1P13

データマイニングによるシトクロム P450 基質の代謝位置予測手法の開発

○光山倫央、荒川正幹、船津公人
（東大院工）

1. 緒言

新規医薬品開発の際には、候補物質の活性の有無だけでなく、その薬物動態も検査する必要がある。薬物動態検査の中でも、代謝物同定や代謝速度検査など、代謝に関わる性質の試験には特に多くの時間を割かなければいけないのが現状である。その理由として、人体内に約 100 種存在する主要な薬物代謝酵素シトクロムP450のアイソザイムの基質特異性がいずれも低く、様々な異物の代謝に関わるため、未だにアイソザイム特異性、基質の位置選択性などが完全には解明されていないことが挙げられる。そのため、医薬品開発プロセスにおける代謝試験の効率化を図るべく、P450 の構造や特性の解明を目指した研究が、実験や量子化学計算を中心に盛んに行われている。特に近年ではこの問題に対し、新たにデータマイニング手法の適用も試みられている[1]。データマイニングとは「大量のデータから機械学習や回帰分析等の手法を用いてルール等の新規かつ有効な知識を得る行為の総称」である[2]。この手法は実験や量子化学計算と比較しコストと分析時間の点で優位性があり、研究開発期間の大幅な短縮、コストの削減が期待される。

2. 目的

本研究では、P450 の特性の中でも基質の位置選択性、すなわち基質分子のどの部位が代謝反応を受けるか、ということに注目し、その知見を得ることを目指す。既往の研究に、各基質の構造データを用いて代謝位置予測モデルを構築したものがあるが[1]、理解可能なルールの導出には至っていない。基質の代謝位置に共通するルールを見出す事が出来れば、代謝物同定や代謝安定度の調整などが容易になり、創薬分野へ大きな貢献を果たす事になる。そこで本研究では P450 基質の位置選択性データにデータマイニング手法を適用し、良いルールの導出を行うことを目的とする。ここで言う良いルールとは、理解が容易で、精度の高いルールのことである。

3. 手法

データマイニングの手順は、データの用意、変数選択やオートスケーリングなどの前処理、学習、モデル評価の 4 つからなる。ここでは変数選択手法と学習手法について説明する。

- ●変数選択手法：本研究では、新たに Rough set theory を用いた変数選択手法を開発し、一般的な手法である情報利得などと比較する。
- ●学習手法：学習手法は、以下のものを用いる。まず、最も一般的なデータマイニング手法の一つである決定木を用いる。ここで、決定木などの学習手法を単独で用いるのではなく、複数の手法を用いて結果を統合するアンサンブル学習を用いることで、多くの場合予測精度が向上する事が知られている。そこで本研究では主要なアンサンブル学習手法であり、その有用性が示されている boosting、bagging、multi boosting(MB)の 3 手法をそれぞれ決定木に適用したものや、random forest（RF）、比較的理解可能なルールを導出しやすい alternating decision tree(ADT)などを用いて結果を比較する。また、上記の手法はみな命題学習器と呼ばれ、適用範囲が単一の表のみに限られる。そのため原子同士の関係を無理やり単一の表で表す必要があり、列の数が多くなる、0 の値が入る箇所が極端に多くなるといった問題が生じる。本研究ではルール導出手法として上記手法に加え、複数の表にまたがるデータからもルールを導く事が出来る帰納論理プログラミング(ILP)も用いる。

4. 結果と考察

4.1 解析データ

P450 のアイソザイムの中で薬物代謝との関係が特に深いものの一つとして 2C9 が挙げられる。Sheridan らは、2C9 の基質 92 分子について、その分子の構造と代謝位置のデータを収集した[2]。本研究ではそのデータを確認、修正後最終的にその中の 50 分子を用いた。各基質分子中のそれぞれの原子が一サンプルである。目的変数の値は、その原子が文献中で代謝されると記載のある部位であるなら 1、それ以外は 0 とした。また説明変数には、隣接する sp^3 炭素の数や水素結合ドナーの数などのフラグメント数、原子が末端にあるかどうかなど、分子の 2 次元構造のみから得ら

れる構造記述子を中心に、7 種類の記述子を用いた。サンプル数は 1031 個、説明変数の数は 1981 個となった。

4.2 予測モデル構築の結果

表1に、各学習手法ごとに様々な変数選択パターンを試した場合の最も良いモデルの結果を示す。この結果より、RF や MB などで特に高精度の予測モデルが構築可能であることが示された。さらに本研究では予測モデルを構築することに留まらず、得られたモデルからルールを得る事を目指している。そこでこれらのモデルを解釈することで、ルールを得ることを試みた。4.3 に、モデルからルールを導出した結果を示す。

表 1　予測モデルの精度

	RF	Bag	Boost	MB	ADT	決定木
検出率	0.51	0.55	0.51	0.57	0.40	0.44
精度	0.86	0.71	0.84	0.77	0.70	0.72

4.3 ルール導出の結果

各手法の中で実際にルールを得ることが出来たのは決定木、ADT および ILP の 3 手法であった。その中で最も評価値の高いルールは決定木により得られたものであり、その評価値は検出率 23%、精度 84%であった。このルールを理解しやすい形式に変換すると、「末端に位置し、最遠原子との path が 12 以上あるメチル基は代謝される。但し 1 つ先と 2 つ先に結合している原子の少なくとも 1 つは 2 級の sp^3 炭素ではないものに限る」 となる。図 1 にそのルールが表す分子構造を示す。このルールはω-酸化や *O*-脱アルキル化反応などを表していると考えられ、実験結果に対して再現性のあるルールであるといえる。この結果より、代謝の位置選択性に関して、分子の 2 次元構造に関する情報のみを用いて良いルールの構築が可能であることが示された。

5. まとめ

本研究では、薬物代謝酵素の位置選択性の問題に対してデータマイニング手法を適用した。その結果、代謝位置に関して実験結果に対して再現性のある良いルールの導出が達成された。学習データの追加、解析手法の改善や用いる変数の変更などで、より高精度のモデルの構築および良いルールの導出が可能となると考えられる。本手法を用いることで、2C9 のみでなくより多様かつ分子量の大きい基質群を代謝する酵素に対しても、ルールの導出が可能となることが期待される。

[参考文献]

[1]R.P.Sheridan, K.R.Korzekwa, R.A.Torrse, J.Med.Chem. 50(2 005), 3173-3184

[2]元田浩, 津本周作, 山口高平, 沼尾正行, データマイニングの基礎, オーム社, 2006

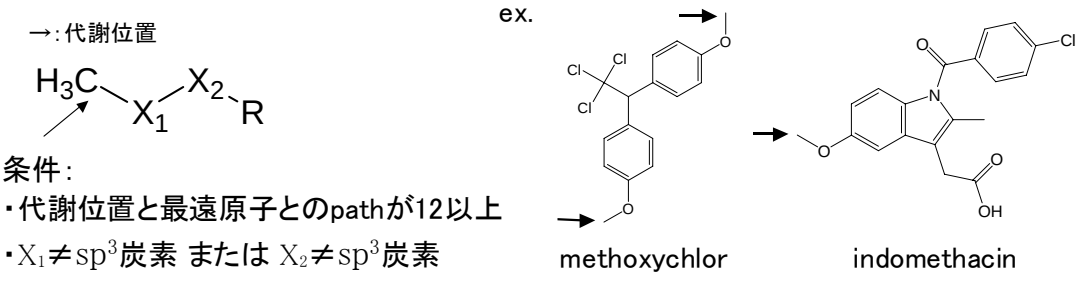

図 1　得られたルール