1ZF-7

特徴選択とサポートベクターマシンを用いた薬物トランスポーター予測システムの構築

池田 和史† 前田 和哉‡ 尾瀬 淳‡ 杉山 雄一‡ 秋山 泰†
†東京工業大学 大学院情報理工学研究科 ‡東京大学 大学院薬学系研究科

1 はじめに

生体の細胞膜上にはトランスポーターと呼ばれるタンパク質が存在しており，必要な物質を体内・組織細胞内に取り込む，生体異物の体外排出を担う，あるいは体内への侵入を防ぐ,など様々な種類の機能を持ったトランスポーターが存在している．薬物トランスポーターは，直接的には薬物の体内動態（吸収，分布，排泄などの過程やその結果としての血中濃度プロファイル）を支配する要因として働き，標的組織における薬物濃度は，薬効・毒性と密接に関係する．すなわち，ある薬物の体内動態を支配するトランスポーターについて理解し，あるいは制御することができれば，その薬物の有効性・安全性の向上，ひいては医薬品としての価値の向上につながるものと期待される．

そこで本研究では，創薬の初期段階における候補化合物のハイスループットスクリーニングを目的として，化合物の構造式のみから得られる基本的な物理化学的特性を使用し，ヒト体内で働く主要な薬物トランスポーター７群に対するクラス分類を行う予測器をサポートベクターマシンを用いて作成した．

2 手法

OATPs, MRPs, OAT1, OAT3, OCTs/MATEs, MDR1, BCRP（略記，詳細は表 1）の７種のトランスポーター群を対象に，薬物が各トランスポーター群の基質になるかどうかを予測するシステムを教師あり機械学習の手法であるサポートベクターマシンを用いて構築する．

2.1 使用した記述子

記述子とは，化合物の構造から計算できるパラメータの総称である．本研究では，薬学の見地から予測に有用だと考えられている **4 種の基本記述子（電荷，分子量，分配係数，血漿中タンパク質非結合率）**と特徴選択により追加するための記述子 904 種 [1] を使用した．

In silico Prediction of Drug Transporters based on Support Vector Machine and Feature Selection
†Kazushi IKEDA ‡Kazuya MAEDA ‡Atsushi OSE ‡Yuichi SUGIYAMA †Yutaka AKIYAMA
†Graduate School of Information Science and Engineering, Tokyo Institute of Technology
‡Graduate School of Pharmaceutical Sciences, The University of Tokyo

2.2 貪欲法による特徴選択

最も予測精度を向上させる記述子ひとつを実際に予測モデルを構築することで発見し，基本記述子に加える．後述する実験では，この操作を複数回繰り返すことで，複数の記述子を追加する．記述子を加えても，予測器の結果がほぼ変化しなくなったところ（変化したサンプル数が全サンプルの 5% 以下）で追加を止める．

2.3 クラスタリングを利用した特徴選択

各記述子同士の相関係数を距離とした空間において，クラスタリングの手法を用いて複数のクラスタを生成し，それぞれのクラスタ重心に近い代表点をひとつ選ぶ．これにより特徴量を一定量にまで絞った上で，その中からさらに数個を選ぶ組み合わせを全通り考え，最も予測精度を向上させる記述子の組み合わせを探索する．クラスタリングの手法には代表的な群平均法，Ward 法の２つを試し，より結果の良かった Ward 法を用いた．

3 実験

3.1 実験に使用したデータ

対象の薬物トランスポーターの基質を化合物データベースと文献から取得した．本実験では確認できた基質 293 個を全て使用する．各トランスポーター群の略記と個数の内訳を表 1 に示す．

表 1: 薬物トランスポーター群毎のデータの内訳

トランスポーター名	略記	データ個数
OATP(1B1/1B3)	OATPs	59
MRP(2/3/4)	MRPs	64
OAT1	OAT1	41
OAT3	OAT3	52
OCT(1/2)/MATE(1/2-k)	OCTs/MATEs	35
MDR1	MDR1	128
BCRP	BCRP	75

3.2 特徴選択を用いた予測器の作成

７種のトランスポーターに対するクラス分類を実現するために one-versus-the-rest 法を用いる．自クラスのデータを正例，それ以外のデータを負例として各クラスで独立に予測器を生成する [2]．SVM のカーネルには Gaussian kernel を用いた．また予測精度を評価するために leave-one-out 法を使用し，precision と recall の調和平均である f 値を指標に用いた．

以上の条件下で，使用する記述子を以下のように変えながら3通りの実験を行った．

実験A） 基本記述子4種のみ
実験B） 基本記述子4種＋貪欲法による数種
実験C） 基本記述子4種＋クラスタリングによる3種（クラスタ数：30）

4 結果と考察

各実験の予測精度（f値）を表2に示す．

表2: 各実験におけるクラス毎の予測精度（f値）

クラス	実験A 基本記述子	実験B 貪欲法	実験C クラスタ
OATPs	0.67	0.69	0.69
MRPs	0.64	0.79	0.77
OAT1	0.46	0.72	0.73
OAT3	0.50	0.67	0.64
OCTs/MATEs	0.55	0.83	0.82
MDR1	0.83	0.88	0.88
BCRP	0.42	0.68	0.60
Total(average)	**0.58**	**0.75**	**0.73**

表2からまず，実験Aの予測器に比べ，実験B,Cの特徴選択を使用した予測器のほうが，全てのクラスにおいて予測精度が上昇していることが分かった．また，各クラスの予測精度の平均でみるとおよそ0.17と大幅に上昇した．また，実験Bと実験Cの比較では予測精度はさほど変わらなかった．実験Cのクラスタリングによる特徴選択は，貪欲法が局所的最適解に陥る危険性を考慮し，組み合わせによる特徴選択の手段として提案したものだが，このような結果になった原因として，貪欲法が局所的最適解に陥っていない，あるいはクラスタ数が少ないために必要以上に探索空間が狭まっていることが考えられる．後者が原因であった場合，対策にはクラスタ数，特徴量数の変更が考えられるが現状の探索空間の広さから考えても全組み合わせを探索することは困難であると考えられる．

5 評価者ベースの負例抽出法の提案

本研究に用いた薬物データには留意すべき点がある．上記した実験では予測器を構築する際に，自クラス以外のデータを全て負例として学習に用いているが，現実には文献に報告が挙がらない正例か負例か判断することが難しい薬物が存在する．このようなクラスラベルに曖昧さがあるサンプルを学習データに負例として使用することは危険である．そこで，これらのデータが学習に使われることを回避するために，複数の専門家の評価を参考にして，学習データから良質な負例を集める方法を以下に提案する．

1. 複数の評価者一人ひとりに，基質報告のない薬物のなかで，本当に基質でないと考えられるものを挙げてもらう．
2. Fleiss' kappa[3] の値によって一致度に貢献しない評価者を外す．
3. 評価者が知識を持たない場合を考える，すなわち評価者がどの薬物も等確率で基質でないと判断すると仮定したとき，判断した人数についての確率分布を求める．（これは二項分布に従う）
4. 二項分布の80% 信頼区間から外れるほど多くの評価者が基質でないと判断したデータを抽出し，知識が集約したサンプルとして学習データに採用する．

5.1 実験と結果考察

前述した実験と同様の環境で，基本記述子のみを使用し，抽出した負例を用いて予測モデルを構築した．表3に提案モデルにおける予測精度を示す．この予測精度の測定に抽出していないサンプルは含めていない．

表3: 抽出した負例によるモデルの予測精度（f値）

クラス	提案モデル
OATPs	0.95
MRPs	0.94
OAT1	0.98
OAT3	0.94
OCTs/MATEs	0.92
MDR1	0.98
BCRP	0.98
Total(average)	**0.96**

表3に示すように，提案モデルの予測精度はどのクラスについても0.9を超え，抽出したデータのほぼ全てが正例が存在する領域から離れた判別可能な領域に存在していることが分かる．

6 まとめ

本研究では，主要な7種の薬物トランスポーター群に対するクラス分類を行う予測器をサポートベクターマシンを用いて作成した．特徴量には，構造式から算出できる基本的な物理化学的パラメータのみを用いて，重要な記述子4種と約900種の記述子から選択したものを組み合わせ，高精度な予測を実現した．また，ラベル付けが難しい曖昧なデータセットからの負例の抽出方法として，複数の専門家による評価を利用する方法を提案した．

参考文献

[1] PreADMET version 2.0, http://www.bmdrc.org/
[2] K. Ikeda, *et al., 25th Annual Meeting of the Japanese Society for the Study of Xenobiotics*, (2010)
[3] J. L. Fleiss, *Psychological Bulletin*, Vol.76, No.5 (1971), p.378-382.