在する．このようなデータをそのまま学習に用いると，少数データのクラスを過学習することが知られているため，与えられたデータをそのまま使用して SVM に学習を行わせるのは適切ではない．そこで本研究ではランダムオーバーサンプリング[11] を用いて正例と負例の偏りを補正する．ランダムオーバーサンプリングは，少数データに対して重複を許してランダムに選出することで疑似的にデータを増やす方法である．本研究では (正例):(負例)=200:200 となるようにランダムオーバーサンプリングを用い，それを 30 回行った平均値を SVM の出力とする．

なお，オーバーサンプリングには knn 法を基にして擬似データを増やす SMOTE 等の手法もあるが[12]，生成された擬似データが実際の化合物として存在しない可能性があるため本研究では利用しないことにした．

4. 既存の予測システム

4.1 矩形領域法 (Rectangular method) を用いた予測システム

矩形領域法は，判別平面に対する高い解釈性の実現のために筆者らが開発した予測法である[13]．矩形領域法は，各特徴量毎に上限と下限 (境界) をそれぞれ設定し，全ての境界条件を満たしている例題を正例とみなす，というものである (図 3)．**探索空間の中から f 値が最大でかつ矩形の体積が最小となる矩形を全探索によって探し出す**．判別領域の形が (超) 矩形になること，境界が各パラメータで独立であることなどから，この手法は判別が視覚的に理解しやすく，アルゴリズムも単純であるといえる．しかし一方で，実装の上では枝刈りを行っているが探索する矩形の候補は膨大なため，次元数の増加に伴い計算量が大きく増加す

⋆1 “SVM^{light}, Thorsten J, Cornell University, http://svmlight.joachims.org/

Vol.2010-BIO-20 No.8
2010/3/5

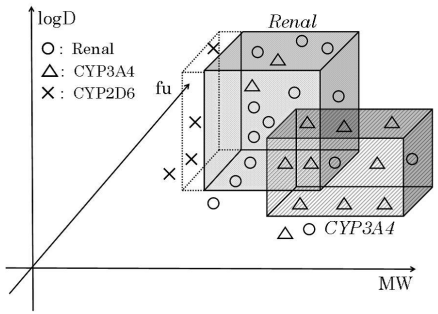

図 3 矩形領域法
Fig. 3 Rectangular method

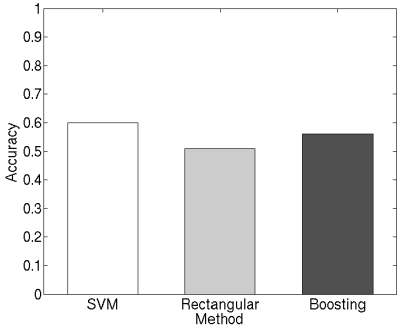

図 4 3 種の予測システムの予測精度
Fig. 4 Prediction accuracy values of the three kinds of prediction system

るという欠点も有する．

また矩形領域法では，各特徴量が連続値であることが望ましいため，化合物を予め電荷状態によって 2 グループ (負の電荷とそれ以外の電荷) に分類した後に矩形を探索していく．そのため本研究における矩形領域法は，各特徴量 fup,MW,logD の 3 次元空間上において直方体の判別領域を作成していることと同義であり，直方体の内部に存在するデータを正例とみなしていることになる．

4.2 ブースティングを用いた予測システム

ブースティング (Boosting) とは，1988 年に Kearns によって提唱された機械学習メタアルゴリズムの一種である．一般的には，得られた弱学習器にその正確さに応じた重み付けを行い，その結果から例題毎に付与される出現確率の見直しを行う．つまり，誤判別された例は重みを増やし，正しく判別できた例は重みを減らす．この一連の動作により，次の弱学習器は今までにうまく判別できなかった例題を重点的に判別するようになる．

ブースティングによる予測システムの研究は，池田ら[4] によって行われたものである．この研究では化合物を予め電荷状態によって 2 グループ (負の電荷とそれ以外の電荷) に分類した後に，ブースティングアルゴリズムを適応させることにする．各グループにおいて，不等式型，挟み込み型の 2 つの弱仮説クラスと，AdaBoost, MadaBoost の 2 つの損失関数を組み合わせて 4 種類の予測システムの構築し，最も精度よく予測できたものを使用する．

不等式型の弱学習器は，ある特徴量に対してある閾値以上 (以下) であれば+1, そうでなければ-1 を与えるものである．不等式型は単純な不等式であるため高速に学習が可能だが，全ての正例について好きな次元を判別面に選べるので，汎化能力はあまり高くないと考える．挟み込み型の弱学習器は，ある特徴量に対して上限値と下限値を設定し，その範囲内であれば+1, そうでなければ-1 を与えるものである．不等式型では，極端に狭い閉区間を作らないようにヒューリスティックな設定を行っているため，汎化能力は不等式型よりも改善されていると考えられる．

5. 評 価 実 験

SVM, 矩形領域法, ブースティングの 3 つの手法を用いた予測システムの予測精度の比較を行う．各手法とも 2.2 節で紹介した 141 個の化合物データセットを用い，入力には基本記述子のみを与える．SVM を用いた予測システムは 10-fold cross validation を，それ以外の 2 種については Leave-one-out cross validation による交差確認を行った．システム全体の予測精度の評価には，式 (2) で表した Accuracy を用いた．

$$\text{Accuracy} = \frac{\text{正しいクリアランス経路のみを予測したデータの総数}}{\text{データの総数}} \tag{2}$$

SVM，矩形領域法，ブースティングによる予測精度の結果を図 4 に示す．これより，SVM による予測が最も高い予測精度を示したことがわかる．SVM を用いた予測システムの予測結果の詳細を表 2 に示す．表の行成分は化合物の実際のクリアランス経路を表しており，列成分は予測結果として返したクリアランス経路を表している．例えば表 2 の (3A4,Renal)=4 は，実際のクリアランス経路は CYP3A4 であるが予測システムでは Renal であると予測さ

表 2 SVM の予測結果 (10-fold cross validation)
Table 2 Result of SVM prediction system (10-fold cross validation)

SVM		Predicted pathway								
		3A4	2C9	2D6	Renal	OATP	Multiple	None	Total	Recall
Observed	3A4	**42**	0	1	4	0	3	2	52	0.81
	2C9	1	**6**	0	0	0	5	0	12	0.50
	2D6	5	0	**6**	4	0	3	0	18	0.33
	Renal	0	2	0	**30**	0	4	2	41	0.73
	OATP	0	2	0	3	**1**	12	0	18	0.06
Total		51	10	7	41	1	27	4	141	
Precision		0.82	0.60	0.86	0.73	1	Accuracy: **0.60**			

れた化合物が 4 個あったことを表す．また，表の対角成分 (太字) が正しく予想できた化合物の数である．これより，SVM による予測では 2 つ以上のクリアランス経路を予測結果とした (Multiple) 化合物が 27 個と多いことがわかる．Multiple の予測結果を詳細に調べたところ，そのほとんどは 2 種類のクリアランス経路を有するという予測結果であり，うち一方は実際にその化合物が有するクリアランス経路であった．以上のことからも，今回構築した予測システムの比較では SVM が最も優れた予測システムであるとみなすことができる．

6. 特徴選択による高精度化

比較実験により，SVM を用いた予測システムが最も優れていることを検証した．ここでは，先行研究[3] と同じく入力に使用する記述子を変化させることで，更なる高精度の SVM 予測システムの構築することを目標とする．まず，薬学の専門家の知見を可能な限り汲むために，基本記述子は必ず入力として加えるものとする．その他の記述子は全部で 881 種類存在するため，考えうる入力の組み合わせは $2^{881} \approx 1.6 \times 10^{265}$ 通りと膨大であることから，そのすべてを探索することは現実的な時間内では不可能である．そこで本研究では貪欲法[14] に基づく特徴選択を行うことで，近似的に最適な記述子の組み合わせを探索した．貪欲法には局所的最適解に陥りやいという欠点があるが，比較的高速に計算ができるという利点を有する．本研究ではデータの学習時にオーバーサンプリング等で多くの繰り返しが行われ，判別境界面の作成に時間を必要とするため，貪欲法は最も妥当な近似解法のひとつであると考えた．貪欲法による特徴選択の手順を以下に示す．

(1) 初期化: $X = \{$ 基本記述子 $\}, Y = \{$ その他の記述子 $\}$ とおく．
(2) 選択: 記述子 $y \in Y$ をひとつ選出する．
(3) 評価: 入力として $X \cup \{y\}$ の記述子を用いて判別領域を学習し，予測精度を測定する．予測精度の測定の際には 10-fold cross validation を 10 回行い，その平均値を取る．
(4) 最大値の選出: 上記 2, 3 を Y に属する全ての記述子に対して行い，予測精度が最大となった時の y を $y\prime$ とする．
(5) 集合への追加: $X = X \cup \{y\prime\}, Y = Y/\{y\prime\}$ とする．ここで/は除外を表す．
(6) 以下 2〜5 を繰り返し，f 値が十分大きくなれば特徴選択を止める．

この特徴選択の (2), (3) において東京工業大学が所有する並列計算機，TSUBAME supercomputing system[⋆1] を用いることで，集合 Y に含まれる記述子が膨大であっても現実的な時間内に特徴選択を行うことを可能とした．

また，この特徴選択は 5 つの SVM が独立して行うものとした．すなわち，選ばれた記述子の組み合わせは 5 つの SVM で一般的にはそれぞれ異なる．我々の先行研究では，全ての SVM が同一の記述子を使用するように特徴選択を行った．

6.1 特徴選択による予測精度の比較

特徴選択により，予測精度がどの程度向上したかを検証した．この実験では，下記の 3 通りの入力を用いた SVM を作成し，10-fold cross validation を 10 回行って予測精度を測定した．また，我々の先行研究で構築した，SVM を用いて特徴選択を行った単一解予測システムとの精度の比較も行った．先行研究では実験データ，特徴選択の手法，基本記述子の扱いに違いがある．詳細は文献[3] に記載されている．

a) 基本記述子のみを用いた予測システム (表 2)
b) 基本記述子とその他の記述子を全て用いた (885 記述子) 予測システム
c) 貪欲法による特徴選択を行った予測システム

入力に用いる記述子を変えた 3 通りの SVM 予測システムと，先行研究の予測精度を図 5 に示す．貪欲法による特徴選択を行った結果，予測精度は基本記述子のみを使用した場合と比べて 0.2 以上大きくなった．また，885 記述子全てを入力に用いた場合と比べても，特徴選択を行ったシステムは高い予測精度を出すことが出来た．さらに，先行研究の SVM 予測システムと比較した場合にも，特徴選択を行ったシステムの予測精度は高いことがわかった．これは，特徴選択の 5 つの SVM が独立して行うよう変更したことで，各 SVM に最適

⋆1 TSUBAME は Dual core AMD Opteron 880 プロセッサ (2.4GHz), 32GB メモリの演算ノードを InfiniBand および Gigabit Ethernet で相互接続した Sun Fire X4600 クラスタシステムを中心としたグリッド計算機環境である．

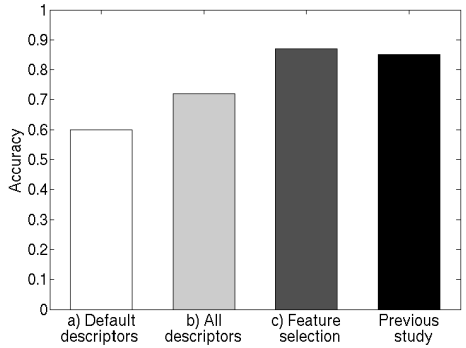

図 5 入力記述子の違いによる予測精度の比較 (10-fold cross validation)
Fig. 5 Prediction accuracy values obtained by different input descriptors (10-fold cross validation)

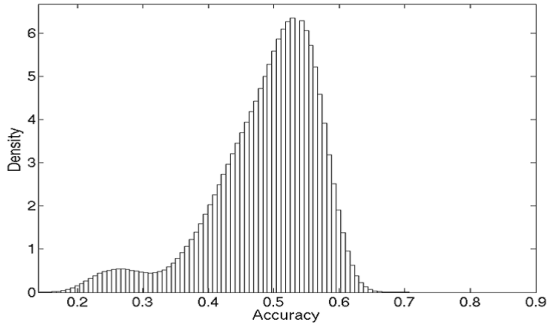

図 6 ランダムな特徴選択を行った予測システムの精度分布
Fig. 6 Distribution of prediction accuracy values obtained random feature selection

表 3 SVM, 特徴選択を行った際の予測結果 (10-fold cross validation)
Table 3 Performance of prediction by SVM and feature selection (10-fold cross validation)

SVM feature selection		Predicted pathway								
		3A4	2C9	2D6	Renal	OATP	Multiple	None	Total	Recall
Observed	3A4	**47**	0	0	0	0	0	5	52	0.90
	2C9	1	**8**	0	0	1	1	1	12	0.67
	2D6	2	0	**13**	0	0	0	3	18	0.72
	Renal	0	0	0	**37**	0	0	4	41	0.90
	OATP	0	0	0	0	**17**	1	0	18	0.94
Total		50	8	13	37	18	2	13	141	
Precision		0.92	1	1	1	0.94	Accuracy: **0.87**			

な記述子が選ばれた影響であると考えられる．

特徴選択を行ったシステムの詳細な予測結果を表 3 に示す．システムの予測精度は Accuracy= 0.87 となり，高い予測精度を示した．また特徴選択の結果，予測システムにおける各 SVM の入力の記述子数は CYP2C9, OATP で 6 個，CYP2D6 で 7 個，CYP3A4, Renal で 8 個と，入力に使用した記述子は比較的少数であった．予測に使用した記述子の数を我々の先行研究と比較すると，先行研究では 12 個の記述子を使用しており，本システムの方が入力に使用する記述子の数が少なくなっていることがわかった．さらに表 3 をみると，表 2 において多くを占めていた Multiple と予測された化合物が少なくなった．これは特徴選択により 5 つの SVM について False positive が少なくなったことにより，より正しい予測結果を返すことが出来たためであると考える．

また，特徴選択によって選ばれた記述子の意味を調べたところ，CYP3A4, CYP2D6, Renal の 3 つのクリアランス経路について，特徴選択の 1 番目に選ばれた記述子はファンデルワールス表面積に関連する記述子であった．ファンデルワールス表面積は酵素タンパク質と化合物の結合強度を考える際に重要な力のひとつで，protein-ligand ドッキングソフトウェア等でも考慮されるパラメータである[15]．そのようなパラメータが貪欲法による特徴選択で選出されるということは，ファンデルワールス力が酵素タンパクと化合物との相互作用を考えるうえで重要であることを示唆していると解釈できる．

6.2 特徴選択の統計的有意性

貪欲法による特徴選択を行うことで予測精度の向上が実現できたが，そもそもその他の記述子の総数が 881 個と多く，ランダムに記述子を選出しても単なる偶然で予測精度が向上する可能性があるかもしれない．そこで貪欲法による特徴選択と，ランダムに特徴選択を行った予測システムとの予測精度の比較を行うことで，貪欲法による特徴選択の統計的有意性を評価した．まず，3000 通りのランダムに特徴選択をした SVM をクリアランス経路ごとに用意した．各クリアランス経路に該当する SVM の入力の記述子数は貪欲法によって特徴選択を行った予測システムのそれと合わせる[⋆1]．それらの SVM を組み合わせて作ることのできる予測システムは $3000^5 = 2.43 \times 10^{17}$ 通り考えられるが，その中から 1.0×10^7 個の予

⋆1 CYP2C9, OATP で 6 個，CYP2D6 で 7 個，CYP3A4, Renal で 8 個

測システムをランダムに作成し，10-fold cross validation を 10 回行い予測精度 (Accuracy) を測定した．そして測定された予測精度の密度分布を求め，比較を行った．

ランダムに特徴選択を行った予測システムの予測精度の確率密度分布をヒストグラムで表したものを図 6 に示す．この確率密度分布より，ランダムに特徴選択を行ったときの予測精度は平均値 $\mu = 0.49$, 標準偏差 $\sigma = 0.082$ という数値を得た．特徴選択を行ったシステムの予測精度 Accuracy = 0.87 はランダムに特徴選択を行った場合の確率密度分布を大きく外れた値であり，偶然ではほとんど起こり得ないことがわかった．

6.3 外部データ (近年の承認薬) による評価実験

本研究に用いた 141 化合物データとは別の 35 個の化合物データをテストデータとして，貪欲法による特徴選択を行った予測システムの評価実験を行った．このデータセットは，2004 年以降に日米欧で承認された医薬品のなかから，主なクリアランス経路が CYP3A4, CYP2C9, CYP2D6, Renal, OATP のいずれかに該当する薬物を抽出したものである．各薬物のクリアランス経路は表 4 に示す．今回評価に用いた外部データは，2004 年以降に承認された薬物であるため，本システムにおいて従来探索していた化合物との分布上の偏りが存在する．この実験では，貪欲法による特徴選択を行った SVM に，これまで用いた 141 化合物データを学習させた．その後 35 個の新規化合物データをテストデータとしてクリアランス経路の予測を行い，予測精度を測定した．

表 4 新規 35 個の化合物データの内訳
Table 4 Newly obtained 35 drug data for each clearance pathway

経路名	3A4	2C9	2D6	Renal	OATP	Total
Anion	1	1	0	3	0	5
Cation & Neutral	18	0	0	12	0	30

評価実験の予測結果を表 5 に示す．予測精度は 10-fold cross validation のそれより低下していることがわかる．前述のように，今回用いた外部データは 2004 年以降に承認された薬物であるため，本来は本システムをそのまま適用しても高精度の結果を得られるとは限らないと考えられた．しかし，予測システムが単一のクリアランス経路であると予測した化合物にのみ注目すると．比較的正しい予測を行っていることがわかる．そこで，**単一解と予測された化合物の中から実際に正しいクリアランス経路を予測できた割合 (以下，Overall-precision と呼ぶ)** をみると，Overall-precision = 0.72 と比較的高い数値であった．この予測精度は薬物動態学の専門家から見ても良好な予測ができているとの評価を受けた．このように，近年の承認薬に対してもある程度の予測精度を出すことが検証出来た．今後更なる化合物データの追加に伴い，本当の創薬の空間が埋まってくるとすると，そのときの予測精度は特徴選択を行った際の値に近くなると期待できる．

表 5 SVM, 特徴選択を行った際の予測結果 (35 個外部データに対する予測)
Table 5 Performance of the proposed prediction system when applied to 35 external data

SVM feature selection		Predicted pathway								
		3A4	2C9	2D6	Renal	OATP	Multiple	None	Total	Recall
Observed	3A4	**9**	0	0	3	1	2	4	19	0.47
	2C9	0	**1**	0	0	0	0	0	1	1
	2D6	0	0	**0**	0	0	0	0	0	-
	Renal	1	0	2	**8**	0	1	3	15	0.53
	OATP	0	0	0	0	**0**	0	0	0	-
Total		10	1	2	11	1	3	7	35	
Precision		0.90	1	0	0.73	0				
Overall-precision		**0.72**					Accuracy: **0.51**			

7. 結　　論

本研究では SVM を用いたクリアランス経路予測システムについて，複数の経路予測が行えるよう拡張を行った．SVM は一般的に過学習を起こしにくいため，ブースティングや矩形領域法といった他の手法と比べると，高い予測精度を示した．

貪欲法に基づく特徴選択を行った結果，Accuracy = 0.87 となり，基本記述子のみを使用したシステムと比較して 0.2 以上の精度向上がみられた．また，885 個全ての記述子を入力として用いた予測システムとの比較でも，特徴選択を行った予測システムは，0.1 以上高い予測精度を示した．さらに我々の先行研究で構築した SVM 予測システム (特徴選択後) と比較したとき，本予測システムはより少ない記述子数で高い予測精度を示した．また，ランダムに特徴選択を行った予測システムとの精度の比較により，貪欲法による特徴選択を行った SVM 予測システムの精度は偶然ではほとんど起こりえない確率であることを実験的に示した．最後に近年の承認薬を集めた新規 35 データでの予測精度を評価したところ，141 データの時と比べると低下していたが，Overall-precision=0.72 と，比較的良好な予測精度であることがわかった．

7.1 今後の課題

本研究で使用した全ての化合物は，主たるクリアランス経路を単一に特定できたものであったが，実際の化合物をみると複数のクリアランス経路を持つ化合物も数多く存在する．

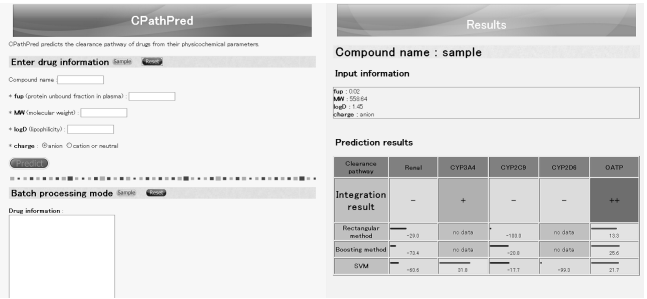

図 7 クリアランス経路予測 Web サービス
Fig. 7 Clearance pathway prediction Web survice

近年では副作用等のリスク分散のため単一の主要クリアランス経路を占める化合物は回避され，複数のクリアランス経路を有する医薬品の開発が好まれている．本研究ではそれらの化合物に対して予測結果の評価を行っていないが，今後はある経路で 5 割の寄与を占め，別の経路でも 3 割の寄与がある，といった化合物もデータとして集まってくると考える．そこで，集まってきた新規化合物データを学習に生かして，複数のクリアランス経路を有する化合物が入力として与えられても高い予測精度であるかを検証するため，化合物データの収集を行いより多くのデータセットを用いた評価実験を行う準備を進めている．

また，本予測システムの最終目標は医薬品開発の現場にて実際に利用することで，より効率的な医薬品開発を行う支援を行うことである．既に我々は基本記述子のみを使用した予測について，矩形領域法，ブースティング，SVM の 3 手法による予測を返す Web サービスを試験的に限定公開している (図 7)．このサービスは，化合物名 (省略可) と基本記述子である fup, MW, logD, Charge の 4 つのパラメータを入力することで，矩形領域法，ブースティング，SVM の 3 手法をそれぞれ用いた場合の予測結果を一覧として得ることができる．

謝辞 2004 年以降に承認された薬物の調査を行い，新規 35 化合物データセットを作成した（株）エーザイの若山直美氏に感謝する．また，本研究の一部は，(財）大川情報通信基金 2008 年度研究助成の支援を得て実施された．

参 考 文 献

1) Kola I, Landis J: “Can the pharmaceutical industry reduce attrition rates?”, *Nature Reviews Drug Discovery,* 3:711-5, (2004).
2) Van de W.H, Gifford E: “ADMET in silico modeling: towards prediction paradise”, *Nature Reviews Drug Discovery,* 2:192-204, (2003).
3) Toshimoto K, Kusama M, Maeda K, Sugiyama Y, Akiyama Y;“Prediction of Drug Clearance Pathway with Machine Learning ”, IPSJ-SIG Technical Report, 2008-BIO-13(10):43-48, (2008).
4) Ikeda K, Toshimoto K, Kusama M, Maeda K, Sugiyama Y, Akiyama Y: “Prediction of Drug Clearance Pathway by Boosting Algorithm”, IPSJ-SIG Technical Report, 2009-BIO-17(10):1-8, (2009).
5) 大村 恒雄, 石村 巽, 藤井 義明 編: “P450 の分子生物学”, 講談社サイエンティフィク, 東京, (2003).
6) 杉山 雄一, 楠原 洋之 編: “分子薬物動態学”, 南山堂, 東京, (2008).
7) Benet L, Oie S, Swharts J. B: “Design and Optimization of Dosage Regimens: Pharmacokinetic Data. In Hardman JG (ed.), Goodman & Gilman’s the pharmacological basis of therapeutics, 9th ed”., McGraw-Hill, New York, (1996).
8) Tummel K. K, Shen D. D: “Design and Optimization of Dosage Regimens: Pharmacokinetic Data. In Hardman JG (ed.), Goodman & Gilman’s the pharmacological basis of therapeutics, 10th ed”., McGraw-Hill, New York, (2001).
9) Tummel K. K, Shen D. D, Isoherranen N, Smith H. E: “Design and Optimization of Dosage Regimens; Pharmacokinetic Data. In L. Brunton (ed.), Goodman & Gilman’s The Pharmacological Basis of Therapeutics, 11th ed”., McGraw-Hill, New York, (2006).
10) Hsu, C. H, Lin, C. J: “A Comparison of Methods for Multiclass Support Vector Machines”, *IEEE Transactions on Neural Networks*, 13:415-425, (2002).
11) Liu Y, An A, Huang X: “Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles”, *In the Proceedings of the 10th Pacific-Asia Conference on knowledge discovery and data mining (PAKDD ’06)*, Singapore, pp.107-118, (2006).
12) Chawla N. V, Bowyer K. W, Hall L. O, Kegelmeyer W. P: “SMOTE: Synthetic Minority Over-sampling Technique”, *Journal of Artificial Intelligence Research*, 16:321-357, (2002).
13) 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).
14) Caruana R, Freitag D: “Greedy attribute selection”, *Proceedings 11th International Conference on Machine Learning*, pp.28-36, (1994).
15) Ewing T. J. A, Makino S, Skillman A. G, Kuntz I. D: “DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases”, *Journal of Computer-Aided Molecular Design*, 15:411-428, (2001).