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　６ 年　 ６ 月 １２ 日現在

機関番号：１４６０３
研究種目：若手研究
研究期間：2020～2023
課題番号：２０Ｋ１９９２２

研究課題名（和文）分子の三次元構造情報を利用した機械学習モデルに基づく分子設計手法の開発

研究課題名（英文）Development of Chemical Structure Generation Method Based on Three-dimensional Molecular Representation

研究代表者
宮尾　知幸（Miyao, Tomoyuki）

奈良先端科学技術大学院大学・データ駆動型サイエンス創造センター・准教授

研究者番号：２０８２３９０９

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）：分子構造の表現に焦点を当てて、コンピュータ上で有機低分子を設計する手法の開発研究を行なった。公共データベースに含まれる活性化合物データを利用した解析から、分子構造の二次元表現が三次元表現より活性値予測という点では優れていた。このような表現を使用したモデルを組み込むことができる構造生成器として仮想反応に基づく生成器を構築した。加えて、異なる実験条件下で計測された実験結果を有効活用するためには、すべてのデータを用いた適切なカーネル関数を利用した非線形回帰モデルが適していることを明らかにした。

研究成果の学術的意義や社会的意義
分子は三次元空間に存在しているため三次元表現（空間的や電子的な情報）を利用した方が薬理活性を予測するモデルの精度は高くなると期待されていたが、今回の検証では、分子構造から活性を予測する場合には、二次元表現が三次元表現より高い予測精度を示した。この知見に基づいた計算コストの低い二次元表現の改良研究や、今回の研究で行ったように、モデルを分子構造生成器に組み込んだ実用的な分子設計などにつながると考える。また、一つの実験グループで取得できるデータ数には限りがある。様々な実験グループから集めたデータを統合してモデル構築する際に、今回の研究成果が指針として役立つと考える。

研究成果の概要（英文）：The development of chemical structure generation methods for organic small molecules was conducted, focusing on molecular representation. Two-dimensional molecular representation was found to be superior to three-dimensional one from a retrospective analysis of potency prediction tasks using a publicly available database. A structure generator utilizing virtual chemical reactions was implemented, where regression models using the representation above can be incorporated. Furthermore, to utilize data sets from different experimental conditions but for the same target, non-linear regression with an appropriate kernel function using simply entire data points was a suitable approach.

研究分野：化学情報学

キーワード：分子表現　構造活性相関　構造生成

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

データ駆動による分子設計では、分子構造から活性・物性を予測する機械学習モデル (定量的構造活性相関（QSAR）モデル、定量的構造物性相関（QSPR）モデル)を利用する。化合物は記述子と呼ばれる数値ベクトルに変換され、記述子と目的変数の関係がモデル化される。研究当初においては、QSAR/QSPR モデルに基づく分子設計の課題として、「分子の二次元構造のみに基づく記述子を利用した分子設計」であり「複数の活性・物性を満足する分子設計ではない」の２点を挙げた。分子は構造式で表現されることもあるが、本質的には三次元空間に存在しているため、方向性のある水素結合などの立体的・空間的情報が考慮できていない可能性がある。また、異なる実験条件からの統合モデル、加えて複数の目的変数に対するモデルに基づく分子設計では、そもそもデータを統合してモデル構築することが適切かを判断する必要があった。以上の２点を解決する機械学習モデリング手法を確立できれば、データ駆動による分子設計を一段階引き上げることにつながる。

２．研究の目的

(1) 必要性を理解した上で、分子構造の三次元情報を考慮した、統計モデルに基づく分子設計手法を開発すること

(2) 複数の異なる実験条件を統合したモデル構築手法の開発、さらには複数の活性・物性を満足する分子構造を統計モデルに基づいて設計する手法開発

３．研究の方法

(1) 分子の三次元構造情報を利用した分子設計法

立体配座に基づく分子表現の必要性を評価するために、複数の活性化合物データを利用した活性予測モデル構築を行う。この結果で三次元構造が必要と判断された場合には、三次元分子表現に基づくモデル構築を行い、モデルに基づき分子設計（構造生成）を行う手法を提案する。

(2) 統合モデルの構築

複数の統計モデル統合手法の開発と適用範囲の把握を行う。実験条件が異なるデータを統合するモデル構築としては、異なるアッセイの情報を統合する効果を阻害定数（IC50）予測モデルの精度比較から検証する。この検証は活性値予測であるので、回帰モデルを構築することになり、適切なモデリング手法を併せて考案する。次に複数の目的を同時に満たす化合物提案のためのモデリング手法を検討する。それぞれのモデルは異なるデータから構築されているため、適用範囲を考慮することが必要になると考える。

４．研究成果

(1) 分子の三次元構造情報と二次元情報の比較

１０種類の活性化合物に対しての活性予測モデル構築を構築した。二次元分子表現としては、Molecular Operating Environment (MOE) software で計算可能な記述子 (MOE 記述子)、原子環境を表現したフィンガープリントである extended connectivity fingerprint with a diameter of 4(ECFP4)、三次元記述子としては、複数のクエリとなる化合物の立体配座に対しての重なりの程度をスコア化した similarity profile (SP)を用いた。SP には、分子形状の重なり(SP(Shape))と水素結合供与基などの化学プロパティの重なり (SP(Color))、その両方 (SP(Combo))の３種類を利用した。これらの分子表現を線形回帰モデリング手法である partial least square regression、非線形回帰モデリング手法である random forest, support vector regression (SVR), multi-layer perceptron neural network と組み合わせることで予測モデルを構築した。各標的化合物に対して、トレーニング：テスト＝７：３にランダムに分割をし、統計的な評価をするために 5 回試行を繰り返した。それぞれのデータセットの大きさ（化合物の数)、結果の一部（４つの標的、SVR）であるモデルの精度を root mean square error (RMSE) 基準で**表１**に示す。

表１：テストデータに対する RMSE

標的マクロ分子	データ数	SP (Shape)	SP (Color)	SP (Combo)	ECFP4	MOE
Acetylcholinesterase	266	1.25	1.13	1.08	0.93	1.06
Kappa opioid receptor	1526	0.75	0.72	0.72	0.68	0.84
Coagulation factor X	1436	0.81	0.76	0.75	0.75	0.88
Muscarinic acetylcholine receptor M3	372	1.76	1.50	1.45	1.23	1.89

ランダム分割のデータに対して、三次元分子表現の二次元表現に対する優位性は確認すること

ができなかった。ECFP4 の精度がどの標的マクロ分子に対しても高く活性予測に関しては原子環境を構造式上で表現した二次元表現を利用すべきとの知見が得られた。
さらに、この結果の妥当性を検証するために、トレーニングデータの数を減らした場合やテストとなる分子それぞれに対して局所モデルを構築するなどの工夫を行なったが、三次元記述子の二次元記述子に対する優位性を示すことはできなかった。つまり、上記の結論は今回試したモデリング手法とデータセットによらず一貫した傾向を観測することができた。つまり、高活性分子を設計するためのモデルとして、分子の二次元構造に基づく特徴量を用いるべきという指針が得られた。

(2) 二次元表現に基づく構造生成器の開発
構造式（二次元情報）に基づく分子構造生成器として、生成器に制約条件や複雑柔軟に組み込むことができる遺伝的アルゴリム（分子改変）による構造生成器を開発した。既往の手法としてはヒュリースティックな構造改変ルールを実装したものが複数提案されているが、今回構築した構造生成器は構造改変ルールを既存の化合物データベースから自動抽出する仕組みを備えている。コンピュータにおける仮想反応を利用していることから、化学的な特徴を捉えた構造の生成が期待された。この構造生成器に、二次元記述子を利用した活性予測モデルを組み合わせることで活性化合物を効率的に生成することができると考える。

(3) 解釈可能な機械学習モデル構築手法の開発
課題を進めていく中で、QSAR/QSPR モデルの解釈性を担保した上で活性化合物を設計する仕組みが重要であると認識した。解釈可能なモデルであれば、分子構造をどのように変換するかの指針を得ることができ、(2) で開発した分子構造生成器にその指針を組み込むことができる。
そこでシンボリック回帰モデルによるモデンリング手法を考案した。シンボリック回帰は数式として目的変数を回帰する。既存のシンボリック回帰はデータセットに対しての適合（フィッティング）に焦点を当てた方法であり、解釈性は必ずしも高くない。そこで、３種類のフィルターを導入したシンボリック回帰モデリング手法を考案した。一つ目のフィルターは関数フィルター（**図１**）であり、階層的に特定の演算が適用された数式は削除される。二つ目は変数フィルターであり、一つの変数が複数項に現れる式を削除する。三つ目のフィルターは値域フィルターであり、テストデータに対しての予測値が想定される値域から外れる式を削除する。このように三つのフィルターを導入することで解釈性の高い数式としての予測モデル構築手法を考案した。このモデルを利用することで、線形モデルに限定することなく解釈可能なモデルを構築することができる。

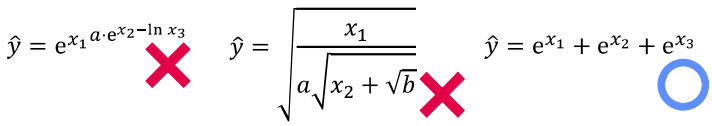

図１．候補式に対して関数フィルターの適用

なお、この手法の有用性は、類縁体評価に焦点を当てた構造活性相関データに加え目的変数を分子構造から計算することができる QED（ドラッグライクネスを表現するスコア）と SA スコア（合成可能性スコア）を用いた検証から確認した。
さらに、シンボリック回帰モデルの汎化性能を向上させるために、回帰係数の変動、記述子の変動に対して式が安定となるための評価関数をモデリングに組み込むとともに、ドメインフィルター（上述）を改良した。改良後のドメインフィルターでは、定義域を探索領域として値域の最大、最小値を数値計算する。この結果、想定する値域から外れる構造をフィルタリングする。これらの改良を適用したモデルは、類縁体化合物の活性予測に対して、既存手法を上回る汎化性能を示した。

(4) 異なる実験条件を統合するモデル開発
研究開始当初は、対象とする複数の物性もしくは活性を予測する複数のモデルを単純に組み合わせることを想定していた。しかし、実際には同一のエンドポイント（測定対象）であっても実験系が異なれば結果の意味は異なり、化学的に意味のあるモデルを構築するためには、一段低い階層（実験系の違い）におけるモデル構築手法を検討する必要があった。そこで、５０%阻害濃度（IC５０）を対象とし、ChEMBL データベースに含まれる特定の標的マクロ分子に対して複数の実験条件下での活性情報が登録されている化合物と IC５０値を機械的に抽出し統合モデルを構築した。モデル構築手法としては、１：IC５０値の補正なく使用、２：複数の実験系にて評価された共通化合物の IC５０値に基づく補正（Scaling）、３：補正ではなく化合物のランキングを使用した評価の三つを試した（上記の１と２は既往研究）。結果として、類似した化合物を含む実験系の全てのデータを用いた Ranking-support vector machine と単純に全ての活性化合物のデータをモデル構築のためのデータとし、Tanimoto kernel による SVR 回帰モデルを構築する方法が予測精度の観点から優れていた。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Raku Shirasawa, Katsushi Takaki, Tomoyuki Miyao	４．巻 9
２．論文標題 Generalizability Improvement of Interpretable Symbolic Regression Models for Quantitative Structure Activity Relationships	５．発行年 2024年
３．雑誌名 ACS Omega	６．最初と最後の頁 9463-9474
掲載論文のDOI（デジタルオブジェクト識別子） 10.1021/acsomega.3c09047	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 －

１．著者名 Takaki Katsushi、Miyao Tomoyuki	４．巻 2
２．論文標題 Symbolic regression for the interpretation of quantitative structure-property relationships	５．発行年 2022年
３．雑誌名 Artificial Intelligence in the Life Sciences	６．最初と最後の頁 100046～100046
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.ailsci.2022.100046	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 －

１．著者名 Matsumoto Katsuhisa、Miyao Tomoyuki、Funatsu Kimito	４．巻 6
２．論文標題 Ranking-Oriented Quantitative Structure?Activity Relationship Modeling Combined with Assay-Wise Data Integration	５．発行年 2021年
３．雑誌名 ACS Omega	６．最初と最後の頁 11964～11973
掲載論文のDOI（デジタルオブジェクト識別子） 10.1021/acsomega.1c00463	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 －

１．著者名 Sato Akinori、Miyao Tomoyuki、Jasial Swarit、Funatsu Kimito	４．巻 35
２．論文標題 Comparing predictive ability of QSAR/QSPR models using 2D and 3D molecular representations	５．発行年 2021年
３．雑誌名 Journal of Computer-Aided Molecular Design	６．最初と最後の頁 179～193
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10822-020-00361-7	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計2件（うち招待講演　1件／うち国際学会　1件）

1．発表者名 Tomoyuki Miyao
2．発表標題 Global Interpretation of Regression Models for Quantitative Structure-Property Relationship
3．学会等名 第7回ケモインフォマティクス秋の学校（招待講演）（国際学会）
4．発表年 2022年

1．発表者名 佐藤彰准，宮尾知幸，Swarit Jasial，船津公人
2．発表標題 二次元分子表現と三次元分子表現を用いたQSAR/QSPRモデルの予測能力の比較
3．学会等名 第43回ケモインフォマティクス討論会
4．発表年 2020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