3E3-2

偏最小二乗法を用いた化学物質の藻類に対する短期毒性予測

Prediction of Short-term Toxicity to Algae of Chemical Substances by Using Partial Least Squares Regression.

菊地 亮太[*1] 桂樹 哲雄[*1] 高橋 由雅[*1]
Ryota Kikuchi Tetsuo Katsuragi Yoshimasa Takahashi

[*1] 豊橋技術科学大学大学院工学研究科 情報・知能工学専攻
Department of Computer Science and Engineering, Toyohashi University of Technology

Hazard prediction models of chemicals using statistical methods and machine learning are desired because safety evaluations of new chemical substances to human health and the environment take a lot of time and cost. In this study, we attempt to predict a short-term toxicity of chemical substances to algae by using TFS-PLS method. TFS is a kind of molecular fingerprints, which was developed in our laboratory. We described molecules in training dataset by TFS and employed them for making regression models for the toxicity prediction by using partial least squares method (PLS). We analyzed the experimental results of 341 compounds taken from the results of Eco-toxicity tests of chemicals conducted by Ministry of the Environment in Japan. The TFS-PLS models validated the regression ability and the prediction capability by Leave-One-Out Cross-Validation method.

1. はじめに

新規化学物質について，人体や環境への安全性を正確に調べるには，コストや時間がかかる．そこで一般化学物質の生態環境毒性予測に関して，近年，OECD の化学品安全性管理プログラムによって QSAR（Quantitative Structure-Activity Relationship）の積極的な活用が推奨され，活発な研究が進められている．特に，魚類やミジンコに対する急性毒性に関しては疎水性パラメータ logP(P はオクタノール-水分配係数)との密接な関係が知られており，様々な化学物質に対する統計的な予測モデルが報告されている．しかし，統計的手法や機械学習などを利用した様々な研究にも関わらず，藻類に対する毒性予測においては必ずしも良好な予測モデルが報告されていない．このことから，藻類に対する急性毒性については，より精度の高い予測手法やモデルの開発が切望されている．

本研究では，別途，当研究室で考案し，機械学習による薬化合物の活性クラス分類などでの有用性が示されている Topological Fragment Spectra(TFS) [Takahashi 98]を用いた詳細な構造情報の記述と偏最小二乗法(Partial Least Squares : PLS)を用いた藻類に対する短期毒性予測モデルの構築を試みた．

2. データセットおよび方法

2.1 データセット

データセットとして，環境省から公開されている藻類に対する速度法，72 時間半数成長阻害濃度(72h-EC50)の試験データ（実施年度，平成 7 年から平成 26 年度）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．

2.2 Active QSAR モデリング

筆者らは先に，QSAR によるデータ予測の精緻化のためのアプローチのひとつとして，予測対象であるクエリ化合物の化学構造とよく似た（類似性の高い）化学構造を有する近傍サンプルを事例 DB から探索し，これを訓練集合としてその都度 QSAR モデルを生成し，データ予測を行う Active QSAR モデリングを提案している[西野 07]．この方法を適用することで，それぞれのクエリで，データセット中にある類似した構造のみを使った局所的なモデリングをすることが可能なため，利用可能なすべてのデータを用いて一括モデリングを行う従来の QSAR モデリングを上回る予測精度が期待できる．(図 1, 2)

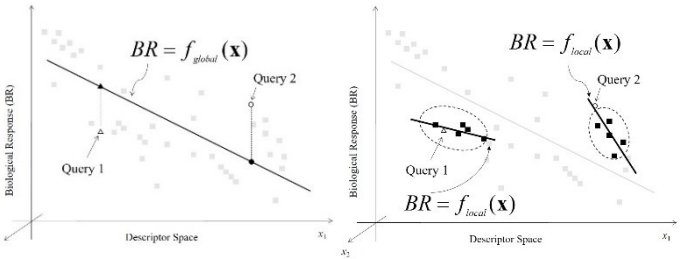


図 1 QSAR

図 2 Active QSAR

2.3 構造プロファイリング

TFS（Topological Fragment Spectra）は，化学構造の構造類似性評価を目的に，当研究室で考案された化学物質の構造情報を離散・数量化し多次元数値ベクトルとして記述する手法である．以下に，TFS の生成方法と部分構造を次数の和で特徴付けた時の TFS 生成例を図 3 に示す．

（1）対象とする化学構造の可能な部分構造をすべて列挙
（2）列挙した各部分構造に対する数値的な特徴付け
（3）特徴付けの値とその出現頻度による多次元パターン表現

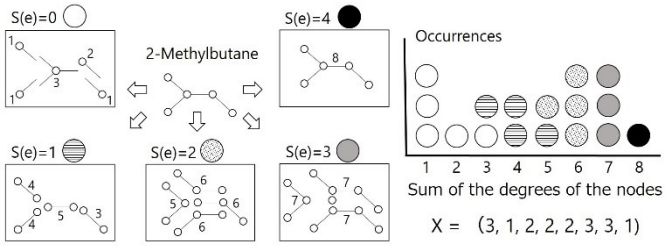


図 3 TFS の次数和に基づくフラグメント生成例

連絡先：菊地亮太，〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1 豊橋技術科学大学大学院 情報・知能工学系，
Tel: 0532-44-6878, kikuchi@mis.cs.tut.ac.jp

2.4 PLS - Partial Least Squares Regression

PLS 回帰は，Wold [Wold 01]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．

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

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

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．

3. 結果と考察

はじめに，比較のため，魚類やミジンコに対して良好な近似／予測性能が知られている疎水性パラメータ logP を用いた QSAR 解析を試みた．次に，TFS を記述子とした PLS モデリング（TFS-PLS）について検討し，最後に TFS-PLS 法を Active QSAR モデリングと組み合わせて用いた場合の近似精度について検討を行った．

3.1 logP による古典的 QSAR 解析

各化合物の logP の値には Ghose らの原子フラグメント法[Ghose 89]による推算値を用いた．図4に logP を記述子とした単回帰モデルによる，データセット中の logP が推算可能な全 313 化合物に対する計算値と実測値の相関プロットを示す．得られたモデルによる実測値との相関係数，決定係数および RMSE 値はそれぞれ 0.417，0.174，1.062 であった．RMSE の値に注目すれば，その計算値の標準誤差が log スケールで 1.0 以上であることから必ずしも良好な近似精度を示すものでない．

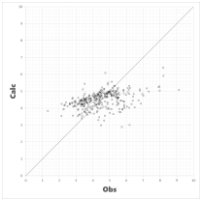
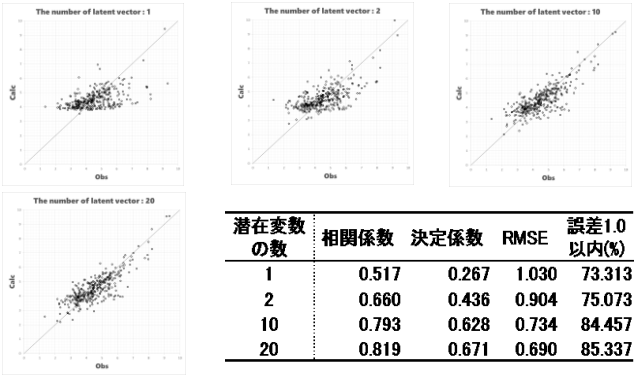


図 4 logP 単回帰モデルによる実測値と計算値の相関プロット

3.2 TFS-PLS モデリングによる結果

QSAR によるデータセット中の一括モデリングで作成された TFS-PLS モデルの潜在変数の数を 1,2,10,20 と変化させた時の回帰能力の変化を図5のプロットで示す．QSAR による TFS-PLS では回帰能力に関して，logP 単回帰モデルを大きく上回る精度が得られた．予測に際しての潜在変数の数については，オーバフィッティングを避けるためのさらなる検討が必要である．



潜在変数の数	相関係数	決定係数	RMSE	誤差1.0以内(%)
1	0.517	0.267	1.030	73.313
2	0.660	0.436	0.904	75.073
10	0.793	0.628	0.734	84.457
20	0.819	0.671	0.690	85.337

図 5 QSAR による TFS-PLS 結果(左から潜在変数の数が 1,2,10,20 の回帰プロット)

3.3 Active QSAR モデリングによる結果

QSAR と同様に潜在変数の数を変化させながら，回帰能力の変化を確認した．モデリングの際の近傍サンプル数は潜在変数の数の 5 倍取得する．Active QSAR では，潜在変数の数が 1 でも良好な回帰能力をもつモデルが作成できていることがわかった．また，近似精度も上述の一括モデリングによる一般的な QSAR モデルと比較して大きく向上しており，Active QSAR と組み合わせることで，その有用性が大いに期待できることを示唆している．

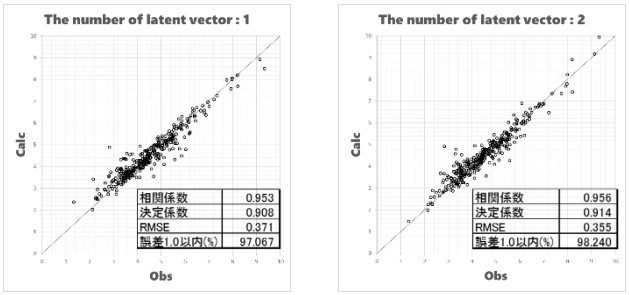


図 6 Active QSAR による TFS-PLS 結果（左から潜在変数の数が 1,2 の回帰プロット）

4. まとめ

本研究では藻類の短期毒性予測問題に対し，構造特徴の記述に TFS 法を用いて多次元特徴ベクトルとして記述するとともに，偏最小二乗法を用いた Active QSAR モデリングによって，従来の古典的 QSAR モデルに比べ，格段に良好な近似が可能であることを示した．引き続き，提案手法での予測能力の検証と更なる予測精度の向上に向けた工夫を進めていきたい．

謝辞

本研究は JSPS 科研費 15K00015 および日本化学工業協会 LRI 研究助成を受けて実施したものである．

参考文献

[Takahashi 98] Takahashi, Y., Ohoka, H., Ishiyama, Y.: Structural Similarity Analysis Based on Topological Fragment Spectra, Advances in Molecular Similarity, Vol.2, pp.93-104 (1998)

[西野 07] 西野達也, 藤島悟志, 高橋由雅: 構造類似性に基づく Active QSAR モデリング, 2007 年度人工知能学会全国大会(第 21 回)論文集 (2007)

[Wold 01] Wold, S., Sjöström, M., Eriksson, L.: PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.*, Vol.*58*, No.2, pp.109–130 (2001)

[Ghose 89] Ghose, A. K., et al.: Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certai, *J. Chem. Inf. Comput. Sci.*, Vol.29, pp.163–172 (1989)