様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２１ 日現在

機関番号：１４２０１
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ１６０７５

研究課題名（和文）生物学的利用率の予測モデル構築：精査済みIn vitroデータからの転移学習

研究課題名（英文）Development of bioavailability prediction model: Transfer leaning model constructed from curated in vitro data

研究代表者
江崎　剛史（Esaki, Tsuyoshi）

滋賀大学・データサイエンス学系・准教授

研究者番号：２０７１７８０５

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：創薬の初期段階で薬物の性質を予測し、有望な化合物の合成や実験に必要な費用と労力を削減する効率的な創薬が求められている。現在市販されている医薬品の多くは経口投与薬であるため、経口投与可能な化合物を予測することは重要である。「薬としてのポテンシャル」を予測するために、経口投与可能な化合物を予測するBA（生物学的利用率）の予測モデルを構築し、創薬の効率化に貢献することを目的とする研究を開始した。BAに関係の深いと推定される代謝安定性、膜透過性、水溶性の大規模なデータの収集と精査、複数の予測手法や記述子の検討、転移学習モデルの構築が行い、BAを予測できる深層学習モデルの構築を行った。

研究成果の学術的意義や社会的意義
近年の科学技術の発展や研究開発費の高騰などに伴い、新薬創出におけるアカデミアの担う役割の重要性が認識されつつある。しかし、アカデミアで合成された薬効を示す化合物の多くが、薬としては使いにくい課題を抱えており、効率よく創薬に繋がっているとは言い難い。創薬初期において様々な予測モデルが構築されてきたが、最終的に必要なのは「ヒトに使えるポテンシャル」の判定に寄与することであり、in vivoの試験結果であるヒトBAが非常に有効なパラメータであると考えた。本計画は、創薬の効率化に対する非常に重要なアプローチであり、今後の予測モデル構築にも有効となる新たな視点を獲得できると考えられる。

研究成果の概要（英文）：Efficient drug discovery that predicts the properties of drugs in the early stages and reduces the costs and efforts involved in synthesizing and experimenting with promising compounds are highly sought after. Predicting compounds that can be administered orally is crucial since a significant number of commercially available drugs are administered orally. To predict the "potential as a drug," a prediction model for the biological availability (BA) of orally administrable compounds has been developed, aiming to contribute to the streamlining of drug development. The research involved the collection and analysis of extensive data related to BA, including estimated factors such as metabolic stability, membrane permeability, and water solubility. Multiple prediction methods and descriptors were explored, and a transfer learning model was constructed to build a deep learning model capable of predicting BA.

研究分野：ケモインフォマティクス

キーワード：機械学習　ADMET　in silico

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

創薬には膨大な労力と費用が掛かっており、1 つの医薬品研究開発にかかる年月は 10-18 年、費用は 2500 億円以上にのぼるとの報告がある。従来は、ある標的分子に作用する可能性がある化合物候補を合成し、薬物動態や安全性に関する試験を行い、薬としてのポテンシャルが低い（溶解性や膜透過性が低い、または毒性があるなど）と判断された化合物を除外していく。そして残った化合物は合成して構造を最適化し、再度試験を行う。この作業を繰り返し、治験薬として創り上げていく。この従来法では、大量の化合物を準備することから始まり、多くの実験を行って薬を絞り込むため、膨大な労力と費用がかかる。近年、創薬の効率化を図るための手法として、化合物の構造から薬としての性質を予測する in silico モデルが注目されてきている。化学構造だけから薬の性質を予測できれば､化合物を合成する前に｢薬としてのポテンシャル｣を推定し、可能性がある化合物だけを合成し、実験できるため、合成や実験に必要な費用や労力を大きく削減できる。

現在市販されている医薬品の 9 割が経口投与薬であるという報告があることから、創薬の初期段階で経口投与可能な化合物を予測できれば、研究開発に係る期間の大部分を短縮することにつながると予想される。これらを予測するためには、薬が腸管から血液に届く割合を示す、生物学的利用率（Bioavailability、BA）を予測することが有効である。BA は経口投与した薬物が腸管で吸収され、小腸や肝臓の初回通過効果を受けずに最終的に循環血に届く割合であり、薬の効果に直結するパラメータである。BA は、吸収率（Fa）、腸管および肝利用率（Fg、Fh）の積で表されるため、どれか 1 つでも低い値であると BA は低くなる。BA が低い薬は血液に届きにくく効果が期待できないため、大量に投与することで BA の低さはカバーできるが、ほとんどが排泄されるため非常に効率が悪い。そのため、ヒトの BA を創薬初期の段階で予測することが、効率的な医薬品開発に必要不可欠である。しかし、ヒト BA を予測できるモデルは高額な商用ソフトウェアにしか組み込まれておらず、全ての創薬研究者が手軽に利用できる環境にないのが現状である。そのため、BA を精度良く予測できる無償のモデルが望まれている。

創薬を支援するための情報科学的手法は多く開発されているが、実際に継続的に使われているものは少なく、これらを創薬の現場に浸透させることは情報科学における重要な課題である。本研究課題では創薬分野の研究者と議論を重ね、成果物が創薬において継続して使われるツールとなることを目指した。

２．研究の目的

創薬の初期段階で化合物の「薬としてのポテンシャル」を予測することは、創薬を効率的に進めるために非常に重要である。近年は研究開発費の高騰などに伴い、新薬創出におけるアカデミアの担う役割の重要性が認識されつつあるが、アカデミアで合成した化合物の多くが、水に溶けない・毒性があるといった課題があり、現状創薬に繋がっているとは言い難い。情報科学の進歩により、化合物の構造のみから様々な薬物動態パラメータの予測モデルが構築されてきたが、そのほとんどが商用ソフトウェアでのみ使用可能である。従って、誰でも使用可能な「薬としてのポテンシャル」を予測するモデルを構築することが急務である。

そこで本研究の目的として、BA の予測モデルを構築して公開することを設定し、アカデミアを含む創薬の効率化に貢献することを考えた。

３．研究の方法

本研究では期間中に、Sol や Papp の予測手法、組み込む層の検討を繰り返して行い、BA の予測に最適な深層学習構造を決定する。研究期間である 3 年間のうち、3 つの計画を立て、初年度でデータ収集（1）と記述子の検討（2）、中間年度はデータの拡充（1）とモデルの最適化（3）、最終年度は構築したモデルの改良（3）を継続しながら公開に向けて取り組んだ。

(1) 最大規模の精査されたデータを用いた予測モデルの構築

今まで報告されていたヒト BA の予測モデルは、文献から BA のデータだけを集めて機械学習を行っており、データ数は多くて数百程度であった。これらの文献から収集しているデータは元の論文を確認すると、BA ではなく Fa や皮下注射での利用率など異なるデータが学習に使用されているモデルもある。そこで本研究課題では､複数のデータベース(ChEBML､PharmaPendium､GOSTAR､ADMEDatabase）から独自に精査して収集した大量データを使用してモデルを構築することを提

案した。収集するデータは、吸収や代謝に関係のある in vitro 試験の溶解性（Sol）、Caco-2 細胞を使用した膜透過性（Papp）、ヒト肝ミクロソームを使用した代謝安定性（CLint）、そして in vivo 試験で得られたヒト BA とする。既に CLint と Papp は精査したデータを収集しており（Esaki T., et al., Mol. Inf., 2019; Esaki T., et al., J. Pharma. Sci., 2019）、他の in vitro 試験データも既に大部分が収集できている（現時点の化合物数が Sol: 57,859、Papp: 4,415、CLint: 5,278）。ヒト BA のデータは ChEBML、PharmaPendium、GOSTAR からそれぞれ 500～800 の化合物を収集できていたことから、重複を考慮しても 1,000 を超える化合物数となることが見込まれ、従来の BA 予測モデルと比べて最大規模のデータとなると想定された。構築したモデルの精度が十分でないときは、Fg と Fh の関連性から CYP3A の基質性も組み込むことを検討した。

(2) 記述子（物理化学的性質、Finger Print、Graph Convolution）を算出

化合物の特徴量として、3 種類の記述子を算出して BA の予測に適した方法を検討した。1 つ目は物理化学的な記述子、2 つ目は化合物の部分構造を表す Finger Print、そして 3 つ目は各原子と繋がりをグラフとして記述子する Graph Convolution とした。物理化学的な記述子と Finger Print は創薬の in silico モデル構築に従来から使われており、Solubility や CLint の予測で有効性が示されている。しかし、近年はこれらだけでは十分に予測できないパラメータもあり、記述子の限界が叫ばれている。そこで、Graph Convolution を用いて従来法とは異なる原子間の情報を追加する。BA 予測に有効な記述子と組み合わせを比較し、検討を行った。

(3) in vitro 試験結果を組み込んだ転移学習モデルの構築

複数の関連試験データを組み込むために、本研究課題では深層学習を拡張した、事前学習やマルチタスク学習などの転移学習（Transfer Learning, TL）を組み合わせて適用することとした。事前学習は似た性質を持つパラメータで事前に学習を行い、その後に目的のパラメータでモデルの構築を検討した。マルチタスク学習は、あるパラメータに対して学習した結果を引き継ぎながら次のパラメータを予測する学習を繰り返してモデルの改良を目指した。これらは他パラメータの情報を活かしながら効果的に学習できるという利点がある一方で、学習が非常に複雑になるといった問題点がある。本研究課題で予測を目指す BA は様々な要素が関わる複雑なパラメータであり、高い精度で予測することが困難であるが、これらの学習法を有効に組み合わせ、高精度で予測するモデルの構築を目指した。

４．研究成果

データ収集（1）に関する成果

本研究課題では、複数のデータベース（ChEBML、ADMEDatabase など）から独自に精査して収集した大量データを使用したモデルの構築を目指した。収集したデータは、吸収や代謝に関係のある in vitro 試験の溶解性（Sol）、Caco-2 細胞を使用した膜透過性（Papp）、ヒト肝ミクロソームを使用した代謝安定性（CLint）、そして in vivo 試験で得られたヒト BA とした。初年度の終了時点で、CLint と Papp の精査したデータを、他の in vitro 試験データも既に大部分が収集できた。また、代謝に関する情報を拡充することを目指し、代謝酵素の基質性を評価した実験データの収集も行うことができた。

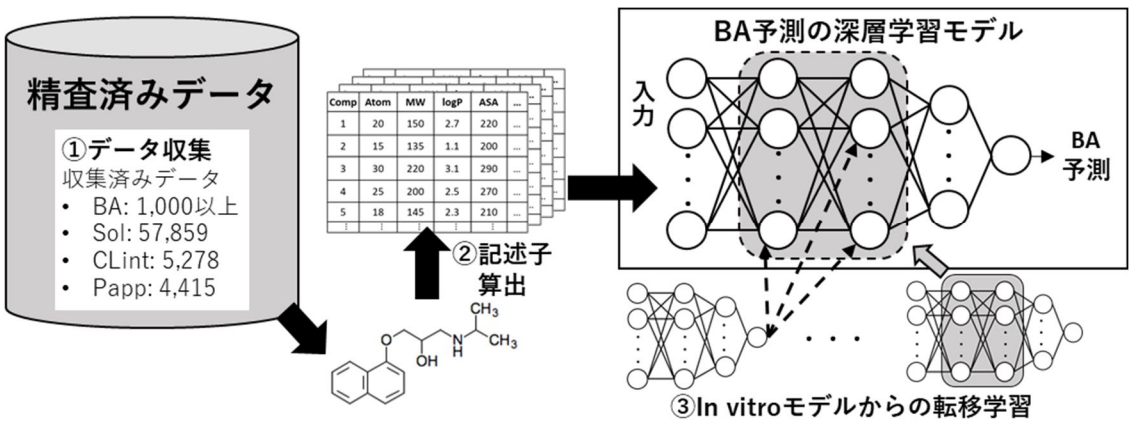

図. BA 予測の深層学習モデルの構築の概要。

記述子の検討（2）に関する成果

化合物の特徴として物理化学的な記述子、化合物の部分構造を表す Finger Print については、予測モデルの実装に向けた検討を行うことができた。それに加え、各原子と繋がりをグラフとして記述子する Graph Convolution を実装することも可能とした。これにより、BA 予測に有効な記述子と組み合わせを比較検討する環境を整えることができた。

これら複数の関連試験データを組み込むために、本研究課題では深層学習を拡張した転移学習を適用することとした。転移学習を実装するためのモデルの骨格を構築することができ、予測モデルの検証を開始することができた。

モデルの改良（3）に関する成果

化合物の特徴として、各原子と繋がりをグラフとして記述子する Graph Convolution の実装を行い、BA 予測を行うことができた。深層学習の枠組みも構築し、最終年度で特徴量とモデル構造を検討し、適したパラメータの選択を繰り返し、最適なモデルの構築に向けて検討を継続している。引き続き、モデルの公開に向けて検討を進めたいと考えている。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　2件／うちオープンアクセス　2件）

１．著者名 Tsuyoshi Esaki, Kazuyoshi Ikeda	４．巻 23
２．論文標題 Difficulties and prospects of data curation for ADME in silico modeling	５．発行年 2023年
３．雑誌名 Chem-Bio Informatics Journal	６．最初と最後の頁 1～6
掲載論文のDOI（デジタルオブジェクト識別子） 10.1273/cbij.23.1	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Esaki Tsuyoshi、Yonezawa Tomoki、Yamazaki Daisuke、Ikeda Kazuyoshi	４．巻 22
２．論文標題 Prediction Models for Fraction of Absorption and Membrane Permeability using Mordred Descriptors	５．発行年 2022年
３．雑誌名 Chem-Bio Informatics Journal	６．最初と最後の頁 46～54
掲載論文のDOI（デジタルオブジェクト識別子） 10.1273/cbij.22.46	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計3件（うち招待講演　3件／うち国際学会　0件）

１．発表者名 江崎剛史
２．発表標題 創薬の加速を目指した人工知能による薬物特性予測
３．学会等名 Digital Pharmacology Conference（招待講演）
４．発表年 2022年

１．発表者名 江崎剛史
２．発表標題 インシリコ創薬を加速するデータとモデルの扱い方
３．学会等名 CBI学会2022年大会（招待講演）
４．発表年 2022年

１．発表者名 江崎剛史
２．発表標題 データ駆動型創薬の加速を目指した薬物特性の予測　－データ収集から予測まで－
３．学会等名 日本環境変異原ゲノム学会（招待講演）
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関