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

平成　３０　年　　６　月　２０　日現在

機関番号：　１３９０４
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００００５

研究課題名（和文）化学構造情報の数値プロファイリングと偏最小２乗法を用いた化学物質の環境毒性予測

研究課題名（英文）Numerical profiling of chemical structure and prediction of environmental toxicity of chemical substances using partial least squares method

研究代表者

高橋　由雅（Takahashi, Yoshimasa）

豊橋技術科学大学・工学（系）研究科（研究院）・教授

研究者番号：００１４４２１２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、化学物質の生態環境毒性予測の問題に対し、トポロジカルフラグメントスペクトル（topological fragment spectra, TFS）から導出されるPLS潜在変数を用いたパラメータ・フリーのQSARモデリング手法（TFS-PLS法）を提案し、実データを用いてその有用性を検証した。また、予測化合物ごとに構造類似性を基礎とした訓練集合のアクティブサンプリングと組み合わせることで、より精度の高い予測結果が得られることを示した。

研究成果の概要（英文）：In this research, for ecological toxicity prediction of chemical substances, we proposed a parameter-free QSAR modeling method (TFS-PLS method) using PLS latent variables derived from topological fragment spectra (TFS). We verified its usefulness with real data. We also showed that a more accurate prediction result can be obtained by combining the method with active sampling of training set based on structural similarity.

研究分野：　分子情報学

キーワード：　環境毒性予測　偏最小二乗法　環境毒性　PLS法　トポロジカルフラグメントスペクトル　TFS　QSAR

１．研究開始当初の背景

化学物質の有害性を評価することは環境への悪影響や健康被害を回避するために極めて重要であり、社会的要請も極めて高い。しかしながら、既存化学物質のすべてについて詳細な安全性評価試験を行うには時間も費用もの膨大なものとなり、直ちにこれを解決することは困難である。このことから、既に評価済みの既存化学物質との種々の特性等を比較することにより、種々の有害性の有無やリスクの度合いを定性的・定量的に推定し、個々の化学物質対する事前の格付けをおこない、より疑わしいものから優先的、効率的なリスク管理を行うことが急務となっている。その解決策の一つとして近年注目されているのが QSAR(定量的構造活性相関)手法の活用である。QSAR は関連する既存化学物質のデータをもとに、化学構造と活性（毒性）との関係を定量的にモデル化し、得られたモデル式を利用して未知の化合物の活性（毒性）を定量的に予測・評価しようとするものであり、既に OECD の化学物質管理プログラムにおいても QSAR や関連するケモインフォマティクス技術の積極的な活用が推進されている。[1]。

一般化学物質の毒性等、有害性予測においては、対象となる化合物が広範囲におよび、その構造的な多様性のため、必ずしもに良好な近似／予測結果を得るのは難しいのが実状である。こうした問題に対し、近年、特定の官能基などに注目した化合物の分類を基礎とした手法（カテゴリーアプローチ）[2]が注目を集めている。しかし、多様な骨格構造や多種の官能基を有する化合物群などに対しては、予測精度が大きく低下したり、カテゴリー化そのものが困難な場合も少なくない。このことから、より精度の高い予測結果を得るための手法やモデルの精緻化のため技術が改めて望まれる。

２．研究の目的

筆者らは、先に、予測結果の精緻化をはかることをねらいとし、化学物質の構造類似性を基礎に、事例データベースを利用しながら、対象化合物の化学構造に応じて、その都度、構造的に類似した化合物の事例を収集するといったアクティブサンプリングにより、クエリ近傍の局所空間に対する QSAR モデルを動的に生成し、予測を行う方法(active QSAR モデリング) を提案し、その有用性を明らかにした。[3]

QSAR は物理化学的パラメータなどを用いて統計的に得られたモデル式により、データ未知の物質の物性や毒性などを予測する方法である。これら必要なパラメータを収集あるいは種々の推算ツールを活用して事前に準備し、モデル式の開発を行う。しかし、これらのモデル式を利用して、LC50 など、目的とする毒性データの予測を行おうとする場合、その手順が煩雑であったり、必要パラメータの入手が律速となり、利用が困難な場合も少なくない。

本研究では、精度の高い予測結果を得ることをねらいとし、QSAR モデリングに対し、筆者らが別途提案したフラグメントスペクトル法（topological fragment spectra, TFS）[4]を用い、予測化合物ごとに構造類似性を基礎とした近傍空間における局所モデルを動的に生成するアクティブサンプリングの手法と PLS（偏最小二乗法）を組み合わせ、トポロジカルフラグメントスペクトルから導出される PLS 潜在変数を用いたパラメータ・フリーの QSAR モデリング手法を提案するとともに、その有用性について実データを用いて検証する。

３．研究の方法

(1)アクティブサンプリングによる動的 QSAR モデリング

事例（訓練集合）となる化合物群に対し、事前に構造特徴の離散数量化のためのプロファイリングを行い、これを化学構造、活性（毒性）データとともにデータベース化し、作成したデータベースを背景に以下の手順に従ってクエリごとにモデルの生成とデータの予測を動的に行う。

① 予測対象サンプルをクエリとして提示する。
② データベース化合物群と同じプロファイリング手法にもとづく化学構造情報の離散数量化（構造特徴パターンベクトルの生成）を行う。
③ 構造特徴空間でクエリと類似性の高い化合物を予め作成した事例データベースから探索し、モデル生成のためのデータ（訓練集合）を収集する。
④ 上記で得られたクエリ近傍サンプルからなる訓練集合をもとに予測モデルを作成する。
⑤ 作成されたモデルを用いて、クエリの目的データを予測する。

本手法は、クエリが与えられた時点で、構造特徴のプロファイリング、類似性検索による近傍データのサンプリングと訓練集合の作成、予測モデルの生成、データの予測といった一連の処理をその都度行うため、予測精度の向上が期待できるほか、データベースの追加・更新に伴う対象化合物の構造の多様化に対しても柔軟に対応できる。

(2)構造プロファイリング

化合物の構造プロファイリングには TFS（Topological Fragment Spectra）法を用いた。TFS とは化学物質の構造式から可能な部分構造を列挙し、その数値的な特徴づけにもとづいて化学物質のトポロジカルな構造プロフィールを多次元数値ベクトルとして表現しようとするものである。その生成手順は、(1)構造情報の記述表現に際しては化学構造式を原子を点(頂点)、結合を辺と見

なし、原子や結合の種類の相違を区別する重み付きグラフとして取り扱い、与えられた構造式に対応する化学グラフ(水素原子は省略)から可能なすべての部分グラフを列挙する。 (2)次に、得られた個々の部分グラフの数値的特徴づけを行う。(3)特徴づけられた個々の部分グラフ（構造フラグメント）集合に対し、特徴量に従って度数を調べ、ヒストグラムを生成する。生成されたヒストグラムは一種のデジタルスペクトルと見なすことができ、多次元数値ベクトルで表すことができる。

(3) PLS - Partial Least Squares

PLS は Wold ら[4]によって提案された回帰手法であり、変数間に高い相関関係がある場合にも予測的なモデルを構築することができるため、現在の QSAR 研究における有力な手法のひとつとなっている。

PLS モデルは **X** 変数ブロックおよび **Y** 変数ブロックごとの個別の内部相関（ブロック内相関）と、双方の変数ブロック間を結びつけるブロック間相関から構成される。ここで言う **X** 変数ブロックと Y 変数ブロックのブロック内相関は次式で表すことができる。

$$\mathbf{X} = \Sigma \mathbf{t}_h \mathbf{p}'_h + \mathbf{E} \qquad (1)$$

$$\mathbf{Y} = \Sigma \mathbf{u}_h \mathbf{q}'_h + \mathbf{F} \qquad (2)$$

ここで **t,u** は主成分の得点（score）ベクトル、**p,q** は負荷（loading）ベクトルと呼ばれる。ここでのねらいは、**Y** の情報を可能な限り記述する，すなわち|**F**|をできるだけ小さくし、かつ同時に **X** と **Y** との間の有意な関係を獲得することである。変数ブロック間の関係は **Y** ブロックの得点 **u** と **X** ブロックの得点 **t** との関係を調べることによって知ることができる。最も簡単なものは線形の関係である。

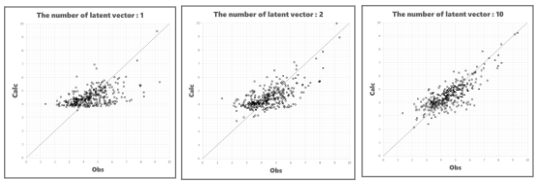
$$\mathbf{u}_h = b_h \mathbf{t}_h \qquad (3)$$

b_h は重線形回帰や主成分回帰における回帰係数の役割を果たす。しかしながら、上述の関係は完全に各変数ブロックごとに別々のアルゴリズムとして記述されている。そこで、各変数ブロックが相互の情報を交換しながらブロック間の相関情報を獲得する必要がある。そのための方法として、 PLS アルゴリズムでは **X** ブロックの得点 **t** と **Y** ブロックの得点 **u** を交換することによりこれを実現している。特徴として、データ行列 **X** の次元を縮約した潜在変数を用いて **Y** を予測することができるため、高次元特徴ベクトルとして表される TFS にも有用である。

４．研究成果

(1) 初年度（平成 27 年度）研究では、化学構造式の離散・数量化によって得られる多次元数値ベクトル（トポロジカルフラグメントスペクトル,TFS）を構造特徴変数とし、PLS 解析（偏最小二乗法）によって導出される数学的な合成変数である潜在変数（隠れ変数とも呼ばれる）を説明変数に用いて、目的とする毒性値の近似・予測モデルを生成するためのアルゴリズムの設計、ならびに必要なプログラムの開発を行うとともに、計算機への実装を行った[6]。提案手法（TFS-PLS 法）は、生成された予測モデルの利用に際しても化学構造情報（原子結合表）のみを必要とし、その他の如何なる物理化学的パラメータも必要としない、全くの parameter free の手法である。

(2) 平成 28 年度研究では，前年度（平成 27 年度）に提案した TFS-PLS 法を用い、藻類に対する短期毒性予測モデルの構築を試みた。データセットには、72 時間半数成長阻害濃度(72h-EC50)の試験データ（環境省より公開）を用いた。試験化合物に重複がある場合は新しい年度の試験結果を採用した。塩や混合物を除外し、さらに毒性値が試験上限値表記（不等号付の値）されているものを除外した全 341 化合物の毒性試験データを用いた。TFS-PLS 法による一括モデリングでの性能を調べたところ、潜在変数 20 程度で精度の向上が飽和してくるのが認められた。また、TFS-PLS による回帰性能は，logP 単回帰モデルを大きく上回る精度が得られた（図 1）。さらに、構造類似性を基礎とした Active QSAR モデリングでは、潜在変数=1 の場合にも良好な回帰モデルが得られ、近似精度も上述の一括モデリングによる結果と比べ大きく向上することを明らかにした（図 2）。[7]



潜在変数の数	相関係数	決定係数	RMSE
1	0.517	0.267	1.030
2	0.660	0.436	0.904
10	0.793	0.628	0.734
20	0.819	0.671	0.690

図1 TFS-PLS 一括モデリングの結果(左から潜在変数の数が 1,2,10,20 の回帰プロット)

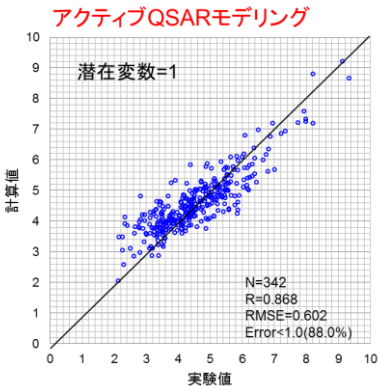


図2 TFS-PLS アクティブ QSAR モデリングの結果(潜在変数＝1のモデルによる回帰プロット)

(3) 前述の TFS-PLS 法＋Active QSAR モデリングによる魚類急性毒性(96h-LC50)予測のついても検討を行った。データセットには同じく環境省より公開されている 367 化合物の魚毒性試験データ (96h-LC50)を用いた。本研究では、アクティブサンプリングにおける最近傍サンプルの類似度に一定の閾値を導入することで、生成モデルの予測安定性を向上させることが可能であることを示した。一方、全データを用いた一括モデリングに比べ、実験値に対するより良好な近似は達成されたものの、Leave-one-out 法による予測実験の結果は期待を大きく下回るものとなった。これらの原因を調査したところ、訓練集合として用いた化合物の構造多様性が極めて大きく（環系骨格構造の 78%がそれを一度しか現れない)、leave-one-out 試験では、多くの場合に、類似性の高い良好な訓練サンプルの参照が困難であることが分った。このことから、より大規模なデータセット（環境省が公開している魚類 96 時間半数致死濃度（LC_{50}）に関する試験データ[3]に加え、産業技術総合研究所で開発を進めている汎用生態リスク評価管理ツール MeRA の有害性データベースに収録されている同データを含め、合計 1698 化合物の毒性試験データ）を用い、改めて提案手法の有用性について検証を試みた。その結果、改めて構造類似性を基礎とした訓練集合のアクティブサンプリングによる局所モデリングの有用性が示されるとともに、事例データベースにクエリとの構造類似性の高い化合物データが収録されている場合、明らかに良好な予測結果が得られることを示した[8]。

＜引用文献＞

① OECD, QSAR project, http://www.oecd.org/env/ehs/oecdquantitativestructure-activityrelationshipsprojectqsars.htm
② Recent Advances in QSAR Studies: Methods and Applications, Eds. Tomasz P., Jerzy L.、Mark T. D. C., 2010, Springer.
③ 高橋由雅，大山美香，第 21 回環境化学討論会，2A-07 (2012)
④ S. Wold et al., *Chemometr Intell Lab.*, Vol. 58, 2001, 109-130.
⑤ Takahashi Y., Ohoka H., Ishiyama Y., *Advances in Molecular Similarity*, Vol. 2, 93-104, 1998.
⑥ Yoshimasa Takahashi, Yoshitaka Inagaki, Ryota Kikuchi, PEACH; Prediction of environmental affects of chemicals, The 21th European Symposium on Quantitative Structure-Activity Relationships, 2016, Sep., Verona, Italy
⑦ 菊地亮太、高橋由雅、偏最小 2 乗法を用いた化学物質の藻類に対する短期毒性予測、2016 人工知能学会全国大会、2016 年 6 月、北九州市
⑧ 泉原拓、石井秀昇、桂樹哲雄、高橋由雅、アクティブ QSAR モデリングを用いた TFS-PLS 法による魚毒性予測、第 45 回構造活性相関シンポジウム講演要旨集、pp. 85-86、2017 年 11 月、土浦市

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計　5　件）

① 泉原拓、石井秀昇、桂樹哲雄、高橋由雅、アクティブ QSAR モデリングを用いた TFS-PLS 法による魚毒性予測、第 45 回構造活性相関シンポジウム講演要旨集、pp. 85-86、2017 年 11 月、土浦市
② Yoshimasa Takahashi, Ryota Kikuchi, Algal toxicity prediction of chemicals using TFS-PLS method in conjunction with active QSAR modeling, EuroTox2017, 2017, Sep.; *Toxicology Letters*, **280(S1)**, 316(2017)
③ Yoshimasa Takahashi, Yoshitaka Inagaki, Ryota Kikuchi, PEACH; Prediction of environmental affects of chemicals, The 21th European Symposium on Quantitative Structure-Activity Relationships, 2016, Sep., Verona, Italy
④ 菊地亮太、高橋由雅、偏最小 2 乗法を用いた化学物質の藻類に対する短期毒性予測、2016 人工知能学会全国大会、2016 年 6 月、北九州市
⑤ Ryota Kikuchi, Tetsuo Katsuragi, Yoshimasa Takahashi, Fish toxicity prediction of chemicals using TFS-PLS method in conjunction with active QSAR modeling, Proc. of the 44th Symposium on Structure-Activity Relationships, pp. 85-86, 2016, Nov., Kyoto.

６．研究組織

(1)研究代表者

高橋由雅　（TAKAHASHI Yoshimasa）
豊橋技術科学大学・大学院工学研究科・教授
研究者番号：００１４４２１２