SAR and QSAR in Environmental Research
Vol. 21, Nos. 5–6, July–September 2010, 403–413

Taylor & Francis
Taylor & Francis Group

Development of an ecotoxicity QSAR model for the KAshinhou Tool for Ecotoxicity (KATE) system, March 2009 version

A. Furuhama[a]*, T. Toida[a], N. Nishikawa[a], Y. Aoki[a], Y. Yoshioka[b] and H. Shiraishi[a]

[a]*Research Center for Environmental Risk, National Institute for Environmental Studies (NIES), 16–2 Onogawa, Tsukuba 305–8506, Japan;* [b]*Faculty of Education and Welfare Science, Oita University, 700 Dannoharu, Oita 870–1192, Japan*

(*Received 18 February 2010; in final form 28 April 2010*)

The KAshinhou Tool for Ecotoxicity (KATE) system, including ecotoxicity quantitative structure–activity relationship (QSAR) models, was developed by the Japanese National Institute for Environmental Studies (NIES) using the database of aquatic toxicity results gathered by the Japanese Ministry of the Environment and the US EPA fathead minnow database. In this system chemicals can be entered according to their one-dimensional structures and classified by substructure. The QSAR equations for predicting the toxicity of a chemical compound assume a linear correlation between its log *P* value and its aquatic toxicity. KATE uses a structural domain called *C*-judgement, defined by the substructures of specified functional groups in the QSAR models. Internal validation by the leave-one-out method confirms that the QSAR equations, with $r^2 > 0.7$, RMSE ≤ 0.5, and $n > 5$, give acceptable q^2 values. Such external validation indicates that a group of chemicals with an in-domain of KATE *C*-judgements exhibits a lower root mean square error (RMSE). These findings demonstrate that the KATE system has the potential to enable chemicals to be categorised as potential hazards.

Keywords: QSAR; ecotoxicity prediction; classification; chemical substances; domain; KATE

1. Introduction

Quantitative structure–activity relationships (QSARs) are potential tools for predicting the activity and properties of chemicals, including their physicochemical attributes, health effects, ecotoxicity and biological activity. QSAR models can estimate and predict such activity and can thus be used to categorise chemicals in terms of their potentially hazardous nature. A recent review has demonstrated that acute aquatic toxicity [1] can be predicted using QSAR and describes the available databases of ecotoxicity data.

Prediction of toxicity by QSAR does not require lengthy experiments, nor the use of animals, plants or cells. QSAR models have therefore been utilised for the assessment of new and existing chemicals for conformity with regulatory requirements in countries within the Organisation for Economic Co-operation and Development (OECD) [2]. In Japan, under the Chemical Substances Control Law (CSCL), the Ministry of the Environment (MoE) is responsible for evaluating the adverse effects of chemicals on

*Corresponding author. Email: ayako.furuhama@nies.go.jp

ISSN 1062–936X print/ISSN 1029–046X online

DOI: 10.1080/1062936X.2010.501815
http://www.informaworld.com

ecosystems, and uses tests involving aquatic organisms such as *Oryzias latipes* (fishes) or *Daphnia magna* (daphnia), in addition to algae data available from the MoE website [3]. The Japanese National Institute for Environmental Studies (NIES) was established to apply QSAR models to acute ecotoxicity, and has developed a QSAR prediction system using the MoE ecotoxicity database. This system, published in March 2009, is known as the *KAshinhou Tool for Ecotoxicity* (KATE) [4].

The present paper focuses on the theoretical and methodological aspects of the KATE system, and QSAR equations classified by chemical substructure are introduced. We shall then present the cross-validation ('leave-one-out') results, and the toxicities calculated by KATE, and by alternative systems such as TIssue MEtabolism Simulator (TIMES) [5,6] (developed by Zlatarov at the Laboratory of Mathematical Chemistry, Bourgas University, Bulgaria), and by ECOSAR™ [7] (developed by the US Environmental Protection Agency (EPA)) using the same end-point data set as that in KATE. The validity of KATE will be discussed using the applicability domain, log *P*, and *C*-judgements.

2. Overview of KATE

2.1 *End-point*

KATE uses experimental data on chemical substances to predict aquatic toxicity. The end-points of interest are the 96-hour median lethal concentration (LC_{50}) in fish after acute toxicity tests, and the 48-hour median effective concentration (EC_{50}) in daphnia obtained after acute immobilisation tests. Training sets for QSAR development were derived from the results of ecotoxicity tests (*Oryzias latipes* LC_{50} and *Daphnia magna* EC_{50}) obtained by the MoE [3], as well as the results of acute toxicity tests from the US EPA fathead minnow (*Pimephales promelas*) database [8,9]. In the KATE system, the 96-hour LC_{50} data for *Oryzias latipes* and fathead minnow were combined to reinforce the number of reference datasets. The QSAR equations in KATE for the fish and daphnia end-points were designed using 535 and 258 chemicals, respectively.

2.2 *Classification of chemicals*

Chemical substances can be classified according to the substructures that give rise to specific chemical properties (Appendix 1 of the supplementary material which is available on the Supplementary Content tab of the article's online page at http://dx.doi.org/10.1080/1062936X.2010.501815). The rules for daphnia and fish end-points are identical, except for the following five classes: *amines aromatic or phenols1*, *amines aromatic or phenols3*, *amines aromatic or phenols4*, *amines aromatic or phenols5*, and *primary amines*. According to KATE, the toxicity of a chemical containing amino functional groups might be different in daphnia from its toxic behaviour in fish.

Forty-four classes are proposed for each end-point of KATE QSAR models. Table 1 shows the QSAR class name, and the detailed class features are listed in Appendix 2 of the supplementary material (available online). The chemicals in the KATE *unclassified* class were not categorised within any of the rules in Appendix 2. Additional classification rules or fragment definitions are required in further studies to reduce the number of chemicals described as *unclassified*. It should be noted that the concept of *unclassified* within KATE does not always include reactive chemicals, and thus differs from the *reactive unspecified* category in the TIMES software.

Table 1. QSARs for fish acute toxicity estimated by the equation: $\log(1/LC_{50}[\mathrm{mM}]) = a * \log \mathrm{P} + b$.

Class name	*a, b*	*n*	*RMSE*	r^2, q^2	*log P range*	*[1]
Hydrocarbons aromatic	0.630, −0.883	43	0.368	0.826, 0.803	[0.60, 5.17]	
Dinitrobenzenes	0.568, 0.551	12	0.669	0.331, 0.170	[0.56, 3.60]	
Nitrobenzenes	0.678, −0.693	9	0.300	0.875, 0.760	[0.82, 5.10]	
Amines aromatic or phenols1	−0.005, 2.671	7	0.354	0.001, 0.887	[−0.30, 4.47]	
Amines aromatic or phenols2	0.012, 1.863	7	0.307	0.003, 0.737	[3.67, 8.47]	C
Amines aromatic or phenols3	0.214, 0.945	16	0.305	0.272, 0.106	[0.15, 3.68]	
Amines aromatic or phenols4	0.725, −0.779	56	0.321	0.900, 0.890	[0.51, 7.54]	
Amines aromatic or phenols5	0.544, −0.612	22	0.324	0.661, 0.600	[0.35, 3.50]	
Primary amines	0.529, −0.622	23	0.406	0.803, 0.741	[−2.04, 3.60]	
Secondary and tertiary amines	0.592, −0.595	10	0.512	0.731, 0.605	[−1.43, 2.79]	C
Hydrazines	0.417, 1.832	4	0.413	0.884, 0.639	[−1.68, 4.70]	
Amides and imides	0.746, −1.026	17	0.601	0.696, 0.607	[−0.48, 3.80]	
Esters aliphatic	0.638, −0.600	13	0.393	0.722, 0.651	[0.18, 3.65]	
Esters aromatic	0.513, −0.157	9	0.253	0.856, 0.790	[1.94, 5.53]	
Aldehydes	0.484, 0.279	15	0.557	0.272, 0.111	[−0.34, 2.47]	
Acids	0.728, −1.652	9	0.355	0.816, 0.667	[0.33, 4.20]	
Acids acrylic	0.122, 0.045	3	0.039	0.607, 0.271	[0.35, 1.33]	
Conjugated systems1	0.753, 2.084	4	1.012	0.463, 0.111	[−1.11, 2.20]	
Conjugated systems2	0.436, 0.901	17	1.007	0.264, 0.066	[−0.38, 4.10]	
Thiols aromatic	NO-QSAR					
Thiols aliphatic	0.371, 0.732	4	0.291	0.910, 0.633	[−0.17, 6.12]	C
Sulfides	0.753, −1.336	8	0.259	0.699, 0.573	[2.46, 4.16]	C
Disulfides	0.386, 0.845	6	0.666	0.210, 0.012	[1.74, 4.44]	C
Carbamates	0.004, 1.894	11	0.519	0.000, 0.645	[−0.47, 4.60]	
Pyrethroids	NO-QSAR					
Acrylates	0.158, 1.498	6	0.155	0.474, 0.022	[−0.21, 2.36]	
Methacrylates	0.465, −0.031	6	0.417	0.657, 0.293	[0.47, 4.54]	
Epoxides	0.323, 1.055	4	0.272	0.755, 0.283	[0.08, 3.98]	C
Barbitals or thiols other	1.583, −2.560	4	0.291	0.657, 0.927	[1.47, 2.10]	
Esters phosphate	0.691, −0.111	11	0.856	0.389, 0.175	[2.23, 5.33]	
N or P cations	0.274, 0.956	9	0.579	0.791, 0.628	[−8.36, 6.69]	C
Halides1	0.254, 1.325	6	0.971	0.112, 0.078	[0.45, 4.50]	
Halides2	0.824, −0.318	8	0.560	0.879, 0.810	[−0.06, 5.04]	
Halides3	0.783, −1.291	42	0.263	0.879, 0.868	[1.25, 4.89]	
Metals	NO-QSAR					
Nitriles aliphatic	0.839, −1.154	6	0.254	0.938, 0.901	[−0.34, 3.12]	N
Ketones	0.864, −1.602	21	0.345	0.891, 0.867	[−0.24, 4.09]	N
Alcohols or ethers aliphatic	0.853, −1.958	23	0.321	0.950, 0.924	[−0.77, 5.82]	N
Phosphates	0.891, −1.926	3	0.257	0.865, 0.485	[2.83, 4.59]	N
Hydrocarbons aliphatic	0.753, −1.286	15	0.289	0.824, 0.785	[2.42, 5.56]	N
Ethers aliphatic	0.749, −1.806	11	0.190	0.972, 0.962	[−0.54, 4.25]	N
Ethers aromatic	0.870, −1.466	10	0.233	0.922, 0.892	[1.16, 4.21]	N
Neutral organics	0.842, −1.674	88	0.384	0.924, 0.919	[−0.77, 5.82]	
Unclassified	0.744, −0.898	25	0.714	0.712, 0.660	[−1.35, 5.50]	

*[1] C: an equation is generated by calculated Clog *P*. N: a member of the *Neutral organics* class.
Note: *n*, RMSE, r^2 and q^2 denote the number of chemicals in a class, the root mean square error, the squared correlation coefficient, and the leave-one-out version of the squared correlation coefficient, respectively. The log *P* range shows minimum and maximum log *P* values.

2.3 *Neutral organics*

Neutral organics is an aggregate of the chemicals in defined classes in the KATE system. It comprises the classes: *nitriles aliphatic*, *ketones*, *alcohols or ethers aliphatic*, *phosphates*, *hydrocarbons aliphatic*, *ethers aliphatic* and *ethers aromatic*. In the OECD Environment Monograph [10], *neutral organic* compounds of minimal toxicity were divided into the groups: *aliphatic alcohols*, *aliphatic ketones*, *aliphatic ethers* and *alkoxyethers*, *aliphatic halogenated hydrocarbons*, *saturated alkanes* and *halogenated benzenes*. Some of the *neutral organics* compounds defined in the OECD monograph were categorised differently from those in KATE.

2.4 *QSAR equations*

The QSAR equations in the KATE model express the correlation between the octanol/water partition coefficient (log P) of a compound and its aquatic toxicity, using simple linear regression analysis. Measured log P values were used to derive the QSAR equations, except for the equations labelled C in Tables 1 and 2. In cases where experimental log P values were not available, an equation was constructed from the calculated Clog P value obtained by the Daylight toolkit [11]. The LC_{50} and EC_{50} values in the equation were expressed in terms of the common logarithm of the inverse of millimoles per litre (mmol L^{-1}, or mM). The equations and the statistical information obtained are shown in Tables 1 and 2. Where there were fewer than three sets of reference data within one class, QSAR prediction could not be performed. In such cases the class name was the only information obtained from KATE, and the label NO-QSAR is indicated in Tables 1 and 2. The equation for a class named *pyrethroids* was not constructed, since the log P values in the reference data were gathered in higher ranges [6.1, 6.5].

2.5 *Domains in KATE*

KATE offers two 'judgements' to verify whether or not a predicted chemical substance falls within the applicability domain of a QSAR class. The first is the log P judgement, based on the log P range defined by the reference chemical data of the class concerned. This has been categorised as a descriptor domain [12,13]. The interpolated log P range for each class is listed in Tables 1 and 2.

The second is the C-judgement, which is categorised as a structural domain and is defined by the substructures shown in Appendix 3 of the supplementary material (available online). The substructures are based on functional groups having similar concepts to those used by Schultz et al. [13], rather than on atom-centred fragments [12,14]. Schultz et al. applied the structural domain to one QSAR equation for aromatic compounds, and the out-of-domain revealed well-known electrophoric mechanisms in the structural space(s) [13]. In the KATE system the classification rules (described in Section 2.2) play a role in constructing such structural space(s). The definition of the applicability domain of C-judgement depends on whether all the substructures of the chemical under test are found in reference chemicals in the class, or secondly, whether all substructures in the test chemical are present in reference chemicals in either *neutral organics* or the class concerned. The first of these definitions is stricter than the second. The reliability of the log P and C-judgements is assessed later in Section 4 (Results and discussion).

Table 2. QSARs for the daphnia acute toxicity estimated by the equation: $\log(1/EC_{50}[\mathrm{mM}]) = a * \log P + b$.

Class name	*a, b*	*n*	*RMSE*	r^2, q^2	*log P range*	*[1]
Hydrocarbons aromatic	0.607, −0.414	26	0.351	0.808, 0.762	[0.65, 5.17]	
Dinitrobenzenes	0.408, 0.632	5	0.561	0.343, 0.090	[0.56, 3.60]	
Nitrobenzenes	0.547, −0.164	4	0.238	0.915, 0.675	[1.17, 5.10]	
Amines aromatic or phenols1	0.085, 2.441	7	0.443	0.057, 0.375	[−0.33, 3.41]	
Amines aromatic or phenols2	0.097, 1.152	6	0.277	0.239, 0.031	[3.67, 8.47]	C
Amines aromatic or phenols3	0.132, 1.748	16	0.406	0.119, 0.018	[0.04, 3.91]	
Amines aromatic or phenols4	0.576, −0.042	28	0.297	0.838, 0.814	[1.32, 6.06]	
Amines aromatic or phenols5	0.552, 0.114	12	0.260	0.802, 0.728	[1.18, 3.91]	
Primary amines	0.189, −0.059	4	0.248	0.390, 0.095	[−1.31, 1.49]	
Secondary and tertiary amines	0.133, 0.200	4	0.150	0.517, 0.040	[−1.50, 1.45]	
Hydrazines	0.190, 1.987	5	0.289	0.766, 0.360	[−2.46, 4.70]	C
Amides and imides	0.212, 0.585	8	0.593	0.151, 0.135	[0.23, 3.80]	
Esters aliphatic	0.666, −0.819	6	0.324	0.927, 0.762	[0.25, 5.41]	
Esters aromatic	0.459, −0.417	3	0.010	1.000, 0.998	[1.60, 4.72]	
Aldehydes	0.521, 0.295	5	0.555	0.616, 0.084	[0.42, 4.47]	C
Acids	0.222, −0.113	7	0.644	0.133, 0.298	[0.08, 4.20]	
Acids acrylic	0.057, 0.248	3	0.143	0.025, 0.947	[0.35, 1.33]	
Conjugated systems1	0.630, 1.393	5	0.321	0.957, 0.916	[−1.76, 4.65]	C
Conjugated systems2	0.213, 0.906	11	0.775	0.097, 0.047	[0.17, 3.70]	
Thiols aromatic	NO–QSAR					
Thiols aliphatic	0.427, 1.410	4	0.786	0.647, 0.049	[−0.17, 6.12]	C
Sulfides	NO–QSAR					
Disulfides	1.041, −0.724	3	0.480	0.865, 0.635	[1.74, 4.44]	C
Carbamates	0.046, 2.991	4	0.688	0.008, 0.523	[0.94, 4.60]	
Pyrethroids	NO–QSAR					
Acrylates	0.003, 1.401	4	0.069	0.002, 0.646	[−0.21, 2.36]	
Methacrylates	0.461, −0.422	5	0.301	0.824, 0.653	[0.47, 4.54]	
Epoxides	0.486, 0.589	4	0.341	0.817, 0.598	[0.08, 3.98]	C
Barbitals or thiols other	NO–QSAR					
Esters phosphate	2.133, −2.376	3	1.477	0.204, 0.526	[3.08, 3.88]	
N or P cations	NO–QSAR					
Halides1	−0.665, 4.825	3	0.350	0.800, 0.998	[2.09, 4.50]	
Halides2	0.880, −0.317	4	0.552	0.860, 0.494	[1.10, 5.04]	
Halides3	0.826, −1.008	24	0.237	0.901, 0.883	[1.47, 4.73]	
Metals	NO–QSAR					
Nitriles aliphatic	NO–QSAR					N
Ketones	NO–QSAR					N
Alcohols or ethers aliphatic	0.641, −1.053	6	0.214	0.958, 0.923	[1.10, 5.82]	N
Phosphates	0.579, −0.634	3	0.103	0.983, 0.922	[1.44, 4.59]	N
Hydrocarbons aliphatic	0.660, −0.555	10	0.268	0.891, 0.797	[2.42, 6.54]	N
Ethers aliphatic	NO–QSAR					N
Ethers aromatic	0.492, 0.285	4	0.437	0.406, 0.088	[2.16, 4.21]	N
Neutral organics	0.696, −0.870	26	0.418	0.857, 0.835	[0.68, 6.54]	
Unclassified	0.537, 0.078	12	1.097	0.475, 0.287	[−1.02, 5.50]	

*[1] *C*: an equation is generated by the calculated Clog *P*. N: a member of the *Neutral organics* class. Note. *n*, RMSE, r^2, and q^2 denote the number of chemicals in a class, the root mean square error, the squared correlation coefficient, and the leave-one-out version of the squared correlation coefficient, respectively. The log *P* range shows minimum and maximum log *P* values.

2.6 *KATE system software*

The KATE software was first made available to the public in January 2008. An updated version of KATE, including standalone personal computer and internet versions, was released in March 2009. The standalone version, called 'KATE on PAS', and the internet version, called 'KATE on NET', adopted the KOWWIN^{TM} [15] of the US EPA, and Clog *P* [11] estimated by the Daylight system, respectively, to estimate the calculated log *P*. Except for the treatment of calculated log *P* values, KATE on PAS and KATE on NET use the same classification algorithm, *fragment identification by tree structure* (FITS), developed by Yoshioka.

In the KATE system, the input is *simplified molecular input line entry specification* (SMILES) and log *P* (if available) for toxicity prediction, and the output is the calculated toxicity concentration (LC_{50} or EC_{50}), the QSAR class found for the predicted chemical, and the domain judgements. If the measured log *P* of a chemical is not available, the calculated log *P* according to the SMILES information (KOWWIN or *C* log *P*) is adopted.

3. Methods of QSAR validation

First, leave-one-out cross validations were examined for training sets used in the QSAR equations of KATE. Secondly, external validations were performed using test set compounds not included in the KATE training sets due to lack of measured log *P* values. The 287 fish 96-hour LC_{50} and 98 daphnia 48-hour EC_{50} from the Japan MoE, along with the US EPA fathead minnow database, were used for comparison of the calculated toxicity by the KATE software version published in March 2009, TIMES v. 2.25, and ECOSAR v. 0.99 h (1999).

It is worth mentioning that the end-points of the data calculated by KATE were not identical to those calculated by TIMES and ECOSAR. Fish (mixed with *Oryzias latipes* and fathead minnow acute toxicity tests) 96-hour LC_{50} and daphnia 48-hour EC_{50} (KATE), *Pimephales promelas* 96-hour LC_{50} and daphnia 48-hour EC_{50} (TIMES), and fish 96-hour LC_{50} and daphnia 48-hour LC_{50} (ECOSAR) were therefore adopted. The input of KATE and ECOSAR were SMILES strings, and calculated log *P* by KOWWIN. In TIMES, only the lists of SMILES strings were used as input values, and quantum chemical calculations were performed using MOPAC AM1 Hamiltonian, using the 'precise' option, without taking other conformers into account.

4. Results and discussion

4.1 *Cross validation*

The QSAR equations were validated by the leave-one-out method obtained from the KATE system. The complete list of results is given in Appendix 4 of the supplementary material (available online). The statistical data are displayed in Tables 1 and 2. The criterion proposed by Hulzebos and Posthumus [16] was evaluated, in which the estimations from models should not deviate from the experimental value by a factor of 10 or above. For fish, 575 of the 628 chemicals met the acceptable criteria, and for daphnia 241 of 290 did so. (In this instance the 628 and 290 chemicals involved some degree of duplication.) Using the QSAR equations in the KATE system, more than 80% of chemicals were predicted within a factor of 10. The classes with less than a 0.7 squared correlation coefficient ($r^2 < 0.7$), and/or more than 0.5 RMSE, tended to increase the

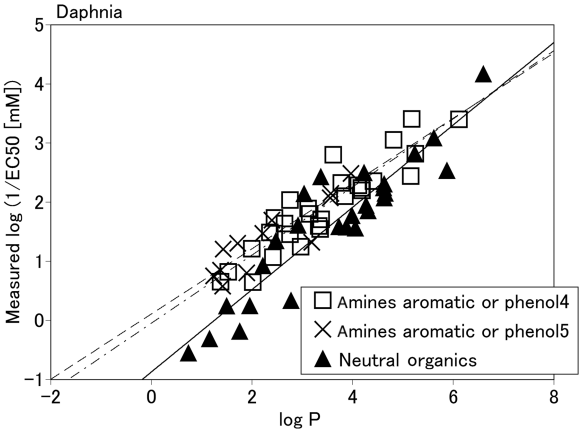

Figure 1. The correlation between log P and the measured toxicity values of chemicals used in KATE as a daphnia end-point. The dotted-dashed, dashed and bold lines are the QSAR equations of *amines aromatic or phenols4*, *amines aromatic or phenols5*, and *neutral organics*, respectively.

number of chemical substances in the *unacceptable* group. For example, the fish *hydrocarbons aromatic* class had 43 reference data, $r^2 = 0.826$, RMSE = 0.368, and only one unacceptable chemical. In other words, 98% of the chemicals were classed as acceptable. On the other hand, the fish *dinitrobenzene* class contained 12 reference data, $r^2 = 0.331$, RMSE = 0.669, and three unacceptable chemicals. In this case, 75% of the chemicals were thus acceptable.

As shown in Tables 1 and 2, each of the classes with $r^2 \geq 0.7$, RMSE ≤ 0.5, and $n > 5$, e.g., the fish *hydrocarbon aromatic* class, had a sufficiently high q^2. Such classes showed QSAR equations similar to those of *neutral organics*. Thus the toxicity of such classes could be explained mainly by the narcotic effect of the chemicals. However, the daphnia *amines aromatic or phenols4* and *amines aromatic or phenols5* groups had a larger intercept b in the QSAR equations than *neutral organics* with a small log P value (see Figure 1). These classes can be explained in terms of polar narcosis or narcosis II [17]. Narcosis II is known to be more toxic than baseline toxicity, i.e., than *neutral organics*, non-polar narcosis, narcosis I, or less inert, as explained by Verhaar et al. [18].

In some cases the q^2 values were much smaller than those of r^2. QSAR equations based on fewer than six reference data require a greater number of reference chemicals.

4.2 *External validation*

Tables 3 and 4 list the statistical data of the TIMES, ECOSAR, and KATE with or without the applicability domains. The complete results are given in Appendix 5 of the supplementary material. First, we will focus on the TIMES, ECOSAR, and all the KATE results, without considering any applicability domains. In fish, the determination coefficient, r^2, and RMSE using KATE ($r^2 = 0.868$ and RMSE = 0.658) were larger and smaller, respectively, than those using TIMES ($r^2 = 0.751$ and RMSE = 0.935) and than by ECOSAR ($r^2 = 0.790$ and RMSE = 0.869). For daphnia, RMSE using KATE (0.993) was smaller than that using TIMES (1.404) and ECOSAR (1.364). However, r^2 using KATE (0.662) showed no noticeable advantage over that by TIMES (0.668) or ECOSAR (0.699).

Table 3. Statistical information comparing measured and calculated fish $\log(1/LC_{50}[\text{mM}])$ of 287 test set compounds. The complete results are shown in Appendix 5-1.

			KATE[*2]					
	TIMES[*1]	*ECOSAR*[*2]	*All*[*3]	*log P*[*4]	*C(1)*[*5]	*C(2)*[*6]	*log P*[*4] *C(1)*[*5]	*log P*[*4] *C(2)*[*6]
Chemicals[*7]	274	242	274	207	152	192	111	144
Predicted[*8]	274	259	318	252	187	233	145	179
r^2	0.751	0.790	0.868	0.833	0.901	0.890	0.886	0.866
RMSE	0.935	0.869	0.685	0.641	0.644	0.655	0.588	0.617
Under[*9] [%]	11.3	10.0	4.7	5.2	5.3	5.6	2.8	3.9
Over[*10] [%]	5.1	8.1	7.2	6.7	8.0	6.9	8.3	7.3

Notes:
[*1]Each chemical is identified by one QSAR class.
[*2]When a chemical is found to belong to more than one QSAR class, all the estimated data are adopted. If only the name of the class is available, such data are omitted.
[*3]Both in-domain and out-of-domain data for log *P* and *C*-judgements are included.
[*4]In-domain of log *P*-judgement.
[*5]In-domain of *C*-judgement is defined as all substructures of a test chemical being found in reference chemicals in the class.
[*6]In-domain of *C*-judgement defined as all substructures of a test chemical being in reference chemicals in either *Neutral organics* or the class.
[*7]The number of compounds that can be predicted.
[*8]The total number of the predicted values by using the training sets. Some chemicals belong to more than one class, and thus *Predicted* is larger than *Chemicals*. r^2, RMSE, Under and Over were calculated based on the *Predicted* number.
[*9]Fractions (%) of the underestimated chemicals. Underestimation is defined as [calculated $\log(1/LC_{50})$ – measured $\log(1/LC_{50})] < -1$.
[*10]Fractions (%) of the overestimated chemicals. Overestimation is defined as [calculated $\log(1/LC_{50})$ – measured $\log(1/LC_{50})] > 1$.

Table 4. Statistical information between measured and calculated *Daphnia* $\log(1/EC_{50}[\text{mM}])$ for 98 test set compounds. The complete results are shown in Appendix 5-2.

			KATE[*2]					
	TIMES[*1]	*ECOSAR*[*2]	*all*[*3]	*log P*[*4]	*C(1)*[*5]	*C(2)*[*6]	*log P*[*4] *C(1)*[*5]	*log P*[*4] *C(2)*[*6]
Chemicals[*7]	93	82	94	58	43	55	25	33
Predicted[*8]	93	85	102	66	46	61	31	39
r^2	0.668	0.699	0.662	0.732	0.793	0.686	0.807	0.801
RMSE	1.404	1.364	0.993	0.784	0.799	0.968	0.639	0.689
Under[*9] [%]	21.5	14.1	9.8	1.5	6.5	8.2	0.0	0.0
Over[*10] [%]	11.8	18.8	14.7	15.2	6.5	11.5	6.5	10.3

Notes: As in Table 3.

Since reference data for the daphnia end-point (258 chemicals) numbered only half of those for fish (535 chemicals), the reference data for each QSAR equation for daphnia would therefore be less satisfactory for predicting toxicity. The addition of reference data and a change in the classification rules can recover the values of the statistical data.

A fraction of $\log(1/LC_{50})$ with an underestimation of less than -1 indicated that, compared with KATE, TIMES and ECOSAR tended to underestimate the toxicities of both fish and daphnia. On the other hand, a fraction of $\log(1/LC_{50})$ showing an overestimation of more than 1 indicated that, compared with TIMES, ECOSAR and KATE tended to overestimate toxicity in both fish and daphnia. Considering these under- and over-estimation fractions, we find that KATE gives a higher predictive ability in acute *Oryzias latipes* and *Daphnia magna* toxicity tests than does TIMES or ECOSAR. If the alert: *Out of domain*, in TIMES, and the applicable $\log P$ range in ECOSAR are considered rigidly, the correlation between measured and calculated toxicity is improved in TIMES and ECOSAR.

Secondly, in fish, the RMSE of one of any in-domains was smaller than if domains were not considered. However, the r^2 in-domain of $\log P$ showed no particular improvement. For daphnia, r^2 and RMSE for one of any in-domains were larger and smaller, respectively, than those without considering domains. In the present study, either the descriptor and/or structural domains were related to the reduction of RMSE and the fraction of underestimated chemicals, especially if both domains were considered simultaneously. Additionally, the stricter structural domain C(1) (shown in Tables 3 and 4) demonstrated better predictive performance than the structural domain C(2). The systematic study of the domain based on the atom-centred fragment (ACF) approach by Kuhne et al. [14] showed that the ACF varied with respect to its size in terms of the path length, and the ACF match mode was specified in terms of degree of strictness. They also demonstrated a clear relationship between predictive performance and the levels of the ACF definition and match mode [14]. Even though the definition of substructures for the domain are different, the improvement by using *C*-judgement is similar in concept to that using the ACF approach. Thus, the $\log P$ range of the equation and *C*-judgement are useful for assessing the applicability of the QSAR results.

5. Summary

We have reported on the KATE system, encompassing a full list of classifications of the QSAR equations and KATE validations. In the KATE system chemicals are classified by their substructure. The QSAR equations express the correlation between $\log P$ and $\log(1/LC_{50})$ or $\log(1/EC_{50})$ of a chemical by simple linear regression analyses. The classes of QSAR equations are characterised by fragments of chemicals, except for the *neutral organics* class. The descriptor and structure domains, $\log P$ and *C*-judgements, in KATE were also introduced.

The cross-validation of the KATE system showed that QSAR equations with higher r^2 and lower RMSE with $n > 5$ gave a reliably higher q^2 than the other QSAR equations in KATE, meaning they had better predictive ability. A comparison of KATE, TIMES, and ECOSAR revealed that KATE was more accurate, due to end-point dependence. The use of $\log P$ and the *C*-judgement improved the statistical data. Thus the KATE system is a powerful tool for predicting acute toxicity in *Oryzias latipes* and *Daphnia magna* when the $\log P$ and *C*-judgement can be confirmed. Also, KATE has the potential to be useful in risk assessment.

The next topics in QSAR development will be to consider the reactivity of chemicals, and to include multi-regression analysis. The quantum chemical parameters, such as partial charges, are candidates for additional descriptors. Other ways of significantly

increasing the reliability of toxicity prediction will be to improve the classification of the substructures, increase the reference data in a QSAR equation, and to refine the *C*-judgement.

Acknowledgements

KATE was researched and developed by the Research Center for Environmental Risk at the NIES, under contract to the Japanese MoE between 2004 and 2008. We also wish to thank the US EPA for permission to use KOWWIN in KATE on PAS, the standalone version of the KATE system. We are grateful to Mr K. Hasunuma and Ms K. Sugiyama for their support and encouragement with the KATE publication.

References

[1] T.I. Netzeva, M. Pavan, and A.P. Worth, *Review of (quantitative) structure–activity relationships for acute aquatic toxicity*, QSAR Comb. Sci. 27 (2008), pp. 77–90.
[2] OECD, *Report on the regulatory uses and applications in OECD member countries of (quantitative) structure–activity relationship [(Q)SAR] models in the assessment of new and existing chemicals*, Environment Health and Safety Publications Series on Testing and Assessment, No. 58, OECD, Paris, 2006.
[3] MoE, Japan ecotoxicity tests data. Available at http://www.env.go.jp/chemi/sesaku/02e.pdf
[4] KATE. Available at http://kate.nies.go.jp Copyright (C) 2008–2009 Ministry of the Environment, Government of Japan, all rights reserved. It is cautioned that these QSAR results may not be used as ecotoxicity test results required for MoE submissions in compliance with the CSCL.
[5] O.G. Mekenyan, S.D. Dimitrov, T.S. Pavlov, and G.D. Veith, *A systematic approach to simulating metabolism in computational toxicology. I. The TIMES heuristic modelling framework*, Curr. Pharm. Des. 10 (2004), pp. 1273–1293.
[6] S.D. Dimitrov, O.G. Mekenyan, G.D. Sinks, and T.W. Schultz, *Global modeling of narcotic chemicals: Ciliate and fish toxicity*, THEOCHEM. 622 (2003), pp. 63–70.
[7] U.S. EPA, ECOSAR^{TM}. Available at http://www.epa.gov/oppt/newchems/tools/21ecosar.htm. See also http://www.epa.gov/oppt/newchems/tools/ecosartechfinal.pdf
[8] US EPA, fathead minnow database. Available at http://www.epa.gov/med/Prods_Pubs/fathead_minnow.htm
[9] C.L. Russom, S.P. Bradbury, S.J. Broderius, D.E. Hammermeister, and R.A. Drummond, *Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas)*, Environ. Toxicol. Chem. 16 (1997), pp. 948–967.
[10] OECD, *Report of the OECD Workshop on quantitative structure activity relationships (QSARs) in aquatic effects assessment, Environment Monographs, No. 58*, OECD, Paris, 1992.
[11] Clog P, Daylight Chemical Information Systems, Inc. Available at http://www.daylight.com/dayhtml/doc/clogP/index.html. The underlying program, CLOG P, is copyrighted by Pomona College and BioByte, Inc., of Claremont, CA.
[12] S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela, and O. Mekenyan, *A stepwise approach for defining the applicability domain of SAR and QSAR models*, J. Chem. Inf. Model. 45 (2005), pp. 839–849.
[13] T.W. Schultz, M. Hewitt, T.I. Netzeva, and M.T.D. Cronin, *Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action*, QSAR Comb. Sci. 26 (2007), pp. 238–254.
[14] R. Kuhne, R.U. Ebert, and G. Schuurmann, *Chemical domain of QSAR models from atom-centered fragments*, J. Chem. Inf. Model. 49 (2009), pp. 2660–2669.

[15] US EPA, KOWWIN™. Available at http://www.epa.gov/opptintr/exposure/pubs/episuite.htm
[16] E.M. Hulzebos and R. Posthumus, *(Q)SARs: Gatekeepers against risk on chemicals?*, SAR QSAR Environ. Res. 14 (2003), pp. 285–316.
[17] G.D. Veith and S.J. Broderius, *Rules for distinguishing toxicants that cause Type-I and Type-II Narcosis syndromes*, Environ. Health Perspect. 87 (1990), pp. 207–211.
[18] H.J.M. Verhaar, C.J. Vanleeuwen, and J.L.M. Hermens, *Classifying environmental pollutants: 1. Structure–activity relationships for prediction of aquatic toxicity*, Chemosphere 25 (1992), pp. 471–491.