Efficient Fish Acute Toxicity Prediction Using QSAR Machine Learning Pipeline

이명훈
숭실대학교 의생명시스템학부

ABSTRACT

기계 학습의 발전으로 수생 생물에 대한 화학 물질의 독성 평가를 위한 정량적 구조-활성 관계(Quantitative Structure-activity Relationship, QSAR) 모델 개발이 지속해서 연구되고 있다. 따라서 본 연구에서는 Japanese rice fish(*Oryzias latipes*)에 대한 선행 연구의 급성 수생 독성 데이터를 기반으로 pLC_{50} 예측을 위한 QSAR 모델 개발 파이프라인을 제안한다. 이를 위해 분자 descriptor 생성과 모델 개발에 필요한 다양한 접근 가능한 소프트웨어를 이용하였으며, 어류의 급성 독성 예측의 주요 descriptor를 분석하였다. 그뿐만 아니라, 선행 연구에서 보고한 성능보다 우수하고 안정적인 모델을 최종적으로 제시하였다.

KEYWORDS

Fish acute toxicity, Machine Learning, QSAR, QSTR, Feature selection

1. INTRODUCTION

과학과 기술의 발달로 전 세계적으로 많은 화학 물질이 사용되어 수생 생물에 대한 독성 중독 가능성이 크게 우려되고 있다[1]. 화학 물질의 독성 평가는 시장에 출시되기 전에 모든 화학 산업에서 필요하다[2]. 전통적으로 화학 물질의 독성은 동물 실험을 이용해 평가되었다. 그러나 이러한 독성 실험은 윤리적으로 문제가 있을 뿐만 아니라 비용, 노동, 그리고 시간이 많이 소요되는 방법이다[3,4]. 따라서 이를 대신할 수 있는 정량적 구조-활성 관계(Quantitative Structure-activity Relationship, QSAR) 모델이 독성 메커니즘을 분석하고 합성되지 않은 미지의 유기물에 대한 수생 독성(aquatic toxicity)을 예측하는 중요한 방법으로 자리 잡았다[5-8]. 특히 많은 선행 연구들이 Fathead minnows(*Pimephales promelas*)의 급성 독성(acute toxicity, LC_{50}, $\log LC_{50}$, $-\log LC_{50}$(pLC_{50}))에 대한 QSAR 모델 연구를 수행하였다. 이는 기계 학습의 발전로 250개 미만의 적은 데이터 세트에서도 효과적인 모델 개발이 가능하여 지속할 수 있었다[9].

본 연구의 목적은 선행 연구인 Furuhama *et al.*[10]에서 공개한 109개의 96-h pLC_{50} 데이터 세트를 이용하여 Japanese rice fish(*Oryzias latipes*)에 대한 급성 독성에 관한 선행 연구보다 개선된 QSAR 모델을 개발하는 것이다. 또한, 이를 통해 급성 수생 독성을 위한 QSAR 모델 개발에 적용될 수 있는 효과적인 파이프라인을 **그림 1**과 같이 제시한다.

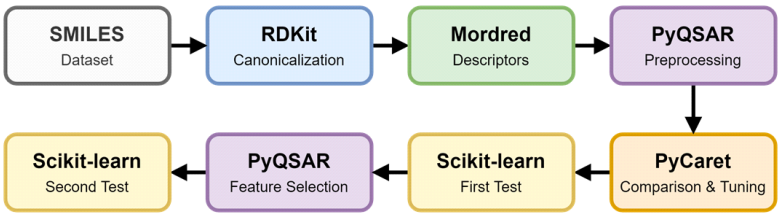


그림 1. 본 연구의 QSAR 모델 개발 파이프라인

2. METHOD

Data preprocessing

QSAR 모델 개발을 위한 데이터 세트는 Furuhama *et al.*에서 공개한 109개의 "Training set"를 이용하였다. 일본 환경부에서 측정한 Japanese rice fish(*Oryzias latipes*)에 대한 급성 독성 실험 데이터이며, 96-h pLC_{50} 값, SMILES, 그리고 관련 descriptor가 QsarDB[11]에 공개되었다. 최종 데이터 세트의 pLC_{50}는 **그림 2**와 같은 분포를 갖는다.

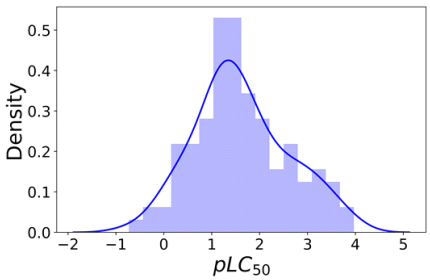


그림 2. Furuhama *et al.*의 데이터 세트에 대한 pLC_{50} 분포

해당 데이터는 SMILES(Simplified Molecular-input Line-entry System)를 RDKit[12]을 이용해 canonicalization하였으며, Mordred [13] 분자 descriptor 계산기를 이용해 2차원의 descriptor를 생성하였다. Mordred는 1,800개 이상의 2차원 및 3차원 descriptor를 계산할 수 있는 분자 descriptor 계산 소프트웨어이다. 다른 소프트웨어(예: Cinfony[14], ChemoPy[15])에 비해 설치와 사용이 쉽고 빠른 계산 속도로 유연하게 사용할 수 있다. 또한, 많이 이용되는 PaDEL-Descriptor는 몇몇 분자에서는 잘못된 값을 생성한다고 보고되었다[13]. 그러나 Mordred의 CPSA 및 MoRSE와 같은 3D 구조 descriptor는 재현이 어려운 최적화 문제를 유발할 수 있다[3]. 따라서 본 연구에서는 Mordred를 이용하여 인접행렬, 거리행렬, SlogP 등 총 1,613개의 2차원 descriptor만을 계산하였다. 그리고 생성된 descriptor는 PyQSAR[16]를 이용하여 유효하지 않으면 삭제되었으며, 최종적으로 1235개의 descriptor가 이용되었다.

QSAR pipeline

최근에는 기계 학습의 발전으로 교육을 받지 않았거나 시간과 노력의 부족한 경우에도 알고리즘 구현이 가능한 AutoML(Automated Machine Learning)이 이용될 수 있게 되었으며,

PyCaret[17]이 그 대표적인 예이다. 따라서 본 연구에서도 PyCaret을 이용한 최적의 모델 선택과 hyperparameter tuning이 이루어졌다. 결과적으로 Bayesian Ridge (BR) 회귀 모델을 채택하였으며, Scikit-learn[18]에 의해 재구현되었다. 모델의 기본적인 성능 평가는 R^2와 σ (standard deviation)로 이루어졌으며, 추가적으로 MSE(Mean Squared Error)와 MAE(Mean Absolute Error)를 이용해 이루어졌다. 모든 성능 지표는 데이터 세트의 구성을 8:2로 각각 다르게 훈련 세트와 검증 세트로 나누어 10회 측정하였다.

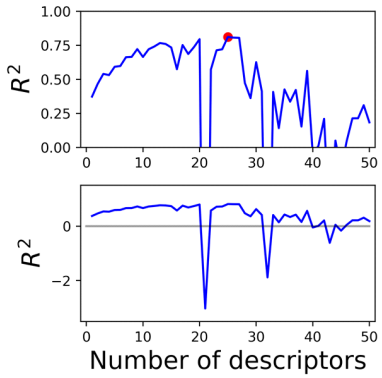


그림 3. PyQSAR를 이용한 feature selection과 descriptor 개수에 따른 BR 모델의 R^2. 위는 유효한 R^2의 범위를 나타내며, 아래는 전체 결과이다. 빨간 점은 최고 성능을 나타낸다.

Feature selection

Mordred에 의해 계산된 1235개의 descriptor는 feature selection을 통해 불필요하거나 중요하지 않은 descriptor을 줄이고 최적의 하위 집합을 찾아 훈련 속도와 예측 정확도를 높이고자 하였다. feature selection은 RFECV(Recursive Feature Elimination with Cross Validation), GA(Genetic Algorithm) 그리고 PCC(Pearson Correlation Coefficient)를 통해 이루어질 수 있다. 본 연구에서는 PyQSAR를 이용해 계층적 클러스터링(Hierarchical Clustering)와 GA를 이용한 feature selection 기능을 이용하였다. 이러한 방법은 속도를 높이고 모델의 성능을 개선하는 데 도움이 된다. 해당 과정은 계층적 클러스터링를 통해 descriptor를 클러스터링하여 탐색을 위한 공간을 줄이고, 각 클러스터에서 GA를 이용한 최상의 descriptor가 선택된다[16]. 이를 통해 1235개의 descriptor를 1개에서 50개까지 개수를 지정하여 BR 모델 기반으로 feature selection 과정을 수행하였다.

3. RESULT AND DISCUSSION

Feature importance

PyQSAR를 이용한 feature selection은 40개 이상의 descriptor가 선택되면 R^2는 0에 가까워지는 것을 **그림 3**에서 확인할 수 있다. 반면에 20개 이하의 descriptor가 선택되는 구간에서는 R^2가 점차 증가하는 것을 알 수 있다. 결과적으로 R^2가 가장 높게 측정된 25개의 descriptor가 최종 선택되었다. 25개의 descriptor는 다음과 같다: AATS2se, AATS3i, AATS4pe, AATSC1pe, AATSC6c, ATS7m, ATSC0c, ATSC3dv, BCUTare-1h, BCUTc-1h, C3SP2, EState_VSA7, GATS1c, GATS2i, GATS6pe, IC3, MATS2d,

MATS6are, NdCH2, PEOE_VSA7, SIC3, SM1_Dzi, VSA_EState5, piPC10, piPC7. 또한, 25개의 주요 descriptor 중에서 PCC > 0.35의 경우는 다음과 같다(**그림 4**): SIC3 0.368, ATS7m 0.360, piPC10 0.343, piPC7 0.341, PEOE_VSA7 0.305.

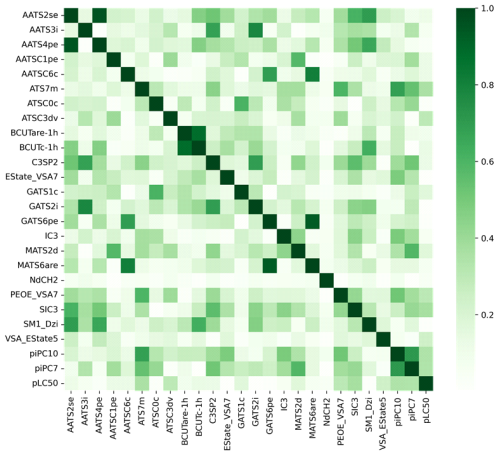
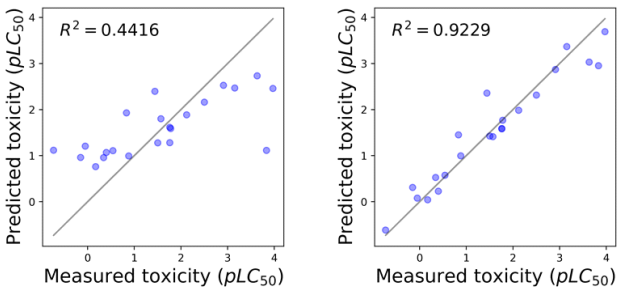


그림 4. PyQSAR에 의해 선별된 25개의 주요 descriptor와 pLC_{50}의 PCC 시각화

Model comparison

본 연구는 1) feature selection 이전의 1235개의 descriptor에 대한 결과, 2) feature selection 이후의 25개 descriptor에 대한 결과, 그리고 3) Furuhama *et al.*에서 공개된 결과를 비교하였다(**그림 5**). feature selection에 따른 결과는 이전의 R^2가 -0.042로 0보다 작으며, 이후 25개의 descriptor를 이용한 경우에는 0.757로 분명하게 개선되었다. 따라서 σ, MSE, 그리고 MAE도 25개만을 이용한 모델이 우수하였다. 이는 feature selection 이전의 결과보다 feature selection 이후의 결과가 분명하게 개선되었다는 것을 확인할 수 있다. 또한, 1235개의 descriptor를 이용하는 것보다 25개의 주요 descriptor를 이용하는 것이 우수하므로 feature selection 과정이 중요하다는 점을 시사한다. 뿐만 아니라, 본 연구의 R^2 0.773은 선행 연구의 0.757보다도 높으며, σ는 0.485에서 0.103으로 모델이 안정화되었다.

(a)



(b)

	Reference	Before	After
R^2	0.757	-0.042	**0.773**
σ	0.485	0.585	**0.103**
MSE	-	0.577	**0.305**
MAE	-	0.722	**0.162**

그림 5. (a) feature selection 전 1235개의 descriptor와 후 25개의 descriptor를 이용한 RB 모델의 예측 결과의 예시. **(b)** 선행 연구와 feature selection 전/후 결과 비교.

4. CONCLUTION

본 연구의 목적은 Japanese rice fish에 대한 급성 독성 예측을 위한 QSAR 모델을 개발하는 것이며, 이를 위해 BR 모델에 Mordred와 PyQSAR를 함께 이용하여 109개의 pLC_{50} 데이터 세트에 대한 성능을 평가하고 주요 descriptor를 분석하였다. 결과적으로 선행 연구보다 개선된 성능을 확인할 수 있었으며, QSAR 모델을 개발을 위한 파이프라인을 **그림 1**과 같이 제시하였다.

화합물의 다양한 물성이 어류에 대한 급성 독성에 영향을 미치지만, 특히 선택된 25개의 descriptor는 Japanese rice fish의 96-h pLC_{50} 예측을 위한 RB 모델 개발에 효과적으로 이용되었다. 또한, 25개의 descriptor 중에서 SIC3, ATS7m, piPC10, piPC7, 그리고 PEOE_VSA7가 PCC에 의해 pLC_{50}와 상관 관계가 있는 주요 descriptor인 것을 확인하였다. 비록 109개의 적은 데이터 세트에서 검증된 결과이지만, 해당 파이프라인은 많은 양의 데이터 또는 pLC_{50} 외의 다양한 분자의 물성 예측에 QSAR 모델 개발로 이용될 수 있다. 결과적으로 본 연구는 파이프라인에 의해 개발된 모델이 기존의 선행 연구 결과보다 우수하고 안정적임을 확인할 수 있었다. 해당 QSAR 개발 파이프라인은 Github repository에서 사용할 수 있다 (https://github.com/mhlee216/QSAR_Machine_Learning_Pipeline).

REFERENCES

[1] A. Stenzel, U. K. Goss and S. Endo, Determination of polyparameter linear free energy relationship (pp-LFER) substance descriptors for established and alternative flame retardants, Environ. Sci. Technol., 2013, 47, 1399-1406.

[2] X. Yu, Prediction of chemical toxicity to Tetrahymena pyriformis with four descriptor models, Ecotoxicol. Environ. Saf., 2020, 190, 110146.

[3] B. Peric, J. Sierra, E. Marti, R. Cruanas and M. A. Garau, Quantitative structure-activity relationship (QSAR) prediction of (eco)toxicity of short aliphatic protic ionic liquids, Ecotoxicol. Environ. Saf., 2015, 115, 257-262.

[4] V. Drgan, S. Zuperl, M. Vracko, F. Como and M. Novic, Robust modeling of acute toxicity towards fathead minnow (Pimephales promelas) using counter-propagation artificial neural networks and genetic algorithm, SAR QSAR Environ. Res., 2016, 27, 501-519.

[5] W. R. Brogan III and R. A. Relyea, Multiple mitigation mechanisms: effects of submerged plants on the toxicity of nine insecticides to aquatic animals, Environ. Pollut., 2017, 220, 688-695.

[6] C.-W. Cho and Y.-S. Yun, Application of general toxic effects of ionic liquids to predict toxicities of ionic liquids to Spodoptera frugiperda 9, Eisenia fetida, Caenorhabditis elegans, and Danio rerio, Environ. Pollut., 2019, 255, 113185.

[7] S. K. Heo, U. Safder and C. K. Yoo, Deep learning driven QSAR model for environmental toxicology: effects of endocrine disrupting chemicals on human health, Environ. Pollut., 2019, 253, 29-38.
[8] D. Wang, Q. Ning, J. Dong, B. W. Brooks and J. You, Predicting mixture toxicity and antibiotic resistance of fluoroquinolones and their photodegradation products in Escherichia coli, Environ. Pollut., 2020, 262, 114275.
[9] X. Chen, L. Dang, H. Yang, X. Huang and X. Yu, RSC Adv., 2020, 10, 36174-36180.
[10] A. Furuhama, K. Hasunuma and Y. Aoki, SAR QSAR Environ. Res., 2015, 26, 301-323.
[11] V. Ruusmann, S. Sild and U. Maran, J. Cheminform., 2015, 7, 32.
[12] G. Landrum, RDKit, http://www.rdkit.org, (accessed November 14, 2021).
[13] H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, J. Cheminform., DOI:10.1186/s13321-018-0258-y.
[14] O'Boyle, N.M., Hutchison, G.R. Cinfony - combining Open Source cheminformatics toolkits behind a common interface. Chemistry Central Journal 2, 24 (2008). https://doi.org/10.1186/1752-153X-2-24.
[15] D.-S. Cao, Q.-S. Xu, Q.-N. Hu and Y.-Z. Liang, Bioinformatics, 2013, 29, 1092-1094.
[16] S. Kim and K.-H. Cho, Bull. Korean Chem. Soc., DOI:10.1002/bkcs.11638.
[17] PyCaret, https://pycaret.org/, (accessed November 14, 2021).
[18] Pedregosa F, Varoquaux, Ga"el, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825-30.