Korean Chem. Eng. Res., 54(5), 621-629 (2016)
http://dx.doi.org/10.9713/kcer.2016.54.5.621
PISSN 0304-128X, EISSN 2233-9558

PCBs 독성 예측을 위한 주요 분자표현자 선택 기법 및 계산독성학 기반 QSAR 모델 개발

김동우 · 이승철 · 김민정 · 이은지 · 유창규†

경희대학교 환경학 및 환경공학과 환경연구센터
17104 경기도 용인시 기흥구 덕영대로 1732
(2016년 4월 19일 접수, 2016년 6월 29일 수정본 접수, 2016년 7월 4일 채택)

Development of QSAR Model Based on the Key Molecular Descriptors Selection and Computational Toxicology for Prediction of Toxicity of PCBs

Dongwoo Kim, Seungchel Lee, Minjeong Kim, Eunji Lee and ChangKyoo Yoo†

Department of Environmental Science and Engineering, Center for Environmental Studies, College of Engineering, Kyung Hee University, 1732, Deogyeong-daero, Giheung-gu, Yongin, Gyeonggi, 17104, Korea
(Received 19 April 2016; Received in revised form 29 June 2016; accepted 4 July 2016)

요 약

EU의 REACH 제도 도입에 따라 각종 화학물질에 대한 독성 및 활성 정보 확보를 위해 화학물질의 분자구조 정보를 기반으로 화학물질의 독성 및 활성을 예측하는 정량적구조활성관계(QSAR)에 대한 연구가 최근 활발히 진행되고 있다. QSAR 모델에 사용되는 분자표현자는 매우 다양하기 때문에 화학물질의 물성 및 활성을 잘 표현할 수 있는 주요한 분자표현자를 선택하는 과정은 QSAR 모델 개발에 있어 중요한 부분이다. 본 연구에서는 화학물질의 분자구조 정보를 나타내는 주요 분자표현자의 통계적 선택 방법과 부분최소자승법(Partial least square: PLS) 기반의 새로운 QSAR 모델을 제안하였다. 제안된 QSAR 모델은 130종의 폴리염화바이페닐(Polychlorinated biphenyl: PCB)에 대한 분배계수(log P)와 14종의 PCBs에 대한 반수 치사 농도(Lethal concentration 50%: LC_{50}) 예측에 사용되고, 제안된 QSAR 모델 예측 정확도는 기존의 OECD QSAR Toolbox에서 제공하는 QSAR 모델과 비교하였다. 관심 화학물질의 분자표현자와 활성정보 간의 높은 상관관계를 갖는 주요 분자표현자를 선별하기 위해서, 상관계수(r)와 variable importance on projections (VIP)기법을 적용하였으며, 화학물질의 독성 및 활성정보를 예측하기 위해 선별된 분자표현자와 활성정보를 이용해 부분최소자승법(PLS)를 사용하였다. 회귀계수(R^2)와 prediction residual error sum of square (PRESS)을 이용한 성능평가결과, 제안된 QSAR 모델은 OECD QSAR Toolbox의 QSAR 모델보다 PCBs의 log P와 LC_{50}에 대하여 각각 26%, 91% 향상된 예측력을 나타내었다. 본 연구에서 제안된 계산독성학 기반의 QSAR 모델은 화학물질의 독성 및 활성정보에 대한 예측력을 향상시킬 수 있고 이러한 방법은 유독 화학물질의 인체 및 환경 위해성 평가에 기여할 것으로 판단된다.

Abstract – Recently, the researches on quantitative structure activity relationship (QSAR) for describing toxicities or activities of chemicals based on chemical structural characteristics have been widely carried out in order to estimate the toxicity of chemicals in multiuse facilities. Because the toxicity of chemicals are explained by various kinds of molecular descriptors, an important step for QSAR model development is how to select significant molecular descriptors. This research proposes a statistical selection of significant molecular descriptors and a new QSAR model based on partial least square (PLS). The proposed QSAR model is applied to estimate the logarithm of partition coefficients (log P) of 130 polychlorinated biphenyls (PCBs) and lethal concentration (LC_{50}) of 14 PCBs, where the prediction accuracies of the proposed QSAR model are compared to a conventional QSAR model provided by OECD QSAR toolbox. For the selection of significant molecular descriptors that have high correlation with molecular descriptors and activity information of the chemicals of interest, correlation coefficient (r) and variable importance of projection (VIP) are applied and then PLS model of the selected molecular descriptors and activity information is used to predict toxicities and activity information of chemicals. In the prediction results of coefficient of regression (R^2) and prediction residual error sum

† To whom correspondence should be addressed.
E-mail: ckyoo@khu.ac.kr

of square (PRESS), the proposed QSAR model showed improved prediction performances of log P and LC_{50} by 26% and 91% than the conventional QSAR model, respectively. The proposed QSAR method based on computational toxicology can improve the prediction performance of the toxicities and the activity information of chemicals, which can contribute to the health and environmental risk assessment of toxic chemicals.

Key words: REACH regulation, PCBs, Quantitative Structure Activity Relationship(QSAR), Molecular descriptors, Partial least square, Computational toxicology

1. 서 론

2007년도부터 EU의 화학물질 관리제도인 Registration, Evaluation, Authorization and restriction of Chemicals (REACH) 법령과 규정안이 EU 전체 회의에서 통과함에 따라 EU 내 제조 및 수입되는 물질들은 제조 및 수입량, 그리고 위해성에 따라 등록, 평가, 허가 및 제한을 받게 되었다[1]. 이를 시작으로 중국, 대만, 일본, 미국 등의 여러 국가들도 기존의 화학물질 관리제도에서 물질 규정에 대한 항목을 추가 및 신설하고 있다. REACH 제도는 대상 물질의 물리화학적 특성, 환경독성 및 인체 유해성 등 다양한 항목에 대해서 관리를 하고 있지만 위 항목들에 대한 실험에 있어 많은 시간과 비용이 소비된다는 단점이 있다[2].

EU 등의 국제 기구에서는 화학물질의 물성 및 독성 측정을 위한 실험에 따르는 부담을 줄이기 위해 정량적구조활성관계(Quantitative Structure Activity Relationship: QSAR)개념을 이용하여 화학물질의 물성 및 독성을 예측하고, 그 결과를 활용할 수 있도록 허용하고 있다[3]. QSAR란 화학물질의 독성 또는 활성 정도는 화학물질의 구조적 특성에 따라 결정된다는 개념으로, 기존의 알려진 화학물질의 독성 및 활성 정보를 기반으로 알려지지 않은 화학물질의 정보에 대한 예측을 가능하게 한다[4]. 현재까지 QSAR와 다양한 예측 기법들을 이용하여 화학물질의 활성치와 분자표현자들과의 관계를 성립하고 화학물질의 독성 및 물성을 예측하고자 하는 연구들이 진행되어 왔다. Song 등은[5] 무지개 송어에 대한 농약의 급성독성의 예측 및 분석을 위하여 선형 모델(다중 선형 회귀 모델) 및 비선형 모델(support vector machine, 인공 신경망) 기반의 QSAR 모델을 개발 및 비교하였다. Ammi 등은[6] 초미세여과와 역삼투압 막분리 공정에서 유기 화합물의 배출을 예측하기 위해 막 재질의 분자표현자 및 분리막 특성에 인공 신경망 개념을 적용한 QSAR 모델을 도입하였다. Kim 등은[7] QSAR 개념과 비슷한 정량적구조특성관계(Quantitative structure-property relationship: QSPR) 개념을 도입하여 난용해성 약품의 물에 대한 용해도를 예측하는 QSPR 모델을 제안하였다. 허나 기존의 연구들은 QSAR 모델을 제시할 때, 대상 화학물질의 사전지식에 근거하여 한정 그룹 내에서만 분자표현자를 선택하였으며 분자표현자 간의 다중공선성을 고려하지 못하였다는 한계점을 갖는다. 따라서 QSAR 모델의 예측력 향상을 위해 특정 그룹에 한정되지 않으며 동시에 분자표현자간의 다중공선성을 고려할 수 있는 새로운 분자표현자 선택에 대한 연구가 필요한 실정이다.

본 연구에서는 향상된 QSAR 모델 개발을 위해 QSAR 모델에 사용되는 분자표현자 선정에 있어 단변량 및 다변량 통계적 방법을 도입하여 다양한 그룹의 분자표현자들을 고려하며 동시에 다중공선성에 대한 문제를 해결하고자 하였다. 또한 선택된 분자표현자들로 기존 QSAR 모델보다 향상된 예측력을 갖는 QSAR 모델을 개발하였다. 관심 활성치와 높은 상관관계를 갖는 분자표현자들을 선택하기 위해 상관계수(r)와 Variable importance in projection (VIP)를 적용하였으며 향상된 QSAR 모델을 제시하기 위해 다변량 통계방법인 부분최소자승법(Partial least square: PLS)을 적용하였다. 본 연구에서 제시된 분자표현자 선택법 및 QSAR 모델의 성능을 파악하기 위해 기존의 QSAR 모델인 OECD QSAR Toolbox와 비교 분석을 수행하였으며 이때 성능 지표로 결정계수(R^2)와 모델예측에러를 나타내는 Prediction residual error sum of square (PRESS)를 이용하였다. 본 연구에서 제안한 분자표현자 선택 방법과 PLS 기반의 QSAR 모델은 기존 모델보다 화학물질의 활성치에 대해 보다 정확한 예측이 가능하며, 이는 각종 화학물질에 대한 인체 및 환경 위해성 검정에 기여할 수 있을 것으로 판단된다.

2. 연구이론

2-1. 정량적구조활성관계(QSAR)

정량적구조활성관계(Quantitative structure activity relationship: QSAR)는 화학물질이 보유하고 있는 물리적, 화학적, 생물학적 활성 정도는 화학물질의 분자구조와 상관관계를 갖는다는 개념이다[8]. 화학물질의 구조와 활성 정도간의 상관관계를 계산할 때 화학물질의 구조는 분자표현자라는 수치적인 값으로 표현된다. 분자표현자는 분자량부터 특정 결합의 개수, 분자간 거리 등 화학물질의 구조를 설명할 수 있는 여러 요인들을 다양한 평준화 기법을 이용하여 수치화한 값이다[9]. 측정된 화학물질의 활성치와 분자표현자간의 상관관계를 수식화한 것을 QSAR 모델이라고 명명하며 이에 대한 개념도를 Fig. 1[10]에 나타내었다. QSAR 모델은 새로운 화학물질의 관심 활성치를 예측하기 위해 사용되며 이때 분자표현자는 QSAR 모델에서 화학물질의 관심 활성치를 설명할 수 있는 주요 변수로 작용된다. QSAR 모델은 식 (1)과 같이 나타낼 수 있다.

$$\text{Properties or activity} = f(\text{Molecular descriptor}) + E_r \quad (1)$$

위 식에서 f는 분자표현자와 활성도 간의 반응함수를, E_r는 예측된 활성도와 측정값 간의 오차를 의미한다. 반응함수를 어떻게 정의하는가에 따라 다양한 QSAR 모델을 개발할 수 있으며, 어떠한 분자표현자를 선택하느냐에 따라 QSAR 모델의 유형을 정할 수 있다[11].

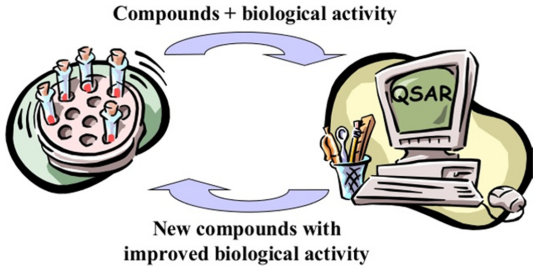

Fig. 1. The concept of QSAR model [10].

2-2. 부분최소자승법(Partial least square regression: PLS)

분자표현자들과 독성 수치 간의 관계를 나타내는 QSAR 모델을 개발하기 위해서 본 연구에서는 독립변수(X)와 종속변수(Y)의 상관관계를 규명하는데 사용하는 부분최소자승법(Partial Least Squares: PLS)을 적용하였다. 이 방법은 두 변수 간의 공분산을 최대로 하는 공간에 각 변수를 사영시키는 과정에서 다차원의 독립변수와 종속변수의 차원을 축소시키며, 이를 통해 독립변수와 종속변수간의 최적의 상관관계를 확인할 수 있다. 다변량 회귀분석에는 다양한 방법이 존재하지만 PLS의 경우 독립변수간의 다중공선성이 존재하더라도 높은 예측력을 갖는 모델을 얻을 수 있는 장점이 있다[12,13].

차원이 축소된 독립변수와 종속변수는 사영된 공간에서 식 (2), (3)와 같이 표현할 수 있다[12].

$$X = TP^T + E = \sum_{a=1}^{l} t_a p_a^T + E \quad (2)$$

$$Y = UQ^T + F = \sum_{a=1}^{l} u_a q_a^T + F \quad (3)$$

본 식에서 P와 Q는 변수 X와 Y의 상관성을 보여주는 loading 벡터이며, T와 U는 score 벡터를 의미한다. 또한 l은 잠재변수(latent variables)의 개수 또는 PLS 차원을, E와 F는 잔차를 의미한다. 두 변수의 상관관계는 각 변수의 score 벡터의 관계를 이용하여 나타내며 이는 식 (4)로 표현할 수 있다[12].

$$U = TB^T + R \quad (4)$$

본 식에서 B는 새로운 공간 상에서 독립변수와 종속변수 사이의 내부 유사성을 나타내는 행렬이며, R은 잔차를 의미한다.

2-3. Correlation coefficient(r)

상관계수(r)은 독립변수와 종속변수 사이의 관계를 판단하는 가장 기본적인 수치이며, 식 (5)과 같이 표현이 가능하다[14].

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (5)$$

위 식에서 X_i, Y_i는 독립변수와 종속변수를 뜻하며, $\overline{X}$, $\overline{Y}$는 각각 독립변수와 종속변수의 평균을 뜻한다. 일반적으로 상관계수의 수치에 따라서 변수간의 관계를 다음과 같이 확인할 수 있다. r의 값이 0.8~1 일 때는 매우 높은 상관관계, 0.6~0.79 일 때는 높은 상관관계, 0.40~0.59 일 때는 보통의 상관관계, 0.20~0.39 일 때는 약한 상관관계, 0~0.19 일 때는 매우 약한 상관관계를 가진다[15].

2-4. Variable importance in projection (VIP)

Variable importance in projections (VIP)는 독립변수 및 종속변수의 구성요소에 대한 loading weight를 반영한 영향을 종합하는 방법으로, 이를 통하여 각 독립변수가 종속변수에 얼마만큼 영향을 미치는지 판단할 수 있다[16]. 통상적으로 유의미한 독립변수를 판별하는데 있어 VIP score가 1.0 이상일 경우를 기준으로 삼지만 0.83~1.21의 범위에서도 유의미한 독립변수를 판별할 수 있다[17]. VIP score는 식 (6)으로 표현이 가능하다[16].

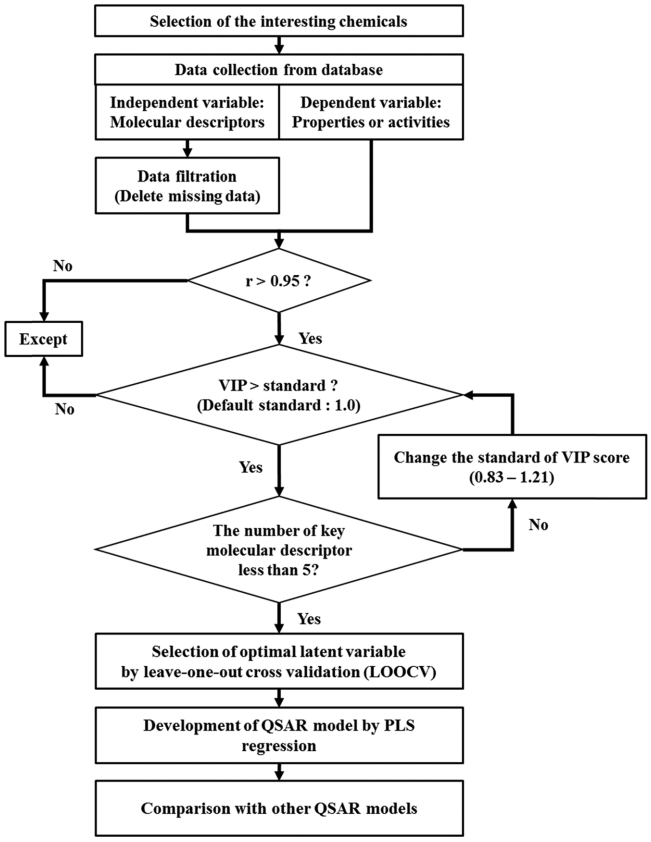

Fig. 2. The flowchart of the proposed QSAR model.

$$VIP_j = \sqrt{\frac{p\sum_{k=1}^{l}(q_k^2 t'_k t_k)(w_{jk}/\|w_{jk}\|)^2)}{\sum_{k=1}^{l}(q_k^2 t'_k t_k)}} \quad (6)$$

위 식에서 p는 분자표현자의 개수이며, w_{jk}는 PLS regression 내에서의 k 번째 구성요소의 j번째 변수에 대한 loading weight를 의미한다. l은 잠재 변수의 개수를 의미하며, $(w_{jk}/\|w_{jk}\|)^2$는 j번째 변수의 중요도를 표현한다.

3. 연구방법

화학물질의 독성치를 예측하기 위한 핵심 분자표현자의 선택방법과 PLS 기반의 QSAR 모델 개발의 과정은 Fig. 2와 같다. 첫 번째 단계로 연구의 대상이 되는 화학물질의 활성치와 분자표현자들을 수집한 후, 활성치와 분자표현자 간에 있어 높은 상관관계를 가지는 분자표현자를 선별하였다. 그 후 활성치와 선별된 분자표현자를 각각 종속변수와 독립변수로 하여 PLS 기반의 QSAR 모델을 개발하였다. 마지막으로, 모델의 잔차와 회귀계수를 이용하여 기존의 QSAR 모델과 제안된 QSAR 모델의 성능을 비교 평가하였다. 본 연구에서의 QSAR 모델 개발 및 분석은 공학용 소프트웨어인 Matlab을 통하여 수행되었다.

3-1. 데이터 수집

연구의 대상이 되는 화학물질에 대한 분자표현자들을 수집하여 이를 독립변수로 지정하고, 관심 물성 및 활성에 대한 정보를 수집

Table 1. The 4885 molecular descriptors from Dragon software

Group name	Numbers	Group name	Numbers
Constitutional indices	43	Ring descriptors	32
Topological indices	75	Walk and path counts	46
Connectivity indices	37	Information indices	48
2D matrix-based descriptors	550	2D autocorrelations	213
Burden eigenvalues	96	P_VSA-like descriptors	45
ETA indices	23	Edge adjacency indices	324
Geometrical descriptors	38	3D matrix-based descriptors	90
3D autocorrelations	80	RDF descriptors	210
3D-MoRES descriptors	224	WHIM descriptors	114
GETAWAY descriptors	273	Randic molecular profiles	41
Functional group counts	154	Atom-Centred fragments	115
Atom-type E-state indices	170	CATS 2D	150
2D Atom Pairs	1596	3D Atom Pairs	36
Charge descriptors	15	Molecular properties	20
Drug-like indices	27	Total	4885

하여 QSAR 모델에 대한 종속변수로 선정하였다. 독립변수에 들어가는 분자표현자는 Dragon software 6[18]를 이용하여 수집하였으며 종속변수에 대한 관심 화학물질에 대한 물성 및 활성정보는 OECD QSAR Toolbox에서 제공하는 정보와 문헌정보를 참고하였다. Table 1은 Dragon software로 수집되는 분자표현자의 그룹과 그 개수를 나타내었다[18]. 분자의 기하학적 구조 및 결합 종류와는 상관없이 가장 기본적인 정보를 나타내는 Constitutional descriptors와 분자의 위상학적 정보에 근거한 Topological indices 등을 포함한 총 29개의 그룹에 해당하는 4885개의 분자표현자에 대한 정보를 수집하였다. 수집된 대상 화학물질의 분자표현자와 물성 및 활성정보는 학습데이터(Training set)와 검증데이터(Test set)로 나뉘어 모델 생성 및 검증에 사용되었다.

3-2. 전처리 및 분자표현자 선택

QSAR 모델을 생성하는 과정에서 활성치를 잘 표현할 수 있는 분자표현자를 선별하는 것은 매우 중요한 과정이다. 또한 분자표현자를 선택하는 과정에서 공선성을 유발할 수 있는 변수들을 선별 및 제거하는 과정은 QSAR 모델 개발에 있어 필수적으로 수행되어야 한다. 변수선별과정은 QSAR 모델의 복잡성을 줄여 예측력 향상 및 계산 시간의 단축을 갖고 온다. 따라서 본 연구에서는 화학물질의 활성도와 높은 상관관계를 갖는 분자표현자를 선별하기 위해서 통계적 선택법을 도입하였다. Dragon software에서 수집된 4885개의 분자표현자들 중 누락된 항목들을 제거한 후, 대상 화학물질의 활성치를 잘 나타낼 수 있는 분자표현자를 선택하기 위해 상관계수(correlation coefficients: r)와 Variable importance in projection (VIP) 방법을 적용하였다. 전체 분자표현자들 중 활성치와 0.95 이상의 매우 높은 상관관계를 갖는 분자표현자들을 우선적으로 선택하였으며, 선택된 분자표현자를 PLS에 적용하여 VIP가 1이상인 분자표현자들을 선택하였다. 모델의 복잡성을 줄이기 위해 최종 선택되는 분자표현자의 수는 5개 이하로 선택하였으며 이때 선택되는 분자표현자가 5개 초과인 경우, 선택된 분자표현자들을 다시 PLS에 적용하여 임의의 VIP 기준(0.83~1.21)에 맞는 분자표현자를 선택하는 과정을 반복하였다(Fig. 2).

3-3. 모델 개발 및 검증

QSAR 모델의 예측도를 높이고 오차를 줄이기 위하여 모델 생성 전후에 있어 다음과 같은 과정을 실시하였다. 우선 학습데이터를 기반으로 PLS 방법을 적용하여 QSAR 모델을 개발하였다. 일반적으로 변수 간의 상관관계를 규정하기 위해서는 다중 선형 회귀 모델을 사용하지만, 이를 QSAR 모델에 적용했을 경우 독립변수간의 강한 상관관계에 따라 발생할 수 있는 다중공선성을 유발할 수 있다[19]. 따라서 본 연구에서는 다중공선성 문제를 피할 수 있고 정확한 예측력을 가진 QSAR 모델을 개발하기 위해 PLS 방법을 적용하였다.

PLS 기반의 QSAR 모델 생성과정에서의 PLS 차원은 leave-one-out 교차검정(leave-one-out cross validation: LOOCV) 방법을 통하여 결정하였다. LOOCV 기법은 학습데이터를 일정 개수의 그룹으로 나눈 후, 하나의 그룹을 검증 대상으로 삼고, 나머지 그룹으로 모델을 생성하는 과정이다. 이 과정을 수행하여 생성된 각각의 모델의 예측력을 Prediction residual error sum of square (PRESS) 값을 이용하여 나타내고, 가장 적은 PRESS값을 나타내는 잠재변수의 개수를 QSAR 모델에서의 PLS 차원으로 설정하였다.

위 과정을 통해 선별된 PLS 차원을 기반으로 PLS 기반의 QSAR 모델을 개발하였으며, 이때 제안된 QSAR 모델의 성능은 회귀계수(R^2)와 PRESS를 이용하여 검정하였다. PRESS는 모든 측정 값과 예측 값 간의 차이의 제곱의 합이며 식 (7)과 같이 표현할 수 있다.

$$\text{PRESS} = \sum_{i=1}^{n}(y_{measured} - y_{predicted})^2 \quad (7)$$

본 식에서 $y_{predicted}$ 은 모델의 예측값, $y_{measured}$은 실제 실험에서 구해진 측정값, n은 데이터의 개수를 의미한다. 제안된 QSAR 모델의 성능평가를 위해 OECD QSAR Toolbox 내의 Trend analysis 기반의 QSAR 모델과 비교·평가 하였다[20].

4. 결 과

4-1. 연구대상물질 선정

본 방법론에 대한 검증을 위해 PCBs (Polychlorinated biphenyl)를

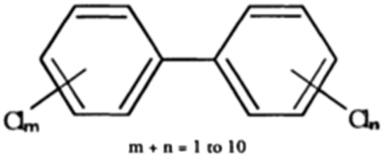

Fig. 3. The molecular structure of polychlorinated biphenyl (PCBs) [20].

연구대상물질로 선정하여 제안된 방법론에 의한 QSAR 모델개발 및 기존의 QSAR 모델과 비교분석을 수행하였다. PCBs는 Fig. 3 [21]과 같은 분자구조를 띄고 있으며, 염소 이온의 치환에 따라 총 209개의 다양한 이성질체가 존재한다. 하지만 다양한 이성질체에 대한 독성치에 대한 정보가 많이 부족한 실정이기에 이에 대한 분석이 필요한 실정이다. PCBs의 여러 활성정보에 있어 본 연구에서 예측하고자 하는 활성치는 화학물질의 일반적인 독성 지표 중 하나인 분배계수(Partition coefficient: P)와 실험군의 50%를 사망시키는 독성 물질의 농도를 뜻하는 반수 치사 농도(lethal concentration 50%: LC_{50})이다. 분배계수의 경우 일반적으로 log 항을 도입하여 log P로 지칭하며, 전체 209개의 PCBs 중 log P 에 대한 정보가 알려진 139개의 PCBs를 본 연구에서의 대상 화학물질로 선택하였다[22]. LC_{50}은 48 시간 기준에서 Branchiopoda의 반수 치사 농도(LC_{50} 48 h)를 관심 활성치로 삼았으며 본 연구에 적용할 때는 –log 항을 추가하여 pLC_{50}로 전환 후 연구에 적용하였다. LC_{50} 48 h는 OECD QSAR Toolbox내에 내장되어 있는 14 개의 PCBs에 대한 LC_{50}를 이용하였다[20]. 또한 데이터를 수집하는 과정에서 두 종속변수인 log P와 LC_{50}을 동시에 보유한 PCBs 데이터는 없었기 때문에 본 연구에서는 log P와 LC_{50}에 대해서 각각 두 개의 PLS 모델을 구축하였다.

4-2. log P 예측을 위한 QSAR 모델

PCBs의 log P를 예측하는 새로운 QSAR 모델을 개발하기 위해 본 연구에서 제시한 방법론을 적용하여 다음과 같이 QSAR 모델 개발을 진행하였다. Dragon software에서 얻을 수 있는 PCBs에 대한 4885개의 분자표현자들 중 계산이 이루어진 2370개의 분자표현자들을 선별한 후, log P와 비교 시 상관계수 0.95 이상의 강한 상관관계를 보이는 135개의 분자표현자들을 선별하였다. Fig. 4(a)는 상관계수를 기반으로 선정한 분자표현자에 VIP를 2회에 걸쳐 적용하여 최종 분자표현자를 선별한 결과를 나타내고 있다. 이 과정에서 135개의 분자표현자들 중 VIP score가 1 이상인 38개의 분자표현자들을 선별하였고, 38개의 분자표현자들 중 VIP score가 1.01 이상인 5개의 분자표현자를 최종 선별하였다.

Table 2은 PCBs의 log P와 가장 높은 상관관계를 갖는 5개의 분자표현자의 이름, group 및 특성을 나타내었다. 선정된 분자표현자들은 H2e, SP01, SP02, SHP2, F06[C-Cl]로 총 5개이며 이 중

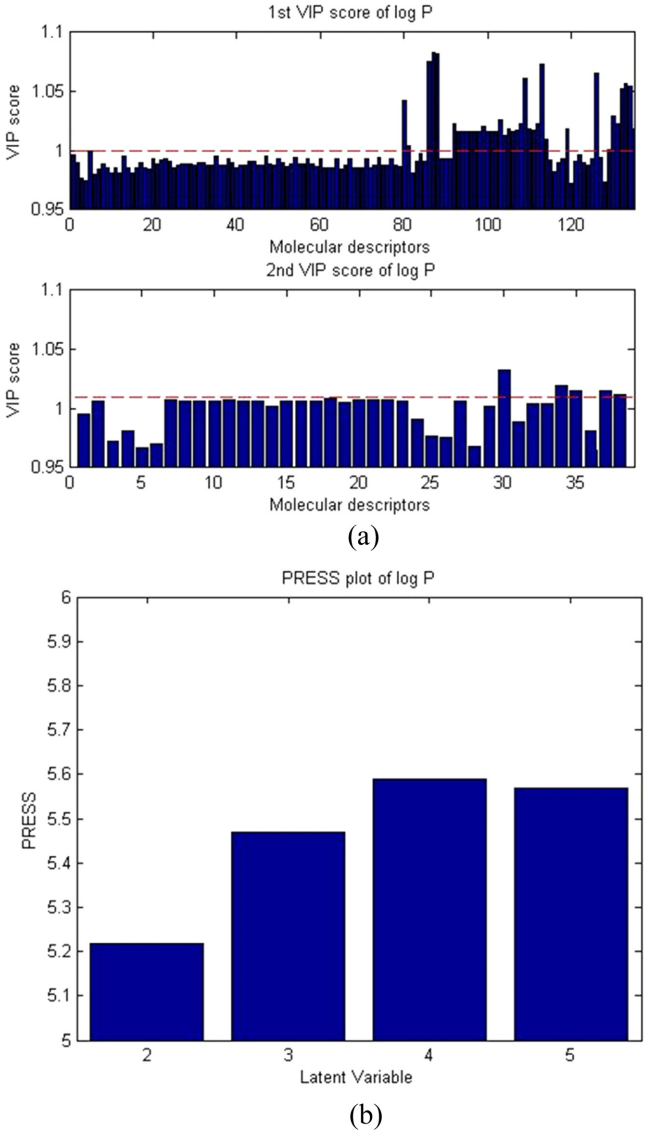

Fig. 4. The preparation of proposed QSAR model for log P; (a) The VIP plots for selecting molecular descriptors which have high correlation with log P; (b) The prediction error sum of square plot for selecting the optimal number of latent variable for log P.

SP01, SP02, SHP2는 Randic molecular profiles에 속하고 H2e와 F06[C-Cl]은 각각 GETAWAY descriptors와 2D Atom Pair group에 속한다. Randic molecular profiles은 원자 사이의 기하학적 거리에 기반하여 분자의 3D 구조 또는 분자 형상(molecular shape)에 따라 계산되는 분자표현자이다[23-25]. GETAWAY descriptors는 Molecular Influence Matrix에 기반하여 화학구조에 대한 특성을 나타낸 분자표현자이다[26-28]. 2D Atom Pairs는 분자를 구성하고 있는 원자와 이

Table 2. The main key molecular descriptors for representation of log P of PCBs

Toxicity	No.	Group	Name	Description
log P	2355	GETAWAY descriptors	H2e	H autocorrelation of lag 2 / weighted by Sanderson electronegativity
	2582	Randic molecular profiles	SP01	Shape profile no. 1
	2583	Randic molecular profiles	SP02	Shape profile no. 2
	2602	Randic molecular profiles	SHP2	Average shape profile index of order 2
	4404	2D Atom Pairs	F06[C-Cl]	Frequency of C - Cl at topological distance 6

를 연결하는 결합의 종류에 따라 정의되는 분자표현자이다[28].

선별된 5개의 분자표현자와 PCBs의 log P를 기반으로 PLS 모델을 생성하기 위한 최적의 PLS 차원을 찾기 위해 leave-one-out 교차검정을 수행하였으며 이에 대한 결과를 Fig. 4(b)에 나타내었다. 이때 PLS 차원의 개수가 2개일 때 PRESS 값이 가장 작았기 때문에 2개의 PLS 차원이 PLS 적용에 있어 최적의 차원개수이다. 교차검정을 통해 선택된 최적의 차원개수과 본 연구에서 제안한 방법을 통해 선택된 분자표현자들을 이용하여 PLS 기반의 QSAR 모델을 개발하였다. 이때 전체 139개의 PCBs의 log P 중 100개의 log P를 학습데이터로 사용하였다.

PLS 기반의 QSAR 모델 생성 후, 분자표현자들과 독성치의 상호관계를 확인할 수 있는 loading plot과 QSAR모델의 성능을 나타내는 Q-Q plot을 Fig. 5에 나타내었다. Fig. 5(a)에서 확인할 수 있는 loading plot은 독립변수(X)와 종속변수(Y) 간의 상관성을 보여 주는 그래프로 두 변수가 근접한 위치에 있을수록 높은 상관성을 갖는

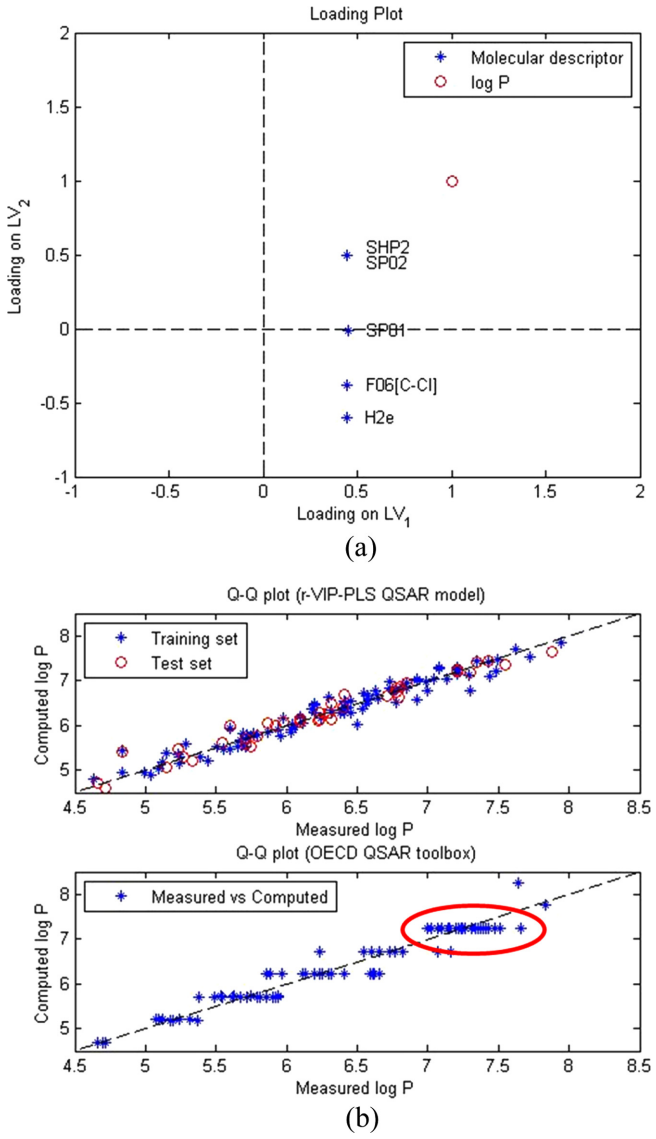

Fig. 5. The result of PLS driven QSAR model and verification for log P; (a) The loading plot of proposed QSAR model for prediction of log P; (b) The Q-Q plots of proposed QSAR model and other QSAR model for log P.

것을 뜻한다. Loading plot의 분석결과, 분자표현자 SHP2와 SP02가 log P와 가까운 위치에 있기에 이 분자표현자들은 log P와 높은 상관관계를 갖는 것을 확인하였다.

제안된 PLS 기반의 QSAR 모델의 성능평가를 위해 기존에 존재하는 QSAR 모델 프로그램 중 하나인 OECD QSAR Toolbox에서 제공하는 Trend analysis와 비교하였다. Trend analysis는 화학물질들의 관심 활성치와 화학물질들의 특성에 대한 경향을 파악한 후, 관심 활성치와 특성에 대한 선형식을 제시하여 타 화학물질의 미확인된 활성치를 예측하는 방법이다. 모델의 Q-Q plot(Fig. 5(b))에서, 제안된 QSAR 모델의 독성치 예측 정도는 기존의 QSAR 모델에 비해 검증 데이터에 있어서도 높은 정확도를 보이고 있는 것을 확인할 수 있다. 이는 QSAR 모델 개발을 위한 분자표현자 선택과정에서 log P와 강한 상관관계를 보이는 분자표현자들이 선택되었기 때문이다. 또한 Fig. 5(b)의 붉은 동그라미로 표시된 것과 같이 기존 QSAR 모델로 예측된 log P는 PCBs의 종류에 관계없이 매우 근사한 값을 나타내는 것을 확인할 수 있다. 이는 기존의 QSAR 모델 개발을 위해 선택된 분자표현자가 관심 활성치를 표현하기에 약한 상관관계를 가지며, 관심 활성치를 예측하기 위한 분자표현자의 개수가 부족하기 때문이다. 따라서 본 연구에서 제시한 분자표현자 선별 방법 및 QSAR 모델 개발법은 기존 방법보다 향상된 예측력을 가졌다고 할 수 있다.

최종적으로 선별된 다섯 개의 분자표현자와 log P, 그리고 선정된 잠재변수의 개수를 기반으로 한 QSAR 모델은 식 (8)과 같은 회귀식으로 도출되었다.

$$\log P = 0.4385 \times (LV_1)_{\log P} + 0.0733 \times (LV_2)_{\log P} + E_{\log P} \quad (8)$$

위 식에서 log P은 전체 PCBs의 log P를 나타내는 벡터이며 $(LV_1)_{\log P}$와 $(LV_2)_{\log P}$은 각각 log P와 관련 있는 첫 번째, 두 번째 잠재변수에 상응하는 값들의 벡터를 의미한다. $F_{\log P}$는 log P에 상응하는 오차에 대한 벡터를 의미한다. 이때 제시된 회귀식에서 $(LV_1)_{\log P}$에 해당하는 계수(0.4385)가 $(LV_2)_{\log P}$에 해당하는 계수(0.0733)보다 크기 때문에 $(LV_1)_{\log P}$가 log P에 미치는 영향이 더 큰 것으로 판단할 수 있다.

4-3. LC_{50} 예측을 위한 QSAR 모델

위의 과정을 거쳐 본 연구에서 제시된 분자표현자 선택법 및 QSAR 모델 개발 방법론이 기존의 QSAR 모델보다 향상된 성능을 보이는 것을 확인하였으며, 이 방법을 바탕으로 PCBs의 여러 독성 기준 중 데이터가 부족한 LC_{50}를 예측하는 새로운 QSAR 모델을 개발하였다. 기존의 방법과 같이 상관계수와 VIP를 적용하여 분자표현자 선택 과정을 수행하였으며 Fig. 6(a)는 VIP 수행 과정을 나타내고 있다. 이때 두 번째 VIP 적용 시 VIP score가 1.05 이상인 분자표현자 5개를 최종 선별하였다.

Table 3은 PCBs의 LC_{50}와 가장 높은 상관관계를 갖는 것으로 확인된 5개의 분자표현자의 이름, group 및 특성을 나타내었다. 선정된 분자표현자들은 VE1_D/Dt, ATSC3m, ATSC3e, SpMax6_Bh(s), RDF025m로 총 5개이며 이 중 ATSC3m, ATSC3e는 2D autocorrelations에 속하고 나머지 분자표현자들은 각각 2D matrix-based descriptor, Burden eigenvalues 그리고 RDF descriptors에 속한다. 2D autocorrelations와 2D matrix-based descriptor 은 위상학적 분자구

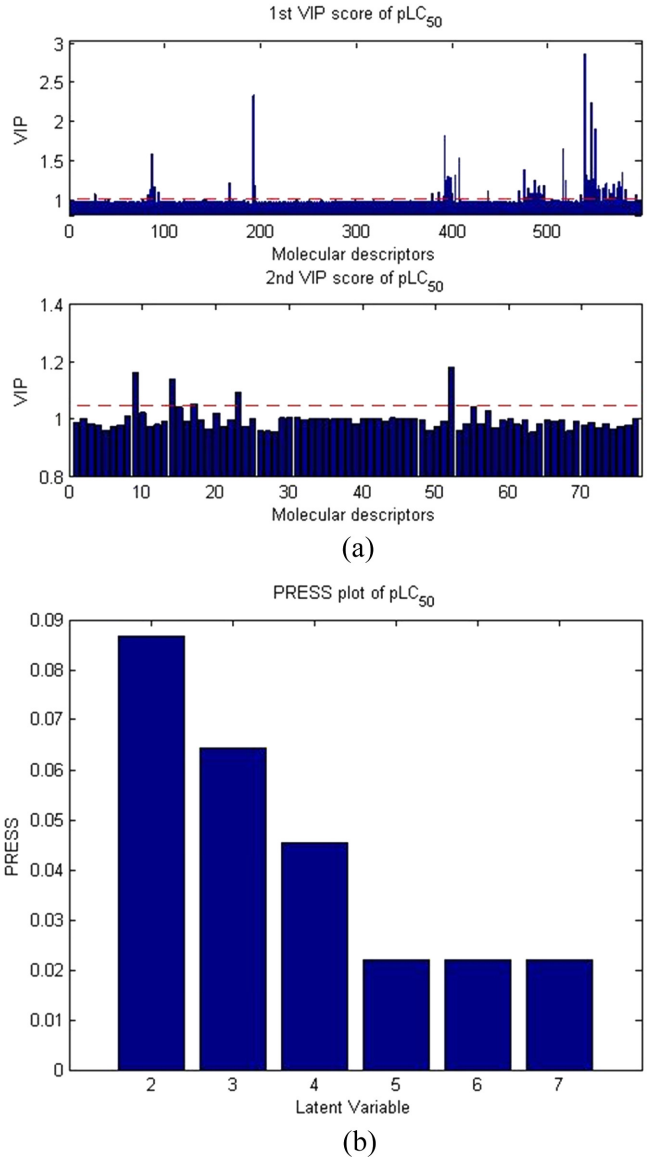

Fig. 6. The preparation of proposed QSAR model for pLC_{50}; (a) The VIP plots for selecting molecular descriptors which have high correlation with pLC_{50}; (b) The prediction error sum of square plot for selecting the optimal number of latent variable for pLC_{50}.

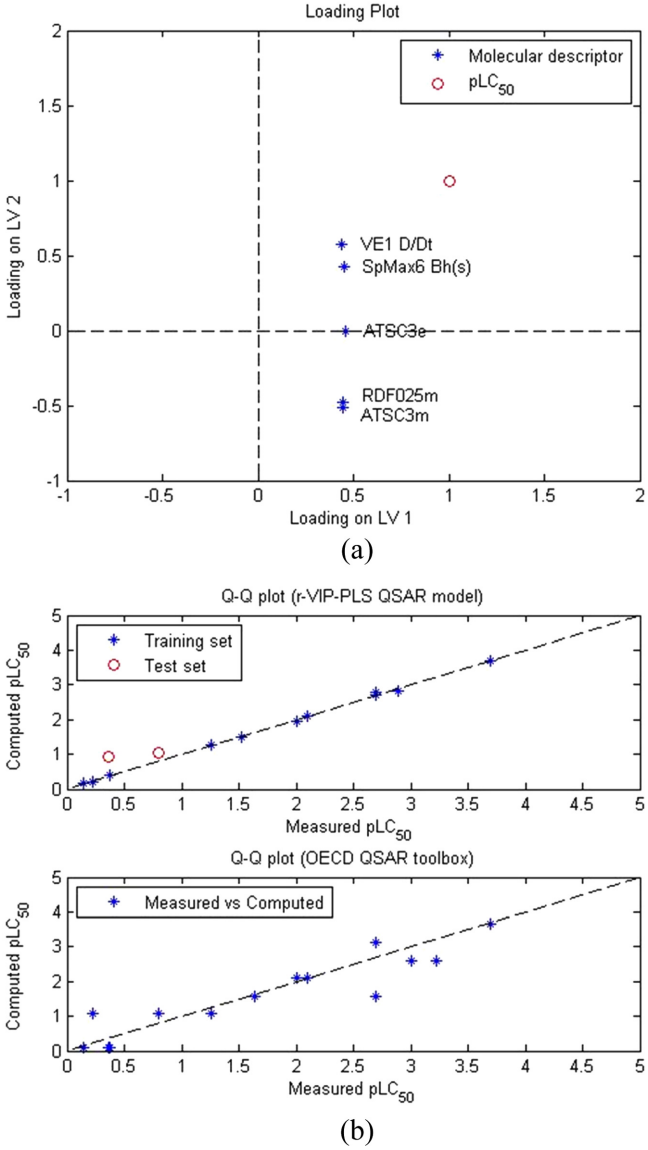

Fig. 7. The result of PLS driven QSAR model and verification for pLC_{50}; (a) The loading plot of proposed QSAR model for prediction of pLC_{50}; (b) The Q-Q plots of proposed QSAR model and other QSAR model for pLC_{50}.

조를 기반으로 하여 계산되는 분자표현자이며, Burden eigenvalues는 각 분자의 hydrogen-filled molecular graph와 Burden matrix에 근거하여 계산되는 분자표현자이다. RDF descriptor는 원자의 구형 부피에서 분자를 찾을 수 있는 확률분포를 설명한 방사분포에 근거한 분자표현자이다[29-32].

Fig. 6(b)는 PCBs의 LC_{50}을 예측할 PLS 모델에 적용될 최적의 PLS 차원을 찾기 위해 leave-one-out 교차검정을 수행한 결과이다. 잠재변수의 개수가 증가할수록 PRESS의 값은 작아지며, 이때 잠재변수의 개수가 5개 이상일 때, PRESS의 값에 대한 변화가 없는 것을 확인할 수 있었다. 따라서 최적의 PLS 차원은 5개로 선정되었으며 선택된 분자표현자를 이용하여 PLS를 수행함으로 LC_{50}을 예측하는 새로운 QSAR 모델을 개발하였다. 이때 전체 14개의 PCBs의 LC_{50} 중 12개의 LC_{50}를 학습데이터로 사용하였다.

Table 3. The main key molecular descriptors for representation of log P of PCBs

Toxicity	No.	Group	Name	Description
pLC_{50}	460	2D matrix-based descriptors	VE1_D/Dt	Coefficient sum of the last eigenvector from distance/detour matrix
	882	2D autocorrelations	ATSC3m	Centred Broto-Moreau autocorrelation of lag 3 weighted by mass
	898	2D autocorrelations	ATSC3e	Centred Broto-Moreau autocorrelation of lag 3 weighted by Sanderson electronegativity
	1090	Burden eigenvalues	SpMax6_Bh(s)	Largest eigenvalue n. 6 of Burden matrix weighted by I-state
	1774	RDF descriptors	RDF025m	Radial Distribution Function - 025 / weighted by mass

Table 4. The number of data, R^2 and PRESS values of proposed QSAR model and OEDC QSAR Toolbox for predicting log P and LC_{50}

		Proposed QSAR model			OECD QSAR Toolbox
		Training set	Test set	Whole set	
log P	The number of data	100	39	139	90
	R^2	0.951	0.959	0.953	0.941
	PRESS	2.750	1.042	3.792	5.150
LC_{50}	The number of data	12	2	14	14
	R^2	0.998	1.000	0.984	0.832
	PRESS	0.029	0.285	0.314	3.680

PLS 기반의 QSAR 모델 개발 후 loading plot과 Q-Q plot에 대해서 Fig. 7에 나타내었다. Fig. 7(a)는 LC_{50}를 예측하기 위해 만들어진 QSAR 모델의 loading plot이며, 이때 분자표현자들 중 VE1_D/Dt와 SpMax6_Bh(s)이 LC_{50}과 높은 상관관계를 갖는 것을 확인하였다. 또한 Fig. 7(b)의 OECD QSAR Toolbox에서 제공하는 QSAR 모델과의 Q-Q plot 비교에 있어서도 제안된 모델이 측정값에 더 잘 수렴하는 것을 확인하였으며, 검증데이터에 대한 예측 또한 측정값과의 차이가 매우 적은 것을 확인하였다.

최종적으로 선별된 다섯 개의 분자표현자와 pLC_{50}, 그리고 선정된 latent variable 개수를 기반으로 한 QSAR 모델은 식 (9)과 같은 회귀식으로 도출되었다.

$$pLC_{50} = 0.4602\times(LV_1)_{pLC_{50}} + 0.1616\times(LV_2)_{pLC_{50}} + 0.0597\times(LV_3)_{pLC_{50}} + 0.2083\times(LV_4)_{pLC_{50}} + 0.1294\times(LV_5)_{pLC_{50}} + F_{pLC_{50}} \quad (9)$$

위 식에서 pLC_{50}은 전체 PCBs의 pLC_{50}를 나타내는 벡터이며$(LV_1)_{pLC_{50}}$부터 $(LV_5)_{pLC_{50}}$는 각 잠재변수에 상응하는 값들의 벡터를 의미한다. $F_{pLC_{50}}$는 pLC_{50}에 상응하는 오차에 대한 벡터를 의미한다. 이때 pLC_{50}에 있어 가장 큰 영향을 미치는 벡터는 계수가 가장 높은 $(LV_1)_{pLC_{50}}$이며 가장 적은 영향을 미치는 벡터는 계수가 가장 낮은 $(LV_3)_{pLC_{50}}$으로 판단된다.

Table 4는 PCB의 log P에 대한 QSAR 모델과 LC_{50}에 대한 모델에 대한 학습데이터 및 검증데이터의 개수와 제안된 QSAR 모델의 성능에 대해서 기존의 OECD QSAR toolbox에서 제공되는 QSAR 모델과의 비교 결과를 R^2와 PRESS를 이용하여 나타내었다. 제안된 PLS 기반의 QSAR 모델은 log P의 경우 전체 데이터를 기준으로 0.953의 R^2 값과 3.792의 PRESS 값을 나타내었다. 이는 OECD QSAR Toolbox에서 측정된 0.941의 R^2값과 5.150의 PRESS 값과 비교 시 제안된 PLS 기반의 QSAR 모델이 기존의 OECD QSAR Toolbox보다 더 좋은 성능을 나타낸다고 할 수 있다. 또한 LC_{50}에 대한 PLS 기반의 QSAR 모델에서는 전체 데이터를 기준으로 0.984의 R^2와 0.314의 PRESS를 나타내었으며, 이는 OECD QSAR Toolbox에서의 0.832의 R^2와 3.680의 PRESS 값과 비교 시 좋은 예측력을 갖는다고 할 수 있다. 이는 기존 모델과 제안된 모델을 비교 시 OECD QSAR Toolbox는 여러 분자표현자들 중에서 단 하나의 특성만을 독립변수로 설정하여 QSAR 모델을 한 반면, 제안된 PLS 기반의 QSAR 모델은 관심 활성치와 강한 활성관계를 갖는 분자표현자들을 QSAR 모델의 독립변수로 설정하였기 때문에 기존의 QSAR 모델보다 더 좋은 예측력을 갖는 것으로 판단된다.

5. 결 론

본 연구에서는 화학물질의 미확인된 물성 및 활성치 예측을 위해 단변량 및 다변량 통계분석을 도입한 분자표현자 선택법과 계산독성학 기반의 QSAR 모델을 제시하였다. 주요 화학물질의 분자표현자들과 활성치에 있어 상관계수 r와 VIP 기법을 적용하여 활성치와 큰 상관관계를 갖는 분자표현자들을 선별하였으며, 선별된 분자표현자들을 이용하여 PLS 기반의 QSAR 모델을 생성하였다. QSAR 모델 생성 시, 최적의 차원개수를 찾기 위해 leave-one-out방법의 교차검증을 사용하였으며, 모델 검정을 위하여 결정계수 R^2과 PRESS을 계산하였다. 제안된 QSAR 모델은 OECD QSAR Toolbox에서 제공하는 QSAR 모델과의 성능 비교를 통해 예측 성능을 검정하였으며, 예측 대상이 되는 화학물질의 활성은 PCBs의 log P와 LC_{50}으로 설정되었다. 본 연구에서 제안된 QSAR 모델은 PCBs의 log P와 LC_{50}를 예측하는 과정에서 기존의 QSAR 모델 보다 각각 26%, 91%의 PRESS를 줄이는 높은 예측력을 나타내었다. 따라서 본 연구에서 제안된 방법은 REACH 제도에 대응하는데 있어 화학물질의 물성 또는 활성 정보에 대한 예측력 향상 및 독성실험에 대한 시간과 비용의 절약이 가능할 것으로 판단되며, 유독 화학물질의 인체 및 환경 위해성 평가 등 다양한 분야에 있어서 적용할 수 있을 것으로 판단된다.

감 사

이 논문은 2015년 정부(미래창조과학부)의 재원으로 한구연구재단의 지원을 받아 수행된 연구로 이에 감사를 드립니다(No. 2015R1A2A2A11001120).

Reference

1. Ahlers, J., Stock, F. and Werschkun, B., "Integrated Testing and Intelligent Assessment - New Challenges Under REACH," *Environ. Sci. Pollut. Res.*, **15**(7), 565-572(2008).
2. Kananpanah, S., Dizadji, N., Abolghasemi, H. and Salamatinia, B., "Developing a New Model to Predict Mass Transfer Coefficient of Salicylic Acid Adsorption onto IRA-93: Experimental and Modeling," *Korean J. Chem. Eng.*, **26**(5), 1208-1212(2009).
3. TGD, E., *Technical Guidance Document (TGD) in support of commission directive 93/67/EEC on risk assessment for new notified substances and commission regulation (EC) No. 1488/94 on risk assessment for existing substances*, Part i to IV, Office for official publications of the European Communities (1996).

4. Devillers, J. and Balaban, A. T. (Ed.), *Topological indices and related descriptors in QSAR and QSPAR*. CRC Press(2000).
5. Song, I. S., Cha, J. Y. and Lee, S. K., "Prediction and Analysis of Acute Fish Toxicity of Pesticides to the Rainbow Trout Using 2D-QSAR," *Anal. Sci. Technol*., **24**(6), 544-555(2011).
6. Ammi, Y., Khaouane, L. and Hanini, S., "Prediction of the Rejection of Organic Compounds (neutral and ionic) by Nanofiltration and Reverse Osmosis Membranes Using Neural Networks," *Korean J. Chem. Eng.*, **32**(11), 2300-2310(2015).
7. Kim, J., Jung, D. H., Rhee, H., Choi, S. H., Sung, M. J. and Choi, W. S., "Aqueous Solubility of Poorly Water-soluble Drugs: Prediction Using Similarity and Quantitative Structure-property Relationship Models," *Korean J. Chem. Eng.*, **25**(4), 865-873 (2008).
8. Coccini, T., Giannoni, L., Karcher, W., Manzo, L. and Roi, R., *Quantitative structure/Activity relationships (QSAR) in Toxicology*. Joint Research Centre, Pavia, Italy (1991).
9. Todeschini, R. and Consonni, V. *Handbook of molecular descriptors*, Vol. 11., John Wiley & Sons (2008).
10. Shi, H., "IAQ monitoring of sub-PCA and health risk assessment of nonlinear QSAR for indoor air pollutants," Master Dissertation, Kyung Hee University, Seoul, Korea (2015).
11. Ock, H. S., "Developing trend of QSAR modeling and pesticides," *Korean J. Pestic. Sci.*, **15**(1), 68-85(2011).
12. Han, I. S. and Shin, H. K., "Modeling of a PEM Fuel Cell Stack Using Partial Least Squares and Artificial Neural Networks," *Korean Chem. Eng. Res.*, **53**(2), 236-242(2015).
13. Lee, C. J., Ko, J. W. and Lee, G. B., "Comparison of Partial Least Squares and Support Vector Machine for the Flash Point Prediction of Organic Compounds," *Korean Chem. Eng. Res.*, **48**(6), 717-724(2010).
14. Montgomery, D. C., Runger, G. C. and Hubele, N. F. *Engineering statistics*. John Wiley & Sons (2009).
15. Pao, S. Y., Lin, W. L. and Hwang, M. J., "In Silico Identification and Comparative Analysis of Differentially Expressed Genes in Human and Mouse Tissues," *BMC genomics*, **7**(1), 1(2006).
16. Mehmood, T., Liland, K. H., Snipen, L. and Sæbø, S., "A Review of Variable Selection Methods in Partial Least Squares Regression," *Chemometr. Intell. Lab.*, **118**, 62-69(2012).
17. Chong, I. G. and Jun, C. H., "Performance of Some Variable Selection Methods when Multicollinearity is Present," *Chemometr. Intell. Lab.*, **78**(1), 103-112(2005).
18. Talete srl, Dragon Version 6.0, http://www.talete.mi.it/.
19. Gholivand, K., Ebrahimi Valmoozi, A. A., Mahzouni, H. R., Ghadimi, S. and Rahimi, R., "Molecular Docking and QSAR Studies: Noncovalent Interaction between Acephate Analogous and the Receptor Site of Human Acetylcholinesterase," *J. Agric. Food Chem*., **61**(28), 6776-6785(2013).
20. OECD QSAR toolbox Version 3.2, http://www.qsartoolbox.org.
21. Robertson, L. W. and Hansen, L. G. (Ed)., *PCBs: recent advances in environmental toxicology and health effects*. University Press of Kentucky(2015).
22. Gramatica, P., Navas, N. and Todeschini, R., "3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs)," *Chemometr. Intell. Lab.*, **40**(1), 53-63 (1998).
23. Randic, M., "Molecular profiles novel geometry-dependent molecular descriptors," *New J. Chem.*, **19**(7), 781-791(1995).
24. Randic, M., "Molecular shape profiles," *J. Chem. Inform. Comput. Sci.*, **35**(3), 373-382(1995).
25. Randic, M. and Razinger, M., "On Characterization of Molecular Shapes," *J. Chem. Inform. Comput. Sci.*, **35**(3), 594-606(1995).
26. Consonni, V., Todeschini, R. and Pavan, M., "Structure/response Correlations and Similarity/diversity Analysis by GETAWAY Descriptors. 1. Theory of the novel 3D Molecular Descriptors," *J. Chem. Inform. Comput. Sci.*, **42**(3), 682-692(2002).
27. Consonni, V., Todeschini, R., Pavan, M. and Gramatica, P., "Structure/response Correlations and Similarity/diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies," *J. Chem. Inform. Comput. Sci.*, **42**(3), 693-705(2002).
28. Carhart, R. E., Smith, D. H. and Venkataraghavan, R., "Atom Pairs as Molecular Features in Structure-activity Studies: Definition and Applications," *J. Chem. Inform. Comput. Sci.*, **25**(2), 64-73(1985).
29. Broto, P., Moreau, G. and Vandycke, C., "Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. Autocorrelation Descriptor," *Eur. J. Med. Chem.*, **19**(1), 66-70(1984).
30. Buckley, F. and Harary, F., *Distance in graphs*. Addison-Wesley Longman(1990).
31. Pearlman, R. S. and Smith, K. M., "Novel Software Tools for Chemical Diversity," *3D QSAR in drug design*, 339-353. Springer Netherlands(2002).
32. Hemmer, M. C., Steinhauer, V. and Gasteiger, J., "Deriving the 3D Structure of Organic Molecules from Their Infrared Spectra," *Vib. Spectrosc.*, **19**(1), 151-164(1999).