

氏名	フン 馮	チュン 春	ライ 来
学位(専攻分野)	博 士 (薬 学)		
学位記番号	薬 博 第 620 号		
学位授与の日付	平成 19 年 3 月 23 日		
学位授与の要件	学位規則第 4 条第 1 項該当		
研究科・専攻	薬学 研究科 医療薬科学 専攻		
学位論文題目	Informatics-based data acquisition and analysis of cytochrome P450-mediated drug metabolism (情報科学的アプローチを駆使したチトクローム P450 薬物代謝酵素に関するデータ収集と解析)		
論文調査委員	(主 査) 教授 橋田 充	教授 高倉 喜信	教授 乾 賢一

### 論 文 内 容 の 要 旨

チトクローム P450 (CYP) は、薬物の酸化的代謝に関わる薬物代謝酵素であり、多くの医薬品の体内動態に深く関与している。実際に臨床の現場では、CYP の遺伝的多型に体内動態変動や競合的代謝阻害による薬物間相互作用など、CYP 代謝に関わる問題も多数報告されている。これらの問題は新薬の開発研究にも大きな影響を及ぼし、最近ではリスク回避の観点から医薬品探索研究の早期から CYP 代謝のスクリーニングが行われるようになってきた。しかしながら、医薬品開発のより望ましい姿として、CYP による薬物代謝を *in silico* で予測する技術を確認し、ライブラリー設計の段階での仮想スクリーニングやリード最適化における論理的分子設計を実現することが必要とされている。一方、CYP は多様な分子種からなるスーパーファミリーを形成し、しかもそれぞれの分子種が多様な薬剤と相互作用することが知られている。したがって、CYP による薬物代謝の予測を精度よく行うためには、大規模な情報の収集と解析が不可欠となる。そこで本研究では、過去に蓄積された膨大な文書情報から化合物と代謝酵素 CYP との相互作用に関する知見を自動的に抽出できるテキストマイニング技術の開発を行い、得られたデータを元にして基質認識に関する大規模な構造活性相関および CYP 分子種間での基質特異性と構造との関係に解析を行った。

#### I. 化合物/酵素間相互作用解析のためのテキストマイニングシステムの開発とチトクローム P450 薬物代謝酵素に関する相互作用情報の収集

化合物と酵素との間の相互作用関係は大別すると、基質、阻害剤、誘導剤の3つのパターンが挙げられる。ライフサイエンス分野でよく用いられる共起性に基づくテキストマイニング法では、2つのキーワードの関係の種類までを知ることができないという欠点がある。そこで、自然言語処理技術を基本として、構文解析から相互作用を表す動詞に着目して所望の情報を抽出する、新しいテキストマイニングシステムを開発した。まず、PubMed 検索システムで利用されている MeSH Term データセットを解析し、約100,000個の化合物名と約35,000個のタンパク質名を抽出し、名前リストを作成した。辞書による抽出漏れや誤認識を改善するために、文書中で化合物名を伴うことが多い単語 (concentration など) や単位 (mg/ml など) に対する専門用語リストを作成するとともに、文脈から化合物名を識別するためのルールベースを作成した。次に、ヒットした化合物と CYP との相互作用に関する情報を抽出するために、品詞に基づく構文解析から動詞句と名詞句に分割し基本単位の文節を再構築するアルゴリズムを作成した。最後に、動詞あるいは動詞性名詞を核としたパターンマッチングにより、再構築した文から化合物と CYP との相互作用パターンを抽出した。開発したシステムのパフォーマンスを評価するために、PubMed から CYP2D6, CYP3A4 に関する論文要旨をそれぞれ100個ランダムに選抜し、抽出されたデータと実際の記載内容との比較を行った。化合物名の識別については、再現率が0.871、精度が0.941であり、基質・阻害剤・誘導剤の識別に対する再現率および精度も、CYP3A4 ではそれぞれ0.867、0.882で、CYP2D6 では0.865、0.92と非常に良好な成績であった。

ヒトにおける薬物代謝で重要と言われている代表的な6種類の CYP 分子種に絞り、PubMed データを解析した結果、相

相互作用に関するレコードがそれぞれ、CYP1A2で1670個(573化合物)、CYP2C9で924個(321化合物)、CYP2C19で813個(263化合物)、CYP2D6で1797個(458化合物)、CYP2E1で1677個(585化合物)、およびCYP3A4で2,700個(634化合物)得られた。これらのレコードを整理すると約2,000個化合物になり、そのうち約1,000個の化合物に対してはPubChemへのリンクを作成できた。CYP分子種間で基質および阻害剤のクロス集計を行い、カイ二乗検定法により相関解析を行った。その結果、基質および阻害剤のいずれの場合にも、CYP2C19とCYP2C9、CYP2C19とCYP2D6の順に高い相関が認められ、CYP間の配列相同性との関係が示された。

## II. 化合物とチトクローム P450 との相互作用に関するケモインフォマティクス解析

化合物とCYPとの相互作用に関するパターン認識は、医薬品候補化合物の分子設計において重要な課題である。第I章で得られた1000個の化合物の化学構造からADMET Predictor 1.3.2を用いて種々の分子記述子を算出し、各CYPに対する相互作用と化合物構造との間の関係を、多目的変数に対応した拡張再帰分割法に基づいて解析した。その結果、CYP2E1に対する基質および阻害剤はいずれも分極率の小さい低分子量の化合物が特徴的に多いのに対し、CYP2D6およびCYP3A4の基質には分極率の高い化合物が多かった。また、CYP2D6の基質にはとりわけ中性でカチオン型率の高い化合物が多い傾向があった。これは、カチオン性の低い化合物がCYP2D6の阻害剤となりにくいという結果とも対応していた。

CYP間での基質選択性の違いは、CYP自体の構造的特徴の違いによるものと考えられる。そこで、CYP1A1、CYP1A2、CYP2B6、CYP2C9、CYP2C19、CYP2D6、CYP2E1およびCYP3A4の8種類を対象として、データマイニングにより化合物のCYP代謝に関するパターン行列の特徴抽出を行い、得られた因子負荷行列と因子スコア行列に対してそれぞれアミノ酸配列構造情報と化合物構造情報との相関解析を行った。CYP代謝のパターン行列を主因子法によって因子分析を行ったところ、因子負荷量から第1主成分がCYP3A4を特異的に区別する成分、第2主成分はCYP3A4とCYP2D6を対極的に区別する成分であることが示され、対応する因子スコアはそれぞれ、化合物の分子サイズ、水素結合受容性とよく相関することが明らかとなった。さらに、配列解析ツールClustalWを用いて8種類のCYPのアミノ酸配列のマルチプルアラインメントを行い、各アラインメント位置においてアミノ酸の疎水的、立体的、あるいは静電的性質と先の因子負荷量との相関解析を行ったところ、いくつかの高い相関性を示すアミノ酸部位を同定でき、それらの中には過去の部位特異的変異実験でCYP活性の変化が示されたアミノ酸が含まれていた。

以上、本研究では、文献情報から大規模な相互作用データを抽出するテキストマイニングシステムを開発し、得られた情報をさらに解析することによって、化合物とCYPとの相互作用に関する構造活性相関およびCYPの基質特異性に影響する構造的因子を見出すことに成功した。これらの知見は薬物代謝の*in silico*予測に有用な情報を提供するとともに、本解析方法は様々な生体分子と薬剤との相互作用解析に容易に拡張できることから、本開発技術も創薬研究の推進に大いに貢献するものと期待される。

## 論文審査の結果の要旨

チトクローム P450 (CYP) は、薬物の酸化的代謝に関わる酵素として多くの医薬品の体内動態に深く関与し、臨床ではCYPの遺伝的多型による体内動態変動や競合的代謝阻害による薬物間相互作用など、CYP代謝に関わる問題も多数報告されている。新薬の開発研究においても、リスク回避の観点から医薬品探索研究の早期にCYP代謝のスクリーニングが行われるようになり、CYPによる薬物代謝の*in silico*予測を用いたライブラリー設計の段階での仮想スクリーニングやリード最適化における論理的分子設計が望まれるようになったが、一方、CYPは多様な分子種からなるスーパーファミリーを形成しかつ各分子種が多様な薬剤と相互作用することから、代謝予測を精度よく行うためには大規模な情報の収集と解析が不可欠となっている。著者は、過去に蓄積された膨大な文書情報から化合物とCYPとの相互作用に関する知見を自動的に抽出できるテキストマイニング技術を開発し、得られたデータより基質認識に関する構造活性相関およびCYP分子種間での基質特異性と構造との関係を解析した。

化合物と酵素の間の相互作用関係は、基質、阻害剤、誘導剤の3つのパターンに分けられる。自然言語処理技術を基本として、構文解析から相互作用を表す動詞に着目して所望の情報を抽出する新しいテキストマイニングシステムの開発を目指し、先ずPubMed検索システムで利用されているMeSH Termデータセットを解析して約100,000個の化合物名と約

35,000個のタンパク質名を抽出し名前リストを作成すると共に、辞書による抽出漏れや誤認識を改善するために、文書中で化合物名を伴うことが多い専門用語リストや文脈から化合物名を識別するためのルールベースを作成した。次に、ヒットした化合物と CYP との相互作用に関する情報を抽出するために、品詞に基づく構文解析から動詞句と名詞句に分割し基本単位の文節を再構築するアルゴリズムを作成し、最後に動詞あるいは動詞性名詞を核としたパターンマッチングにより、再構築した文から化合物と CYP との相互作用パターンを抽出した。PubMed から CYP2D6, CYP3A4 に関する論文をそれぞれ100個ランダムに選抜し、抽出されたデータと実際の記載内容との比較を行った結果、化合物名の識別や基質・阻害剤・誘導剤の識別に対する再現率および精度は非常に良好であった。次に、医薬品候補化合物の分子設計において重要な課題であるパターン認識に関し、1000個の化合物の化学構造から ADMET Predictor 1.3.2 を用いて種々の分子記述子を算出し、各 CYP に対する相互作用と化合物構造との間の関係を多目的変数に対応した拡張再帰分割法に基づいて解析した結果、各 CYP 分子種に特徴的な基質および阻害剤の特性が明らかとなった。一方、配列解析ツール ClustalW を用いて8種類の CYP のアミノ酸配列のマルチプルアラインメントを行い、各アラインメント位置においてアミノ酸の疎水的、立体的、あるいは静電的性質と先の因子負荷量との相関解析を行ったところ、いくつかの高い相関性を示すアミノ酸部位を同定でき、過去の部位特異的変異実験で CYP 活性の変化が示されたアミノ酸が含まれていた。

以上、本研究では、文献情報から大規模な相互作用データを抽出するテキストマイニングシステムを開発し、得られた情報をさらに解析することによって、化合物と CYP との相互作用に関する構造活性相関および CYP の基質特異性に影響する構造的因子を見出すことに成功した。これらの知見は薬物代謝の *in silico* 予測に有用な情報を提供するとともに、本解析方法がさまざまな生体分子と薬剤との相互作用解析に容易に拡張できることから創薬研究の推進に貢献するものと期待される。

よって、本論文は博士（薬学）の論文として価値あるものと認める。

更に、平成19年2月22日論文内容とそれと関連した口頭試問を行った結果合格と認めた。