様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月２５日現在

機関番号：１４３０１
研究種目：基盤研究（B）
研究期間：2009～2011
課題番号：２１３９０００８
研究課題名（和文）　自然言語処理に基づく薬物代謝ネットワークと化合物間相互作用の網羅的情報解析
研究課題名（英文）　Natural Language Processing-Based Comprehensive Data Analysis for Interaction Between Chemicals and Drug Metabolism Network
研究代表者
山下　富義（YAMASHITA FUMIYOSHI）
京都大学・大学院薬学研究科・准教授
研究者番号：30243041

研究成果の概要（和文）：
薬物代謝は医薬品の有効性および安全性に深く関わり，その阻害や誘導は多剤併用時に生じる薬物間相互作用の原因となっている。薬物間相互作用の予測を行うためには，単に薬物代謝酵素のみならず，その活性や発現を制御する生体ネットワーク全体を包括的に理解することが必要となる。本研究では，自然言語処理を利用したテキストマイング技術を開発し，これを利用して薬物代謝酵素やその発現を制御する生体分子と化合物との相互作用に関する情報を収集した。さらに，得られた情報を元に，包括的な構造活性相関解析を実施した。本研究で得られた情報は，安全な医薬品の開発を目指した創薬分子設計に有益な情報を提供するものと期待される。

研究成果の概要（英文）：
Drug metabolism is closely related to efficacy and safety of drugs, of which inhibition or induction is involved in drug-drug interaction following administration of multiple drugs. In order to predict drug-drug interactions, it is required to comprehensively understand not only interaction of chemicals with individual drug-metabolizing enzymes but that with a network system regulating expression and activity of the enzymes. In the present study, we developed a natural language processing-based text-mining technique for systematically collecting information on interaction between chemicals and a drug metabolism-related network. In addition, based on the information collected, we systematically performed structure-activity relationship analyses. These results will present a clue to drug design and discovery intended for development of safer drugs.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	6,800,000	2,040,000	8,840,000
２０１０年度	3,800,000	1,140,000	4,940,000
２０１１年度	3,600,000	1,080,000	4,680,000
年度			
年度			
総　計	14,200,000	4,260,000	18,460,000

研究分野：　医歯薬学
科研費の分科・細目：薬学，物理系薬学
キーワード：薬物代謝，薬物間相互作用，自然言語処理，テキストマイニング，構造活性相関

１．研究開始当初の背景

コンビナトリアルケミストリー，ハイスループットスクリーニングなど創薬基盤技術の革新により早期の医薬品探索研究は飛躍的なスピードアップを遂げたものの，医薬品の承認数は過去 10 数年に渡ってほとんど変化しないむしろ減少傾向にある。これは，医薬品として満たすべき条件（薬理活性，安全性，薬物動態，製剤学的物性など）が多面的で，かつそれらがしばしばトレードオフの関係にあるため，医薬品開発に向けた化合物展開の方向が定まらないことが根本的な問題となっている。

薬物の代謝は，医薬品の作用部位への到達性および滞留性を決定する薬物動態関連因子の一つであり，薬物の副作用とも深く関わっている。しばしば多剤の併用による薬物間相互作用が問題となることがあるが，その原因の多くがチトクロム P450 に代表される薬物代謝酵素によるものであることが知られている。そのため，臨床における薬物治療の実践あるいは新薬の開発のいずれにおいても，薬物代謝によって引き起こされる薬物－薬物間相互作用の分子機構を理解し，これを回避する方法を論理的に見出すことが不可欠となってくる。しかしながら，通常競合阻害のような分子レベルでの直接的なメカニズムに加え，酵素誘導やネガティブフィードバックといった酵素の発現レベルの変動により薬物代謝活性が影響される間接的なメカニズムも考えなければならない。すなわち，相互作用の予測においては薬物代謝酵素の発現を制御する生体分子ネットワークレベルでの包括的な理解が必要とされる。代謝酵素の発現制御に関する研究はこれまでに多くの研究がなされ，例えば AhR，CAR，PXR などの核内受容体の関与がよく知られているが，これらの核内受容体は，互いにクロストークして複雑な相互作用ネットワークを形成している上，代謝酵素と同様極めてブロードな基質選択性を示すため，代謝酵素関連ネットワークの入出力に相当する薬物－薬物間相互作用を俯瞰できるような知識情報ネットワークと呼ぶにふさわしいものは存在しなかった。

２．研究の目的

申請者は，これまでに遺伝的アルゴリズムやニューラルネットワークなどの情報科学的手法を取り入れた構造活性相関解析法を考案するとともに，構造的に多様な化合物に対して薬物動態特性の予測が可能なモデル開発を行う一方，自然言語処理アルゴリズムに独自開発した辞書ならびにルールベースを実装し，文献上に記載されるテキスト情報の中から化合物の名前およびチトクロム P450 との相互作用様式に関する情報を抽出するアルゴリズムを開発してきた。そこで，本研究では，テキストマイニング技術を利用して，薬物代謝酵素に対して直接的に相互作用する基質・阻害剤に加え，酵素の誘導剤に関する情報を収集するとともに，代謝酵素の発現調節にかかわる分子とそれらと相互作用する化合物の情報を網羅的に収集することを計画した。さらに，得られた化合物の化学構造を解析し，相互作用に関係する構造的要因を明らかにすることを試みた。

３．研究の方法

(1) タンパク質と化合物との相互作用に関するテキストマイニングシステムの構築：まず米国 National Center for Biotechnology Information(NCBI)が提供する MeSH Term を解析し，化合物名およびタンパク質名をリストした辞書データベースを構築した。タグ作成，形態素解析および構文解析には，自然言語処理オープンソフトウェア GATE の基本コンポーネントを利用することとしたが、文脈に基づくセマンティック解析のためのルールベースは JAPE 言語を使って作成し，GATE に実装した。基本的には，文を名詞句と動詞句に分け，名詞句に含まれるタンパク質名および化合物名を，動詞句から動詞を抽出し，それらの語順と動詞の時制および意味から相互作用の分類を行なった。また，動詞性名詞に関しても解析対象に含めることで，再現率の向上を図った。

(2) 多目的同時再帰分割解析法の開発と CYP 基質の構造活性相関解析：多特性データをもつ化合物群を分類するための決定木を作成することを目的として，新しい最適な分岐ルールを選択するための評価式を考案した。具体的には，データ分割による情報利得を各特性値に対して計算し，その総和を評価式とした。決定木は，十分成長させた後，交差予測実験による誤分類率を指標に枝刈りを行って最適化した。本方法を用い，主要な CYP 分子種 5 種に絞って多目的構造活性相関を解析した。

(3) 大規模情報可視化による構造活性分類の評価：多目的再帰分割解析により得られる決定木により化合物は階層的に分類されることに着目し，データオブジェクトを長方形枠で入れ子状に囲みながら階層構造を表現する方法を考えた。これにより，決定木の分類精度を視覚的に評価するとともに，多目的分類された化合物に対し構造活性相関について検討した。

４．研究成果

(1) CYP 分子種と相互作用する化合物のテキストマイニングと構造活性相関解析：チトクロム P450 は薬物の酸化的分解を担う酵素ファミリーであり，多くの薬物の消失に関っている。そこで，主要な分子種である CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP2E1, CYP3A4 に焦点を絞り，その情報収集と構造活性相関解析を行った。

まず，各 CYP 分子種の名称でフィルタした PubMed アブストラクト（CYP1A2, 4108; CYP2C9, 2670; CYP2C19, 2275; CYP2D6, 4423; CYP2E1, 4301; CYP3A4, 5,278）に対してテキストマイニング解析を適用し，それぞれ 2391, 1477, 1246, 2482, 2430, 4466 個のレコードを得た。これらから重複するレコードを除き整理した結果，最終的に化合物構造までが既知の基質の数はそれぞれ 216, 145, 136, 217, 156, 379 個，阻害剤の数は 167, 130, 111, 154, 121, 274 個，誘導剤の数は 101, 31, 16, 22, 81, 125 個となった。テキストマイニングの性能を確認するために，コーパスの中から 200 個のアブストラクトをランダムに選択し照合した結果，化合物名の抽出に関しては再現率が 86.3%，正確度が 86.7%であり，化合物-CYP 間相互作用の抽出に関しては再現率が 72.2%,正確度が 83.9%であった。

各 CYP 間で基質等の類似性を評価した結果，基質，阻害剤，誘導剤いずれにおいても CYP2C9 と CYP2C19 との間で類似性が高く，アミノ酸配列構造のホモロジーと対応した結果であった。しかしながら，アミノ酸配列構造の類似性が比較的高くても CYP2E1 はこれらとは異なった認識特性を示すことから，１次構造の類似性だけでは議論できないことが明らかとなった。

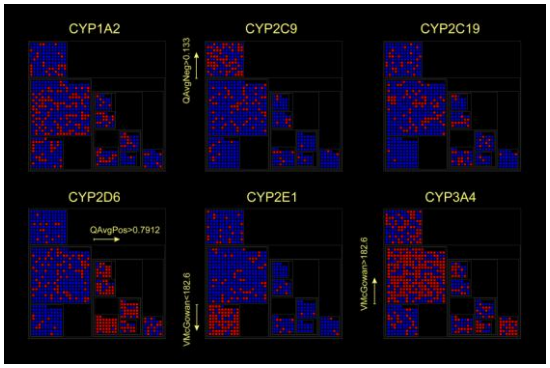

図１ テキストマイニングにより収集された化合物の各 CYP 分子種に対する代謝感受性の予測結果. 赤アイコンが基質であり， 2D6 基質の多くはカチオン性化合物群に，2E1 基質の多くは低分子量化合物群に分類される. 3A4 の基質は多いが，2E1 とは異なり低分子量の化合物は基質になりにくい.

CYP 基質の構造活性相関に関して，精査された 161 個の化合物データを多目的同時再帰分割法により解析した結果，pH7.4 において負電荷を有するものは CYP2C9 の基質となりやすく，逆に正電荷をもつものは CYP2D6 の基質となりやすいことが示された。さらに，分子量の小さいものは CYP2E1 の基質に，大きいものは CYP3A4 の基質になりやすいことも明らかとなった。得られた決定木モデルの妥当性を検証するために，テキストマイニングにより収集され化学構造既知の計 709 個の基質に関し，予測を行った結果，良好な予測結果が得られることが示された（図１）。

(2) CYP 発現・活性を調節するタンパク質のテキストマイニング：タンパク質間相互作用を抽出するテキストマイニング法は共起解析や機械学習法が利用されている。しかしながら，前者の場合は相互作用メカニズムの分類が困難であり，後者の場合はカーネル法を利用するため情報抽出が学習用データセットに強く依存するなどの問題点を抱えていた。そこで，自然言語処理を利用してタンパク質間相互作用に関する情報を抽出するテキストマイニングシステムを開発し， CYP の発現や活性の調節に関わるタンパク質の同定を行った。その結果，代表的な CYP である CYP3A4 に関しては，発現誘導に関わるタンパク質 4 種類（constitutive androstane receptor , D-site-binding protein , glucocorticoid receptor alpha, pregnane X receptor)，抑制性のタンパク質の２種類（ hypoxia-inducible factor-1 alpha , interleukin-6)，自動的には分類されなかったがその他 9 種類のタンパク質を同定できた。テキストマイニングの性能評価では，辞書データベースだけで評価した再現率は 0.327 であったのに対し，辞書と文脈ルールベースとを組み合わせることで再現率が 0.924 と上昇することが示され，自然言語処理による方法が極めて有効であることが確認された。

(3) PXR と化合物の相互作用に関するテキストマイニングと構造活性相関解析：上記の解析で見つかった PXR はリガンド結合性の核内受容体であり,CYP3A4 をはじめとして多くの薬物代謝酵素やトランスポータの発現調節に関わっている。PXR と介した CYP の発現変動による薬物相互作用も知られており，この基質選択性を明らかにすることは重要な課題である。化合物―タンパク質間相互作用のテキストマイニングシステムにおいて，相互作用分類に関する keyverb リストをカスタマイズし解析を行ったところ， PXR に関する PubMed アブストラクト 1727 個の中から 868 個のレコードが抽出された。重複を除き精査したところ，107 個の化合物が見出された。

このうち，human PXR に関するアゴニストは 67 個，非アゴニストは 15 個であった。構造活性相関を明確にするためには，正例と負例データがバランスよく必要なため，既報告の論文や PubChem BioAssay データベースに収載されるスクリーニングデータから更なる情報収集を行い，最終的に 270 個のアゴニストと 248 個の非アゴニストからなる大規模なデータベースを構築した。

構造活性相関解析にあたり，まず，これらの化合物の化学構造から ADMET Predictor を用いて分子記述子を計算した。主成分分析により記述子情報を要約し，化合物全体構造に関してアゴニストと非アゴニスト間での違いを検討したところ，2 次元主成分プロットにおける両者の分布にあまり違いは認められず，マクロな物性はほぼ同様であることが明らかとなった。次に，全データを訓練用データセット(正例 217 個，負例 198 個)と外部テスト用データセット(正例 53 個，負例 50 個)に分割し，再帰分割法に基づく構造活性相関解析を行なった。訓練用データセットの解析を行い，交叉検証法により木の大きさを最適化した結果，５つの分岐ルールからなる決定木が得られた（図２）。訓練用データセットに対する分類正確度は 79.0%であり，図３に視覚的表現からも明らかなように，非常に単純なルールによる高精度の分類が可能であった。分岐ルールに選ばれている記述子の多くは化合物の電気的特性を表すものが多かった。炭素原子のπFukui インデックス（Pi_AFPIc）や分子中の芳香族結合の割合（F_AromB）の値が大きいものが PXR アゴニストになりやすい傾向が認められることから，PXR の活性中心に存在する芳香族アミノ酸とのπ－π相互作用が結合に大きく関係していることが示唆された。また，中性溶液中での分子型分率が小さい，すなわちイオン型が多いものは PXR アゴニストになりにくく，PXR の活性中心の多くが疎水性アミノ酸で構成されているという知見との良好な対応関係が示された。

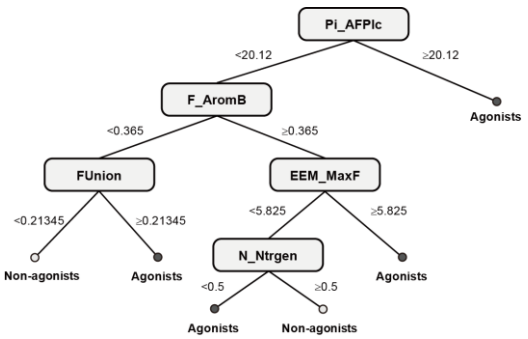

図２ PXR のアゴニストと非アゴニストを分類する決定木．化合物の分子記述子は ADMET Predictor により計算されるものである．

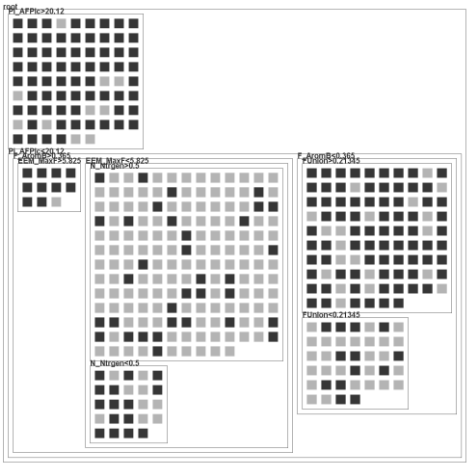

図３　決定木による PXR アゴニストの分類結果の可視化．図２で示される決定木による階層的分類を，アイコンを入れ子状の長方形枠で囲むことで表現している。

５．主な発表論文等

〔雑誌論文〕（計３件）

1. Shuya Yoshida, Fumiyoshi Yamashita, Takayuki Itoh, Mitsuru Hashida, Structure-activity relationship modeling for predicting interaction with pregnane X receptor by recursive partitioning. Drug Metabolism and Pharmacokinetics, in press (2012), doi:10.2133/dmpk.DMPK-11-RG-159, 査読有.

2. Fumiyoshi Yamashita, Chunlai Feng, Shuya Yoshida, Takayuki Itoh and Mitsuru Hashida, Automated information extraction and structure-activity relationship analysys of cytochrome P450 substrates. Journal of Chemical Information and Modeling, 51, 378-385 (2011), doi:10.1021/ci100334z, 査読有.

3. Fumiyoshi Yamashita, Takayuki Itoh, Shuya Yoshida, Mohammad K. Haidar and Mitsuru Hashida, A novel multi-dimensional visualization technique for understanding the design parameters of drug formulations. Computers & Chemical Engineering, 34, 1306-1311 (2010), doi:10.1016/j.compchemeng.2009.07.002, 査読有.

〔学会発表〕(計５件)

1. Fumiyoshi Yamashita, Shuya Yoshida, and Mitsuru Hashida, In Silico Prediction of Pregnane X Receptor Activators by Classification Tree Model, The 8th AFMC International Medicinal Chemistry Symposium, 2011/12/1, 京王プラザホテル(東京都).

2. Fumiyoshi Yamashita, Data Mining and Visualization Techniques for ADME Screening, 2011 Spring International Convention of The Pharmaceutical Society of Korea, 2011/4/22, Busan Exhibition and Convention Center(Busan).

3. Fumiyoshi Yamashita, Data Mining and Visualization Techniques for ADME Screening, FIP Pharmaceutical Sciences World Congress, 2010/11/17, The Ernest N. Morial Convention Center(New Orleans).

4. Fumiyoshi Yamashita, Chunlai Feng, Shuya Yoshida, Mitsuru Hashida, Automated knowledge acquisition and structure-activity relationship analysis regarding cytochrome P450 metabolism. 18th European Symposium on Quantitative Structure-Activity Relationships, 2010/9/22, The Rodos Palace International Convention Centre(Rhodes).

5. 吉田秀哉, 山下富義, 橋田充, CYP－化合物間相互作用情報のテキストマイニングと構造活性相関解析,日本薬物動態学会第２５回年会, 2010/10/8, 2010, 大宮ソニックシティ (埼玉県).

〔図書〕(計２件)

1. 山下富義, In silico によるバイオアベイラビリティの予測,薬物の消化管吸収予測研究の最前線, (杉山雄一 監修), メディカルドゥ, 30-36 (2010).

2. 山下富義,薬物体内動態と DDS 機能のコンピュータシミュレーション, 図解で学ぶ DDS(橋田　充監修,高倉喜信編),じほう, 17-19 (2010)

〔その他〕
http://dds.pharm.kyoto-u.ac.jp/Dds_Home/index.htm

６．研究組織

(1)研究代表者
山下　富義 (YAMASHITA FUMIYOSHI)
京都大学・大学院薬学研究科・准教授
研究者番号：30243041

(2)研究分担者
伊藤　貴之 (ITO TAKAYUKI)
お茶の水女子大学・大学院人間文化創成科学研究科・准教授
研究者番号：80401595

(3)連携研究者
馮　春来 (FENG CHUNLAI)
京都大学・大学院薬学研究科・教務補佐員
研究者番号：40456835