EHT Environmental Health and Toxicology

Open Access

Volume: 30 Suppl, Article ID: s2015007, 10 pages
http://dx.doi.org/10.5620/eht.s2015007

• Original Article

eISSN: 2233-6567

Assessment of quantitative structure-activity relationship of toxicity prediction models for Korean chemical substance control legislation

Kwang-Yon Kim[1], Seong Eun Shin[1], Kyoung Tai No[1,2]

[1]Materials Design Team, Bioinformatics and Molecular Design Research Center, Seoul; [2]Computational Systems Biology Laboratory, Department of Bioengineering, Yonsei University, Seoul, Korea

Objectives For successful adoption of legislation controlling registration and assessment of chemical substances, it is important to obtain sufficient toxicological experimental evidence and other related information. It is also essential to obtain a sufficient number of predicted risk and toxicity results. Particularly, methods used in predicting toxicities of chemical substances during acquisition of required data, ultimately become an economic method for future dealings with new substances. Although the need for such methods is gradually increasing, the-required information about reliability and applicability range has not been systematically provided.

Methods There are various representative environmental and human toxicity models based on quantitative structure-activity relationships (QSAR). Here, we secured the 10 representative QSAR-based prediction models and its information that can make predictions about substances that are expected to be regulated. We used models that predict and confirm usability of the information expected to be collected and submitted according to the legislation. After collecting and evaluating each predictive model and relevant data, we prepared methods quantifying the scientific validity and reliability, which are essential conditions for using predictive models.

Results We calculated predicted values for the models. Furthermore, we deduced and compared adequacies of the models using the Alternative non-testing method assessed for Registration, Evaluation, Authorization, and Restriction of Chemicals Substances scoring system, and deduced the applicability domains for each model. Additionally, we calculated and compared inclusion rates of substances expected to be regulated, to confirm the applicability.

Conclusions We evaluated and compared the data, adequacy, and applicability of our selected QSAR-based toxicity prediction models, and included them in a database. Based on this data, we aimed to construct a system that can be used with predicted toxicity results. Furthermore, by presenting the suitability of individual predicted results, we aimed to provide a foundation that could be used in actual assessments and regulations.

Keywords: Applicability, Assessment, Act on the Registration, Evaluation, Authorization, and Restriction of Chemical Substances, Quantitative structure-activity relationships, Validity

Correspondence: Kyoung Tai No
50 Yonsei-ro, Seodaemun-gu,
Seoul 120-749, Korea
Tel: +82-2-393-9550
Fax: +82-2-393-9554
E-mail: ktno@bmdrc.org

Received: January 9, 2015
Accepted: April 22, 2015.
Published online: June 12, 2015

This article is available from: http://e-eht.org/

Introduction

The number of chemical substances currently in global circulation is about 100000, and according to statistics from the Ministry of Environment, approximately 40000 chemical substances are used in the Republic of Korea (hereafter Korea). Moreover, over 90% of the nearly 400 new chemical substances that are introduced each year, are distributed without the necessary infor-

mation about their harmfulness [1]. Accordingly, the human and environmental risks resulting from these substances is increasing daily and, therefore, makes the safety, toxicity, and risk assessment of chemical substances even more essential for the safety and health of the nation. The European Union (EU) Regulation Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), is an example of a recent management and regulations system for chemical substances [2]. This regulation makes registration mandatory for substances exceeding 1 tonne that are manufactured or imported to the EU per year [2]. REACH also requires the registration of chemicals intentionally extracted from completed products as well as the individual chemicals that constitute the mixtures. Therefore, essentially all items exported into the EU are subject to REACH. Recently, the intensification of the so-called 'no data, no market' policy, has rendered conducting industrial activity impossible without material data. Following this trend of environmental regulation, Korea also recently legislated the Act on the Registration and Evaluation, etc. of Chemical Substances (K-REACH) and the implementation is impending [3]. Presently, the data required to register a substance with K-REACH consists of up to 50 categories depending on the amount being manufactured or imported. This requirement includes data for the physicochemical characteristics, human toxicity, and biotoxicity. The greatest problem is the expense of measuring the physical characteristics and toxicity. For example, the EU REACH, which was put into action before the K-REACH, only accepts good laboratory practice (GLP) analysis results when there is no existing data. However, there are only a few designated GLP analysis institutions that can perform the required human and environmental toxicity tests and therefore, the number of categories is also restricted. In terms of the expense, it is reported to cost an average of 50 million Korean won (KRW) and 2130 million KRW to produce data for 1 to 10 and 1000 tonnes of material, respectively [4].

For this reason, alternative non-testing methods have recently received much attention. The non-testing methods (NTM), such as the quantitative structure-activity relationships (QSAR), SAR, and read-across, offer a practical and economic approach that reduces the costs and number of test animals. In addition, these methods are thought to be able to provide the most fundamental toxicity or risk data for previously unestablished or newly developed substances. These methods are even suggested in REACH as a category that must be used in assessments prior to experiments on new vertebrate animals in cases where there is no existing data [5]. However, the following major conditions must first be satisfied: (1) it should be derived from a scientifically validated QSAR model; (2) the substance should be within the applicability domain of the QSAR model; (3) the predicted results

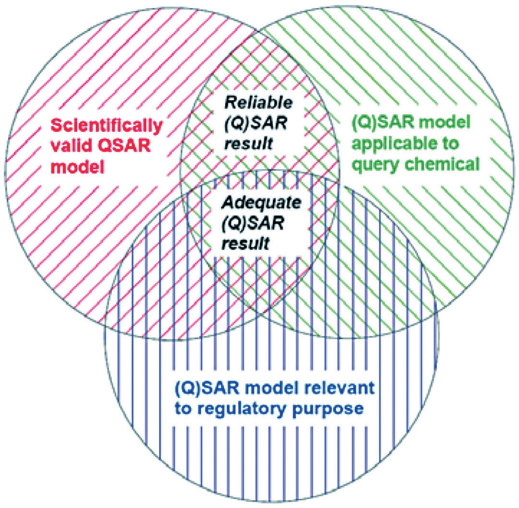

Figure 1. Interrelated concept of quantitative structure-activity relationships (QSAR) validity, reliability, applicability, adequacy, and regulatory relevance.

should be appropriate to the objectives of the classification, labeling, or risk assessment; and (4) adequate and reliable documents should be provided for the relevant techniques [6]. Therefore, to utilize the QSAR and other predictive techniques that are currently in development or being used, for particular groups of substances and objectives such as K-REACH, the criteria listed above must be tested and fulfilled individually to determine the order of priority for selection or application. Specifically, to use NTMs the predictive model should fulfill the three categories of criteria including scientific validity, applicability to query chemicals, and relevance to regulatory purpose as seen in Figure 1 [7].

Of these methods, the predictive QSAR method has received considerable attention because it allows the rapid and easy calculation of the toxicity and properties of a particular substance in large quantities, for the endpoint that requires predicting [8]. Read-across is another predictive method that is a classic data-gap filling technique. This non-testing approach is used to predict the endpoint for a substance that is thought to have endpoint information that is analogous to 1 or several other substances. The analogy may be differentiated based on similarities in physical properties or toxicities, or by groups or categories of substances that show particular patterns. In addition, the categorization process that creates these groups of analogous substances is the decisive factor in the reliability and accuracy of the read-across method. When the substance that is being predicted shows a different analogy, the categorization must be performed differently. As a result, this process is highly influenced by the researcher's experience with the technique and knowledge of the physicochemical and toxicological properties of the target substance groups. Furthermore, predictions for a large number

of diverse substances using this process are known to be difficult. Conversely, in the QSAR method, the various molecular descriptors are calculated from the structure of the individual substances after collecting information about substances that have existing measured data. The method is, therefore, a formulation of the descriptors that have the greatest influence over particular physical or toxicological endpoints as well as coefficients representing that influence, which are derived following 'training' with statistical methods [9]. Currently, the use of nonlinear statistical methods as a way of increasing the prediction accuracy is rapidly increasing. For the QSAR method, after developing the predictive model, only the structure representing the substances to be predicted and the pertinent data may be imputed. This process allows the descriptors used in the model to be calculated and applied to the numerical model. The series of processes are usually automatized in programs that run QSAR models. Therefore, the advantage of predictions based on the QSAR method is that a large number can be performed very easily. However, the downside to this convenience of implementation is that QSAR models tend to be used as a 'black box', which carries the risk that only the predicted results will be obtained without any information about the reliability or future applicability. This differs from the read-across method where an understanding of the target substance, endpoints, and analogous categories occurs naturally in the course of the implementation. This situation can be worse when NTMs are used by non-specialists in response to regulations. Therefore, it is essential to systematically and objectively construct data that shows the performance of QSAR methods, which can predict endpoints of interest in advance. In addition, the constructed data should also confirm the reliability and range of applicable substances, which enables comparison with the predicted results. This is an essential task that must be carried out prior to using NTMs to satisfy K-REACH requirements.

Therefore, the aim of our study is to establish methods to evaluate the suitability of currently used QSAR models for meeting the objectives and applications of K-REACH, which is due to be presented soon. We also intend to apply these methods practically, in the production of a comparative evaluation of the reliability of predictive models. Furthermore, we intend to accumulate these comparative results to provide a foundation for the future use of NTMs.

Materials and Methods

Selection of Usable Predictive Quantitative Structure-activity Relationship Models

In order to establish the QSAR methods to be use in predicting the physical and toxicological data currently required by K-REACH, we first determined the best-known programs and packages worldwide that apply QSAR methods. Our search was focused on the US and Europe where research and the use of QSAR methods are most abundant. To ensure that the results would be suitable for application in future policymaking activities and by the public, we targeted QSAR models used in open software packages. In addition, we included additional paid commercial programs for comparisons of the two types. After identifying the main QSAR-based toxicity prediction programs currently being operated, we found that the majority were used in Europe. We then selected 10 publicly well-known open and commercial QSAR prediction software packages including the estimation program interface (EPI) Suite [10], VEGA [11], TEST [12], ToxPredict [13], Toxtree [14], OCHEM [15], DEMETRA [16], CORALSEA [17], Chembench [18], and the commercial program TOPKAT [19]. We selected one or more QSAR models from each software package for each endpoint that needed to be predicted. Because different software packages can be used to achieve different goals when predicting physicochemical characteristics, we selected a software and QSAR models that would predict endpoints related to environmental and human toxicity. Our selections are displayed in Table 1.

For each predictive model that was selected, we secured as much information as we could from when it was developed, and confirmed the scientific reliability. Essentially, our aim was to collect as much information as possible about each model. The elements listed in the Organization for Economic Cooperation and Development (OECD) principles, served as the basic guideline for the data collection and are as follows:

(1) A defined endpoint.
(2) An unambiguous algorithm.
(3) A defined applicability domain (AD).
(4) Appropriate measure of goodness-of-fit, robustness and predictivity.
(5) A mechanistic interpretation, if possible.

Selection and Prediction of Target Substances

With the cooperation of the affiliated Korea Environment Corporation, we secured a catalogue of 1164 substances that are expected to be prioritized for registration with K-REACH as of 2013, from which we selected the substances that were applicable to the predictive QSAR models. We then deduced the predicted toxicity values for each endpoint by applying them to the previously selected QSAR models. The predicted values were prepared for performance evaluations using the various predictive models later in this study, and for comparison with existing experimental toxicity values, to be collected in the future.

Table 1. Selected QSAR softwares and models for predicting toxicity endpoints defined in K-REACH

	Environmental toxicity									
Software	EPI Suite	VEGA	TEST	ToxPredict	Toxtree	OCHEM	DEMETRA	CORALSEA	Chembench	TOPKAT
Short-term fish testing	2	1	6	2			1	1		1
Short-term *Daphnia* toxicity	1		5				1	1		1
Growth inhibition of algae	1		6			1			1	
Long-term daphnia toxicity	1									
Long-term fish testing	2									
Ready biodegradability	1	1			1					1
Inherent biodegradability	7									
Hydrolysis	1									
Adsorption/Desorption	2									
Bioaccumulation	2	5	6							
Short-term toxicity to earthworms	1									
Short-term toxicity to plants	1									

	Human toxicity									
Software	EPI Suite	VEGA	TEST	ToxPredict	Toxtree	OCHEM	DEMETRA	CORALSEA	Chembench	TOPKAT
In vitro skin irritation/corrosion					1					3
In vitro eye irritation/corrosion					1					3
Skin sensitization		1			1				2	2
In vitro mutagenicity		1	4	1	1	1			2	1
In vitro cytogenicity mammalian cells				1						
In vivo mutagenicity mammalian cells					1					
Inhalation acute toxicity										1
Acute oral toxicity			4					1	1	1
Dermal acute toxicity	1									1
Short-term repeated dose toxicity (28 d)				1						1
Sub-chronic repeated dose study (90 d)										1
Developmental toxicity		1	6							
Carcinogenicity		1		6	1					13

Numbers indicate the number of models in each software.
QSAR, quantitative structure-activity relationships; K-REACH, Act on the Registration, Evaluation, Authorization, and Restriction of Chemical Substances.

Evaluation and Comparison of Robustness and Usability of Predictive Quantitative Structure-activity Relationship Models

Our main interest was to quantify and evaluate the performance of the predictive methods such as QSAR. There are many known methods that can be used to achieve this, which are currently still being studied. Typically, the accuracy, sensitivity, and specificity, which are based on the Cooper statistics, are used to evaluate the performance of a particular predictive model. Taking the most simple form of a binary classification as an example, the different indices were defined as follows:

$$\text{Accuracv A} = (TP+TN)/TOTAL$$

$$\text{Sensitivity S} = TP/(TP+FP)$$

$$\text{Specificity SP} = TN/(TN+FN)$$

Where, TOTAL is the total number of compounds, true positive (TP) is the number of toxic compounds predicted correctly as toxic, true negative (TN) is the number of non-toxic compounds predicted correctly as non-toxic, false positive (FP) is the number of non-toxic compounds predicted incorrectly as toxic, and false negative (FN) is the number of toxic compounds predicted incorrectly as non-toxic. For the quantitative models, we presented the goodness-of-fit of the model using the coefficients of determination (R^2) and validation (Q^2), which are evaluated through the training, test, and external validation sets of the model. There are no statistically clear thresholds and they differ according to the development conditions of the model and field of use, but it is usually accepted that an $R^2 > 0.8$ and $Q^2 > 0.6$ are the required conditions.

Conversely, the evaluation of the QSAR model's robustness and usability remains qualitative. Recently, in the Alternative

Table 2. Criteria used within ANTARES to score QSAR models

Main criteria	Description	Score
Data quality	Relevance: 0, 1 for borderline, 2 for exact Quality: 0 for no-info, 1 for good quality (applied only if relevance >0)	0-3
Chemical number	0: <100, 1: 100-500, 2: 500-5000, 3: >5000	0-3
Descriptors/fragments	0: no info 1: only partial info available 2: possible ambiguities depending on the chemical format 3: full description, equation available	0-3
Explicit and verified the algorithm	0: no info 1: only partial info available 2: possible ambiguities 3: full description, equation available	0-3
Applicability domain	0: no info 1: only partial info available 2: explained, but to be applied manually 3: full description, and model provide tool	0-3
Performance	0: R2<0.65, 1: 0.65≤R2≤0.85, 2: R2>0.85 +1 added if the training set is available	0-3
Validation	0: Q2<0.60, 1: 0.60≤Q2≤0.80, 2: Q2>0.80 +1 added if the external validation set is available	0-3
Output	+1 if univocous +1 if usable as it is +1 if usable as key study or not	0-3
Cost	0: annual license and cost, 2: perpetual license, 3: free	0-3
Additional criteria	**Description**	**Score**
Batch supported	Possible to calculate properties of a set of chemicals	0-2
Structural format	Preference to models where an explicit structure format is defined	0-1
Verify the presence of the uncertainty	Preference to models, that address uncertainty	0-1
Further adequate and reliable documentation	Preference to models, where this is available	0-1
Usability/user friendly	Preference to models, that are easier to be used	0-1
Comprehension	Preference to models, that are easier to comprehend	0-1
Skill requested to interpret results	Preference to models, that are easier to be interpreted	0-1
Access	Preference to models, that have a better access	0-1
Platform/Software requirements	Preference to models, that have less requirements	0-1
Connection problems	Preference to models, that have a less problems in connections	0-1
Time needed	Preference to models, that are faster	0-1

ANTARES, Alternative non-testing methods assessed for REACH substances; REACH, Registration, Evaluation, Authorization and Restriction of Chemicals; QSAR, quantitative structure-activity relationships.

non-testing methods assessed for REACH substances (ANTARES) project of the EU, efforts were made to develop criteria for the quantitative evaluation of the QSAR models, and a self-contained scoring system was presented [20]. The ANTARES scoring system is broadly divided into the main and additional criteria. The resultant scores of each are displayed in Table 2.

As can be seen from Table 2, the main criteria mostly compare areas related to the reliability of the QSAR model's development process from a quantitative perspective while the additional criteria focus mainly on comparing the usability of the relevant QSAR models. A score between 0 to 3 for the 9 main criteria and between 0 to 2 or 0 to 1 for the 11 additional criteria, was awarded based on the selected criteria for each item. The predictive QSAR models were evaluated by adding up the scores for a possible total of 39 points–27 for the main criteria and 12 for the additional criteria. Although there is room for debate about the scores assessed for each item, this is just one method for quantitatively measuring the characteristics of the QSAR models. In addition, we utilized the ANTARES scoring system in our study for comparing the QSAR models.

Comparing Applicability Domains and Quantitative Structure-activity Relationship Models

The AD is an element that has recently become prominent due to its importance in confirming the application validity of the QSAR-based predictive models. The basic premise of the AD is as follows: a QSAR model uses molecules in a restricted training set to produce a statistical quantification of the correlation between structure and activity in those molecules. Therefore, the QSAR model is only suitable for making predictions about the target molecule if that molecule shows a certain level of similarity to the training set. The AD is used as a scale to determine whether this is the case and, therefore, acts as an important element in evaluating the reliability of the results predicted by a QSAR model before assessing its accuracy. In order to determine if there is a similarity, the scales used are structurally similar, and the molecular descriptor is also similar. Therefore, the AD should in principle, be taken into account from the development stage and presented together when calculating predicted values from a QSAR model. However, in reality the existing QSAR models do not sufficiently account for this. Furthermore, the methods and criteria for deducing the AD are different for each QSAR model and so it is impossible to make a relative assessment and comparison of the ADs presented for different QSAR models. As a result, for predictive models we used these lists to redefine the AD ourselves where it was possible to collect the list of molecules in the training set. In addition, we assessed and compared the inclusion of the target substances within the AD using these criteria. We used the k-nearest neighbor (k-NN) method: first, we investigating the degree of similarity between each molecule in the training set and its most similar k neighbors, and then used this information to determine the AD for the training set. Then, for each target molecule, we calculated the degree of similarity with the most similar k molecules from the training set and compared this with the AD [21]. We used the fragment-based Tanimoto structural similarity index [22] to calculate the degree of similarity, which we compared for the 5 most similar molecules.

Results

Quantitative Structure-activity Relationship Model and Target Substances Collections, and Predictions

We collected information about the QSAR models and additional detailed information for each selected endpoint. We used the QSAR model reporting format (QMRF) proposed by the OECD, to collect the data where available. In addition, when this was unavailable we compiled a QMRF ourselves using as much information as possible pertaining to the detailed items within the QMRF. Each collected or compiled QMRF text was included in a database so that it could be referenced when confirming the results predicted by each QSAR model.

Of the 1164 target substances that we selected, we excluded the forms that were not applicable to the general predictive QSAR models. In particular, we excluded substances with implicitly defined names, mixtures of two or more ingredients, and substances with inorganic atoms other than silicon, which current QSAR models cannot predict. As a result, we confirmed 632 substances that were applicable to the QSAR models. In addition, when we applied these molecules to the selected predictive models, we were able to deduce a predicted toxicity value in 87% of the attempts. All the predicted toxicity values that we deduced were included in a database containing basic and structural data based on the substance, endpoint, and predictive QSAR model.

Calculating Alternative Non-testing Methods Assessed for Registration, Evaluation, Authorization, and Restriction of Chemical Substances Scores

We determined the ANTARES scores for each endpoint of each of the selected predictive QSAR models, using the collected data. The results presented in Table 3 list the scores determined for the main areas in the predictive QSAR models for the short-term toxicity in fish and *in vitro* mutagenicity (*in vitro* gene mutation study in bacteria) endpoints. We took the following measures for items requiring detailed qualitative determinations. For the data quality category within the main criteria, we

Table 3. Evaluated ANTARES scores in main criteria for selected QSAR prediction models

Short-term toxicity in fish	EPI Suite	VEGA	TEST	ToxPredict	DEMETRA	TOPKAT
Data quality	3	3	3	3	3	3
Chemical number	2	2	2	2	1	1
Descriptors/fragments	3	3	3	3	3	3
Explicit and verified the algorithm	3	3	3	3	3	1
Applicability domain	3	3	3	2	3	3
Performance	2	2	2	2	2	3
Validation	2	2	2	2	2	2
Output	3	3	3	3	2	3
Cost	3	3	3	3	3	2
Total	24	24	24	23	22	21
In vitro mutagenicity	VEGA	TEST	ToxPredict	Toxtree	TOPKAT	
Data quality	3	3	3	3	3	
Chemical number	3	3	3	2	3	
Descriptors/fragments	3	3	3	3	3	
Explicit and verified the algorithm	3	3	3	3	1	
Applicability domain	3	3	2	0	3	
Performance	2	2	2	2	2	
Validation	2	2	2	2	2	
Output	2	2	2	2	2	
Cost	3	3	3	3	2	
Total	24	24	23	20	21	

ANTARES, Alternative non-testing methods assessed for REACH substances; REACH, Registration, Evaluation, Authorization and Restriction of Chemicals; QSAR, quantitative structure-activity relationships.

Table 4. Evaluated ANTARES scores in additional criteria for QSAR softwares

Name	EPI Suite	VEGA	TEST	Tox Predict	Toxtree	OCHEM	DEME TRA	CORAL SEA	Chem bench	TOP KAT
Batch supported	2	2	1	1	2	1	1	1	1	2
Structural format	1	1	1	1	1	1	1	1	1	1
Verify the presence of the uncertainty	0	1	1	1	0	0	1	0	0	1
Further adequate and reliable documentation	1	1	1	1	1	1	1	1	1	1
Usability/user- friendly	1	1	1	1	1	1	0	1	1	1
Comprehension	1	1	1	1	1	1	0	1	1	1
Skill requested to interpret results	1	1	1	1	1	1	1	1	1	1
Access	1	1	1	1	1	1	0	1	1	1
Platform/software requirements	1	1	1	1	1	1	0	1	1	1
Connection problems	1	1	1	1	1	1	0	1	1	1
Time needed	1	1	1	1	1	1	0	1	1	1
Total (additional criteria)	11	12	11	11	11	10	5	10	10	12

ANTARES, Alternative non-testing methods assessed for REACH; REACH, Registration, Evaluation, Authorization and Restriction of chemicals; QSAR, quantitative structure-activity relationships.

determined whether the data set was significantly relevant to the pertinent endpoint within K-REACH. Then, we awarded a score when the information about the data quality or actions to maintain quality was described. For the output category, when there was enough output of results to allow the compilation of the QPRF we assessed this as usable in the key study. Within the additional criteria, we awarded points for the presence of uncertainty when the output content included techniques to determine the uncertainty of the predicted results. For the 4 categories including the usability/user-friendly, comprehension, skill required to interpret results, and access we assessed the prediction accessibility and difficulty in interpreting results with the assistance of researchers in the field of mandatory experimental toxicology or collaborators who were risk managers. For the areas of the model performance and validation of the categories with a calculated score, we reflected the data presented in each QSAR model for our study. This was because of the limitation requiring the presentation of the actual development results of

Table 5. In-domain ratio of 642 compounds within applicability domain of QSAR models for acute toxicity of *Daphnia magna* and skin sensitization endpoints

Endpoint	Model	Sub-model	Ratio (%)
Acute toxicity of *D. magna*	TOPKAT		92.55
	TEST	TEST-overall	92.45
		TEST-HC	93.75
		TEST-SM	92.51
		TEST-FDA	92.28
		TEST-NN	88.54
	VEGA		92.82
Skin sensitization	TOPKAT		66.99
	Chembench		64.56

QSAR, quantitative structure-activity relationships; HC, hierarchical clustering method; SM, single model method; FDA, Food and Drug Administration method, NN, nearest neighbor method.

the predictive model.

In addition, the ANTARES scores for the additional criteria were mostly used to compare the usability of the QSAR models. Therefore, rather than a scientific assessment of the QSAR model itself, these can be said to reflect the features of the platform or software that is running the QSAR model. As a result, in our study, the comparative evaluation of the additional ANTARES scoring criteria was calculated for each QSAR program irrespective of the individual endpoints, and those results are displayed in Table 4.

Comparing Applicability Domain Between Individual Predictive Quantitative Stucture-activity Relationship Models

To assess the AD for each predictive QSAR model, we first secured the lists of substances in the training and test sets used at development. Then, using the method described above for the predictive models where these lists were available, we calculated the AD based on the k-NN for each endpoint and each model. After calculating the structural similarity based on the k-NN of the 642 target substances with the substances in the training and test sets for each model, we determined whether the calculated similarity index was located within the AD. We calculated the ratio of the substances located within the domain to the total number of substances and compared this for each endpoint and each model. For the TOPKAT predictive model, we were unable to secure the training and test set, but we determined the AD inclusion using the optimum prediction space based on the principal component analysis, which is provided with the program [23]. As an example of representative endpoints, Table 5 summarizes the calculated and compared AD inclusion ratios for the acute toxicity of *Daphnia magna* and skin sensitization in the environmental and human toxicity fields, respectively.

Discussion

Comparing Calculated Alternative Non-testing Methods Assessed for Registration, Evaluation, Authorization, and Restriction of Chemical Substances Scores

Comparing the ANTARES scores in Table 3 calculated for the main criteria, no particular QSAR model showed superiority. In addition, the predictive models for the short-term toxicity in fish and in vitro mutagenicity both showed results within the range of 21 to 24 out of a maximum of 27 points. The EPI Suite was one of the 3 predictive models awarded the highest scores. This is particularly noteworthy because the high score was awarded for robustness even while enabling the predictions of endpoints across the whole field of environmental toxicity. In addition, there were a limited number of endpoints that could be predicted by the VEGA and TEST. However, predictions could be made for the major endpoints in the environmental and human toxicity fields, and they showed the highest model reliability grades together with the EPI Suite. Conversely, when we investigated the elements in the predictive models that received relatively low assessments, for short-term toxicity on fish, DEMETRA showed a low score. This low score was obtained because the number of chemicals included during model development was less than 5000. In addition, the TOPKAT also showed a low score because of the small number of chemicals used in the development and the algorithm used in the predictive model was not clearly revealed. For the *in vitro* mutagenicity, both Toxtree and TOPKAT received low scores because there was no data presented for the AD. In particular, TOPKAT, the only commercial program used in our study for comparison was awarded a poor assessment in the clarity of algorithm category for both endpoints. However, since this pattern might be similar for other commercial programs, this needs to be considered in the future use of the predictive models.

The additional criteria in Table 4 shows the scores for the other calculated index, which revealed that the majority of the models received a score in the range of 10 to 12 and VEGA and TOPKAT received satisfactory scores in every category. This may be because both VEGA, a java-based web or independent program and TOPKAT, a module within a different modeling package showed an overall high level of completeness from the interface to the speed and the output of results. In contrast, we found that the other prediction programs received a relatively poor assessment in certain areas due to a lack of support for a particular operation or failure to confirm the existence of uncertainty measurements. In particular, DEMETRA was awarded the lowest score and received a poor assessment in the 7 categories pertaining to ease of use. This result suggests that its accessibility and applicability might be relatively limited for general users.

Comparing Applicability Domains of Predictive Quantitative Structure-activity Relationship models

The AD inclusion results for the 642 target substances for the acute toxicity of *D. magna* endpoint TOPKAT are illustrated in Table 5. This result included the TEST and VEGA for comparison, and they all showed a within-domain inclusion rate of approximately 92%. In addition, although there are no clear standards, this demonstrates the overall applicability of these 3 predictive QSAR methods for predicting the toxicity of the targeted endpoint in the majority of substances currently expected to be regulated. Furthermore, TEST is the only detailed model developed by the nearest neighbor method, and it showed an inclusion rate below 90%. In contrast, for the skin sensitization endpoint, TOPKAT and Chembench both showed inclusion rates in the region of 65%. This rate implies that if these models were used later about 1 in 3 molecules would fall outside of the prediction reliability range of the QSAR model. As a result, when these models are applied to diverse substances, caution should be exercised in analyzing the results. In addition, additional testing of the prediction results may be required when deemed necessary using comparative analysis with substances within the training set that have structural similarity.

The results of the reliability analysis above obtained by setting the AD and evaluating the inclusion for each QSAR model can be combined with the previously mentioned ANTARES scoring results. This combination allows these analyses to be applied in determining the priorities and usability of the results calculated by each model. Conversely, many tasks still need to be included. The scientific validation of each model is reflected by the prediction results and the ANTARES scores, and efforts are continually required to update relevant information and the predicted results. These updates are necessary to reflect the changes to these QSAR models, which are constantly being upgraded and renewed. Furthermore, the method of comparing ADs is inevitably influenced not only by changes to the predictive model, but also by the types of substances to which it is applied. As a result, there is a need to change and update the AD continually, based on changes in the predictive model and the range of substances targeted for regulation. In addition, a suitable plan must be instituted to integrate and compare all the performance indices of the predictive models with different scales such as robustness, ease of use, and AD inclusion rate. In addition, the accuracy, sensitivity, and specificity need to be incorporated, which we are planning to do in the future. Efforts must be made to construct a system that allows the integrated management and search of all the determined information within a single database. Finally, it is noteworthy that it is our intention to provide a foundation that can be currently used in the development of assessment and regulation policies for the target substances in this study.

Acknowledgements

This study was supported by Korea Ministry of Environment (MOE) as "MOE's R&D Program on environmental technology development" (no. 412-111-009).

Conflict of Interest

The authors have no conflicts of interest with material presented in this paper.

References

1. Minstry of Environment. Current status and vision of the management of chemical substances in Korea: presentation report of National Institute of Environmental Research (NIER), 2011 [cited 2015 Jan 5]. Avaliable from: http://www.reach.me.go.kr/board/boardDownload.asp?id=2907&fn=1 (Korean).
2. European Chemicals Agency. Guidance on registration; 2012 [cited 2014 Apr 28]. Available from: http://echa.europa.eu/documents/10162/13632/registration_en.pdf.
3. Ministry of Government Legislation. Act on the Registration and Evaluation, etc. of Chemical Substances [cited 2014 Dec 30]. Available from: http://www.law.go.kr/lsInfoP.do?lsiSeq=140402#0000 (Korean).
4. Korea Development Institute. Anaysis report: the effect of domestic industry according to EU-REACH legislation; 2012 [cited 2014 Dec 15]. Available from: http://www.prism.go.kr/homepage/researchCommon/retrieveResearchDetailPopup.do;jsessionid=53B47AF8E1BB6B007845C668E9DAF279.node02?research_id=1480000-200700151 (Korean).

5. Höfer T, Gerner I, Gundert-Remy U, Liebsch M, Schulte A, Spielmann H, et al. Animal testing and alternative approaches for the human health risk assessment under the proposed new European chemicals regulation. Arch Toxicol 2004;78(10):549-564.
6. Simon-Hettich B, Rothfuss A, Steger-Hartmann T. Use of computer-assisted prediction of toxic effects of chemical substances. Toxicology 2006;224(1-2):156-162.
7. ORCHESTRA. Theory, guidance and application on QSAR and REACH; 2012 [cited 2014 Dec 30]. Available from: http://home.deib.polimi.it/gini/papers/orchestra.pdf.
8. Mays C, Benfenati E, Pardoe S. Use and perceived benefits and barriers of QSAR models for REACH: findings from a questionnaire to stakeholders. Chem Cent J 2012;6(1):159.
9. Cronin MT. Prediction of harmful human health effects of chemicals from structure. In: Puzyn T, Leszczynski J, Cronin MT, editors. Recent advances in QSAR studies: methods and applications. New York: Springer; 2010, p. 305-325.
10. United States Environmental Protection Agency. Estimation Program Interface (EPI) Suite. version 4.10 [cited 2014 Dec 15]. Available from: http://www.epa.gov/oppt/exposure/pubs/episuite.htm.
11. Istituto di Ricerche Farmacologiche Mario Negri Milano. VEGA (based on CAESAR project). version 1.0.8 [cited 2014 Aug 25]. Available from: http://www.vega-qsar.eu.
12. United States Environmental Protection Agency. Quantitative structure activity relationship [cited 2014 Aug 25]. Available from: http://www.epa.gov/nrmrl/std/qsar/qsar.html.
13. OpenTox. ToxPredict [cited 2014 Aug 25]. Available from: http://apps.ideaconsult.net:8080/ToxPredict.
14. Joint Research Centre, European Union. Toxtree. version 2.6.0 [cited 2014 Aug 25]. Available from: https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/qsar_tools/toxtree.
15. Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, et al. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J Comput Aided Mol Des 2011;25(6):533-554.
16. Benfenati E, Craciun M, Neagu D. The use of the DEMETRA models. In: Benfenati E, editor. Quantitative structure-activity relationship (QSAR) for pesticide regulatory purposes. Amsterdam: Elsevier; 2007, p. 303-313.
17. Toropova AP, Toropov AA, Benfenati E, Gini G, Leszczynska D, et al. CORAL: quantitative models for estimating bioconcentration factor of organic compounds. Chemometr Intell Lab Syst 2012;118:70-73.
18. Carolina Exploratory Center for Cheminformatics Research. CHEMBENCH. version 1.2.0 [cited 2014 Aug 25]. Available from: http://chembench.mml.unc.edu/modeling.
19. Dassault Systemes. QSAR, admet and predictive toxicology with biovia discovery studio [cited 2014 Dec 8]. Available from: http://accelrys.com/products/datasheets/qsar-admet-and-predictive-toxicology-with-ds.pdf.
20. ANTARES. Deliverable 2: report on the identified criteria for non-testing methods, and their scores; 2010 [cited 2014 Dec 18]. Available from: http://www.antares-life.eu/files/ANTARES_D2.pdf.
21. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 2003;22(1):69-77.
22. Godden JW, Xue L, Bajorath J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. J Chem Inf Comput Sci 2000;40(1):163-166.
23. Gombar VK, Enslein K. Assessment of n-octanol/water partition coefficient: when is the assessment reliable? J Chem Inf Comput Sci 1996;36(6):1127-1134.