様式３

# 論　文　内　容　の　要　旨

| 氏　名　（　森脇　寛智　） | |
|---|---|
| 論文題名 | 定量的構造-活性相関における手法開発 |

論文内容の要旨

定量的構造-活性相関（QSAR）研究の代表的な方法として記述子ベースの QSAR がある。記述子ベースの QSAR ではまず化合物の構造から記述子と呼ばれる化合物の特徴を表す値を複数算出し、その記述子を機械学習の入力とする事で予測モデルの構築を行なう。この様な記述子ベースの QSAR はその簡便さから頻用されるため、これまでに記述子を計算するためのソフトウェアが多数開発されている。しかし、既存のソフトウェアにはそれぞれに十分な量の記述子を計算できない、インストールや使用方法が困難である、商用であるため公開するシステムに使用しにくい、バグが存在する、などといった欠点が存在した。そのため、第一章 分子記述子計算ソフトウェア mordred の開発では我々はこれらの欠点を可能な限り改善した新しい記述子計算ソフトウェアである mordred を開発した。また、他の QSAR 研究の方法として 3D-QSAR が存在する。3D-QSAR では記述子の代わりに分子力場を用いて分子とプローブと呼ばれる仮想的な原子との相互作用エネルギーを用いてモデルの構築を行なう。3D-QSAR の代表的な手法である CoMFA ではデータセット中の化合物同士を重ね合わせる必要があるため骨格に共通部分を持つデータセット以外には適用できず、更に人手での重ね合わせ操作が必要であり、これによって予測精度が左右される。そのため、これまでに VolSurf や GRIND、Anchor-GRIND といった人手での操作を低減あるいは不要とした手法が提案されている。また、多層ニューラルネットワークは近年様々な分野で従来の手法を大きく上回る精度を達成しており盛んに研究が行なわれている。第二章 畳み込みニューラルネットワークによる3D-QSARでは多層ニューラルネットワークの入力として3D-QSARで用いられる分子相互作用場（Molecular interaction field; MIF）を使用する事で、3D-QSARの欠点である位置合わせの必要性を解消する事を試みた。また、構築されたモデルにおいてどの原子がモデルの判断に重要であったかを可視化する手法も合わせて提案した。

第一章で開発した mordred は単一のコマンドで簡単にインストールを行なう事が可能であり、シンプルな Web アプリケーションやコマンドラインアプリケーションに加えて柔軟性に優れた Python 2/3 両対応のライブラリとして使用する事ができる。Python ライブラリとして使用すると、デフォルトでは計算が行なわれない範囲の記述子（例えば14員環を計数する記述子など）の計算を追加で行なう事が出来る。また、記述子同士で単項、二項演算を行なう事が可能であるため、線形モデルで特徴量同士の関係を考慮する際に使用される交互作用項といった、記述子同士の計算から得られる特徴量を新たな記述子として容易に使用できる。またオープンソースソフトウェアで最も多い1,875個の記述子を計算できる PaDEL-Descriptor に匹敵する1,825個の記述子を計算する事が可能である。更に Chi や Detour Matrix、Molecular Framework、Molecular ID といった計算コストが高く既存のソフトウェアではタイムアウトが発生する可能性のあった記述子を中心に計算アルゴリズムの改善を行なった結果、maitotoxin といった巨大な低分子に対しても計算を行なう事が可能になった。アルゴリズムの改善を行なった記述子を単独で計算した場合だけでなく、全ての記述子を並列計算した場合においても既存のソフトウェアである PaDEL-Descriptor と比較して2倍以上高速かつ良好なスケーリング性能を示し、6並列での計算では4倍以上のスループットを達成した。mordred は3条項 BSD ライセンスを採用したオープンソースソフトウェアとして https://github.com/mordred-descriptor で配布している。

第二章では、MIF と畳み込みニューラルネットワークによる QSAR モデルの作成手法の提案を行なった。データセットに Fontaine らによって Anchor-GRIND の研究に使用された benzamidine 骨格を有する第 Xa 因子阻害剤のデータセットを使用した。このデータセットは Fontaine らによってテストセットとトレーニングセットに分割されているため、Anchor-GRIND での正答率（88%）と直接比較を行なう事が可能である。データセットの各化合物に対して RDKit-2017_03_3 を使用し、水素原子の付加、電荷の中和、三次元構造の生成、MMFF94 力場による構造最適化を行なった後、MOPAC-2016 を使用し、更に PM7 法による構造最適化を行なった。各構造をランダムに回転させる事によってデータ拡張を行ない、データセットの構造数を100倍に拡張した。これらの構造に対し、MMFF94 力場を用いて MIF を算出した後、スケーリングを行ない、平均値を0、分散を1とした。学習には mxnet-0.10.0 を使用し、VGG モデルを

参考にネットワーク構造を決定した。ハイパーパラメータは100点のランダムサーチを行ない最適な値を決定した。重要かどうかを判別したい原子を除いた仮想的な分子に対する MIF を計算し、その予測精度がどの程度低下するかにより重要な原子の可視化を行なった。比較のため、分子記述子を使用したモデルを作成した。

提案手法のテストセットに対する正答率は91.7%となった。これによってAnchor-GRIND の88%を上回り、精度の高いグリッドに依存しない 3D-QSAR 手法として使用できる事が示された。また記述子を使用したモデルの正答率（90.3%）と比較しても高い精度を示した。

Protein Data Bank（PDB）に第Xa因子または類似のセリンプロテアーゼであるトリプシンとの複合体構造が登録されている化合物に対する重要な原子の解析を行なった結果、データセット中の共通骨格であるbenzamidine骨格のamidine部分がモデルの判別に重要である事が示された。他の強調された部分についてもそれぞれ結合に重要なトリプシンの残基との間に関連が見られ、たんぱく質の構造をモデル構築に用いていないにも関わらず、モデルの判別に重要な部分とたんぱく質との結合に重要な部分との間での関連が示唆された。この事から、特にたんぱく質構造が未知のデータセットに対する解析に有用であると考えている。

本研究ではQSAR分野の発展に資す事を目的に記述子ベースのQSAR、3D-QSARそれぞれに対するソフトウェア、手法の開発を行った。これらは共にQSARを含む幅広い分野での応用が可能であるため、創薬やレギュラトリーサイエンスへの貢献を期待する。

様式 7

# 論文審査の結果の要旨及び担当者

| 氏　名　（森脇　寛智） | | |
|---|---|---|
| | (職) | 氏　名 |
| 論文審査担当者 | 主 査　教授 | 高木　達也 |
| | 副 査　教授 | 大久保　忠恭 |
| | 副 査　教授 | 小比賀　聡 |

## 論文審査の結果の要旨

定量的構造-活性相関は化合物から生理活性や毒性、物性などを予測する研究であり、候補医薬品の活性予測や環境毒性予測などで幅広く用いられている。QSARは大別して記述子ベースのQSARと3D-QSARに分けられる。申請者は記述子ベースのQSAR、3D-QSARの両方の発展に資する事を目的に手法やソフトウェアの開発を行った。

記述子ベースのQSARで使用される分子記述子を計算するためのソフトウェアはこれまでに多く開発されてきており、その中でもPaDEL-Descriptorは1800個以上の記述子を計算する事が出来、多くのインターフェースを提供し、300回以上引用されている著名なソフトウェアであり、オープンソースの記述子計算ソフトウェアの中では大敵なソフトウェアの一つであると言える。しかし、申請者はPaDEL-Descriptorにいくつかのバグを発見したため、使用に際しては慎重な検証が必要であると考えている。そこで、第一章では新たな分子記述子計算ソフトウェアであるMordredの開発を行った。Mordredでは既存のソフトウェアでは十分な速度で計算を行なう事ができない幾つかの記述子についてその計算アルゴリズムの改善を行なった。例えば、Detour Matrix記述子では入力分子を関節点によって小さなフラグメントに分割する事で効率的に計算を行なう。この改善により一般的な性能の計算機を用いて自然界で最も大きな低分子化合物であるmaitotoxinに対して約1.2秒と現実的な時間で計算を完了させる事が可能となった。また、PaDEL-Descriptorとの性能比較を行なった。原子数に依存した計算速度の増加具合のベンチマークでは特に大きな分子についてPaDEL- Descriptorより遥かに高速かつタイムアウトを起こさず計算が完了した事を示し、並列計算のベンチマークではより良いスケーリング性能や6並列の場合で4倍以上の性能の改善を示した。MordredはPython2及び3のライブラリとして使用可能で、1800を越える記述子を計算でき、CLI、Webインターフェースを提供する。Mordredは3条項BSDライセンスでリリースされている。

3D-QSARモデルは化合物の三次元構造に対して分子力場を使用して計算したポテンシャルエネルギー (Molecular Interaction Field; MIF)などを使用して構築する。代表的な3D-QSARモデルであるCoMFAではデータセット中の全ての化合物について原子座標での重ね合わせが必要とされるため、共通した骨格を持つデータセットについてのみしか適用できず、手動での操作が必要となる。この様な欠点を解消するため、これまでにVolSurfやGRIND、Anchor-GRINDなどが報告されてきた。申請者はMIFを畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) の入力として用いる事で分子の重ね合わせを不要としたモデル構築手法を提案した。データセット中の構造に対し、MMFF94力場を用いて MIF を算出し、VGG モデルを参考にしたニューラルネットワークの学習を行なった。また、重要かどうかを判別したい原子を除いた仮想的な分子に対する MIF を計算し、その予測精度がどの程度低下するかにより重要な原子の可視化を行なった。提案手法のテストセットに対する正答率は91.7%となり、Anchor-GRIND や記述子ベースのQSARを上回り、精度の高いグリッドに依存しない 3D-QSAR 手法として使用できる事が示された。重要な原子の解析を行なった結果、データセット中の共通骨格であるbenzamidine骨格のamidine部分がモデルの判別に重要である事が示された。他の強調された部分についてもそれぞれ結合に重要なトリプシンの残基との間に関連が見られ、特にたんぱく質構造が未知のデータセットに対する解析に有用である。

両者は共にQSARを含む幅広い分野での応用が可能であるため、創薬やレギュラトリーサイエンスへの貢献が期待でき、博士（薬科学）の学位論文に値するものと認める。