SciMatics SciQSAR model for acute toxicity in *Daphnia magna* (48h immobilisation, $EC_{50}$)

1. QSAR identifier

1.1 QSAR identifier (title)

SciMatics SciQSAR model for acute toxicity in *Daphnia magna* (48h immobilisation, $EC_{50}$), Danish QSAR Group at DTU Food.

1.2 Other related models

Leadscope Enterprise model for acute toxicity in *Daphnia magna* (48h immobilisation, $EC_{50}$), Danish QSAR Group at DTU Food.

1.3. Software coding the model

SciMatics SciQSAR version 2.3.0.0.12.

2. General information

2.1 Date of QMRF

January 2015.

2.2 QMRF author(s) and contact details

QSAR Group at DTU Food;

Danish National Food Institute at the Technical University of Denmark;

http://qsar.food.dtu.dk/;

qsar@food.dtu.dk

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

Marianne Dybdahl;

National Food Institute at the Technical University of Denmark;

Sine Abildgaard Rosenberg;

National Food Institute at the Technical University of Denmark;

2.3 Date of QMRF update(s)

2.4 QMRF update(s)

2.5 Model developer(s) and contact details

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Jay Russel Niemelä;

National Food Institute at the Technical University of Denmark;

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

Danish QSAR Group at DTU Food;

National Food Institute at the Technical University of Denmark;

http://qsar.food.dtu.dk/;

qsar@food.dtu.dk

2.6 Date of model development and/or publication

January 2014.

2.7 Reference(s) to main scientific papers and/or software package

Contrera, J.F., Matthews, E.J., Kruhlak, N.L., and Benz, R.D. (2004) Estimating the safe starting dose in phase I clinical trials and no observed effect level based on QSAR modelling of the human maximum recommended daily dose. *Regulatory Toxicology and Pharmacology*, 40, 185 – 206.

SciQSAR (2009) Reference guide: *Statistical Analysis and Molecular Descriptors*. Included within the SciMatics SciQSAR software.

2.8 Availability of information about the model

The training set is non-proprietary and was compiled from the references in 9.2. The model algorithm is proprietary from commercial software.

2.9 Availability of another QMRF for exactly the same model

## 3. Defining the endpoint

### 3.1 Species

Water flea (*Daphnia magna*).

### 3.2 Endpoint

QMRF 3. Ecotoxic effects

QMRF 3. 1. Short-term toxicity to *Daphnia* (immobilisation)

OECD 202 *Daphnia* sp., Acute Immobilisation Test

### 3.3 Comment on endpoint

Water pollution has become a major threat to the existence of living organisms in aquatic environment. A huge quantity of pollutants in the form of domestic and industrial effluents is discharged directly or indirectly into the water bodies and can have severe impact on the biotic and abiotic environment.
In view of the importance of cladocerons (i.e. small crustaceans) as an important link in the food chain in aquatic ecosystems, testing the toxic effects of waste effluents to the small, planktonic crustacean *Daphnia magna* (water flea) is relevant when regulating the discharge of wastewater treatment systems.
*Daphnia magna* is highly sensitive to toxic substances, has short generation time, multiplies very rapidly, easily acclimatizes to laboratory condition, can be cultured in a small space and can be measured in a relatively short period.
The training set consists of data for acute toxicity to *Daphnia magna*. The concentration of the test chemical at which 50% of the Daphnids are immobilised (i.e. can't swim) after 48 hours exposure is used as the endpoint.

### 3.4 Endpoint units

$-\log(EC_{50})$.

### 3.5 Dependent variable

$EC_{50}$ (immobilisation, 48h), in µM.

### 3.6 Experimental protocol

The experimental protocol is described in OECD guideline 202 (2004). Briefly, young daphnids, aged less than 24 hours at the start of the test, are exposed to the test substance at a range of concentrations for a period of 48 hours. Immobilisation is recorded at 24 hours and 48 hours and compared with control values. The results are analysed in order to calculate the half maximal effective concentration ($EC_{50}$, mg/L) at 48 hours of exposure. $EC_{50}$ is the concentration estimated to immobilise 50 % of the daphnids within a stated exposure period (48 hours).

3.7 Endpoint data quality and variability

As data are compiled from multiple sources some degree of variability in data is expected.

## 4. Defining the algorithm

### 4.1 Type of model

This is a continuous (Q)SAR model based on calculated molecular descriptors, and if available the modeller's own or third-party descriptors or measured endpoints can be imported and used as descriptors.

### 4.2 Explicit algorithm

This is a continuous (Q)SAR model made by use of partial least squares (PLS) regression (see 4.5). The specific implementation is proprietary within the SciQSAR software.

### 4.3 Descriptors in the model

Molecular connectivity indices

Molecular shape indices

Topological indices

Electrotopological (Atom E and HE-States) indices

Electrotopological bond types indices

SciQSAR software provides over 400 built-in molecular descriptors. Additionally, SciQSAR makes it possible to import the modeller's own or third-party descriptors or use measured endpoints as custom descriptors.

### 4.4 Descriptor selection

A built-in genetic algorithm (GA) analysis is used by SciQSAR to select the descriptors for the model. The GA method sequentially generates sets of descriptors. Selection of the best set of descriptors is accomplished through an algorithm which simulates mutation and genetic cross-over. Each set of descriptors (i.e. generation) is evaluated and its ''goodness of fit'' is determined by a set of criteria. The algorithm makes use of the initial pool of descriptors to select the set of descriptors with the best regression statistics. The performance of each candidate model is assessed using an automated cross-validation process within SciQSAR. (Contrera *et al.*2004)

17 descriptors were selected from the initial pool of descriptors and distributed on 6 PLS components used to build the model.

4.5 Algorithm and descriptor generation

SciQSAR uses genetic algorithms (GA) to select descriptors for the model (see 4.4) (Contrera *et al.*2004).

For a binary classification problem SciQSAR uses discriminant analysis (DA) to make a (Q)SAR model. SciQSAR implements a broad range of discriminant analysis (DA) methods including parametric and non-parametric approaches. The classic parametric method of DA is applicable in the case of approximately normal within-class distributions. The method generates either a linear discriminant function (the within-class covariance matrices are assumed to be equal) or a quadratic discriminant function (the within-class covariance matrices are assumed to be unequal). When the distribution is assumed to not follow a particular law or is assumed to be other than the multivariate normal distribution, non-parametric DA methods can be used to derive classification criteria. The non-parametric DA methods available within SciQSAR include the kernel and *k*-nearest-neighbor (kNN) methods. The main types of kernels implemented in SciQSAR include uniform, normal, Epanechnikov, bi-weight, or tri-weight kernels, which are used to estimate the group specific density at each observation. Either Mahalanobis or Euclidean distances can be used to determine proximity between compound-vectors in multidimensional descriptor space. When the kNN method is used, the Mahalanobis distances are based on the pooled covariance matrix. When the kernel method is used, the Mahalanobis distances are based on either the individual within-group covariance matrices or the pooled covariance matrix. (Contrera *et al*. 2004)

If the data outcome is continuous, regression analysis is used to build the predictive model. Within SciQSAR several regression methods are available: ordinary multiple regression (OMR), stepwise regression (SWR), all possible subsets regression (PSR), regression on principal components (PCR) and partial least squares regression (PLS). The choice of regression method depends on the number of independent variables and whether correlation or multicollinearity among the independent variables exists: OMR is acceptable with a small number of independent variables, which are not strongly correlated. SWR is used under the same circumstances as OMR but with greater number of variables. PSR is used for problems with a great number of independent variables. PCR and PLS are useful when a high correlation or multicollinearity exist among the independent variables. (SciQSAR 2009)

To test how stable the developed models are, SciQSAR have built-in leave-one-out cross-validation procedures (see 6.).

4.6 Software name and version for descriptor generation

SciMatics SciQSAR version 2.3.0.0.12.

4.7 Descriptors/chemicals ratio

In this model 17 descriptors were used. Dimensionality was reduced by applying 6 PLS components to make this model. No information about the number of descriptors contained in each PLS component is given by SciQSAR.

## 5. Defining Applicability Domain

### 5.1 Description of the applicability domain of the model

The definition of the applicability domain consists of two components; the definition in SciQSAR and the in-house further refinement algorithm on the output from SciQSAR to reach the final applicability domain call.

1. SciQSAR

The first criterion for a prediction to be within the models applicability domain is that all of the descriptor values for the test compound can be calculated by SciQSAR. If SciQSAR cannot calculate each descriptor value for the test chemical no prediction value is given by SciQSAR and it is considered outside the model's applicability domain.

2. The Danish QSAR group

In addition to the general SciQSAR applicability domain definition the Danish QSAR group has applied two further requirements to the applicability domain of the model. First, the logP value of the query compound should fall within the logP interval of the model's training set [-4.47;7]. Secondly, only predictions that falls within the response variable $EC_{50}$ interval (µM) [0.00006;810936.20] of the model's training set are considered reliable and therefore accepted.

### 5.2 Method used to assess the applicability domain

The system does not generate predictions if it cannot calculate each descriptor value for the test compound.

Only compounds with a logP value within the logP interval [-4.47;7] are within the applicability domain. The generated predictions should fall within the response variable interval [0.00006;810936.20] of the training set. Any prediction outside this interval is set to the closest response variable limit (0.00006 or 810936).

### 5.3 Software name and version for applicability domain assessment

SciMatics SciQSAR version 2.3.0.0.12.

### 5.4 Limits of applicability

The Danish QSAR group applies an overall definition of structures acceptable for QSAR processing which is applicable for all the in-house QSAR software, i.e. not only SciQSAR. According to this definition accepted structures are organic substances with an unambiguous structure, i.e. so-called discrete organics defined as: organic compounds with a defined two dimensional (2D) structure containing at least two carbon atoms, only certain atoms (H, Li, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and I), and not mixtures with two or more 'big components' when analyzed for ionic bonds (for a number of small known organic ions assumed not to affect toxicity the 'parent molecule' is accepted). Structures with less than two carbon atoms or containing atoms not in the list above (e.g. heavy metals) are rendered out as not acceptable for further QSAR processing. Calculation 2D structures (SMILES and/or SDF) are generated by stripping off

accepted organic and inorganic ions. Thus, all the training set and prediction set chemicals are used in their non-ionized form. See 5.1 for further applicability domain definition.

6. Internal validation

6.1 Availability of the training set

Yes

6.2 Available information for the training set

CAS

SMILES

6.3 Data for each descriptor variable for the training set

No

6.4 Data for the dependent variable for the training set

All

6.5 Other information about the training set

626 compounds are in the training set.

6.6 Pre-processing of data before modelling

The training set $EC_{50}$ (48h) results were given in mg/L and were converted to -log(μM) before modelling.

Only structures acceptable for Leadscope were used in the final training set. That is only discrete organic chemicals as described in 5.4 were used. In case of replicate structures, one of the replicates was kept if all the compounds had the same activity and all were removed if they had different activity. No further structures accepted by the software were eliminated (i.e. outliers).

6.7 Statistics for goodness-of-fit

SciQSAR's own internal performance test gave the following result for predictions within the applicability domain as defined by SciQSAR (i.e. the first criterion described in 5.1):

R-squared: 0.65

6.8 Robustness – Statistics obtained by leave-one-out cross-validation

SciQSAR's own internal leave-one-out (LOO) cross-validation procedure was used for predictions within the applicability domain as defined by SciQSAR (i.e. the first criterion described in 5.1). This gave the following result:

Q-squared: 0.63

6.9 Robustness – Statistics obtained by leave-many-out cross-validation

Not performed.

6.10 Robustness - Statistics obtained by Y-scrambling

Not performed.

6.11 Robustness - Statistics obtained by bootstrap

Not performed.

6.12 Robustness - Statistics obtained by other methods

Not performed.

## 7. External validation

### 7.1 Availability of the external validation set

### 7.2 Available information for the external validation set

### 7.3 Data for each descriptor variable for the external validation set

### 7.4 Data for the dependent variable for the external validation set

### 7.5 Other information about the training set

### 7.6 Experimental design of test set

### 7.7 Predictivity – Statistics obtained by external validation

### 7.8 Predictivity – Assessment of the external validation set

### 7.9 Comments on the external validation of the model

External validation has not been performed for this model.

## 8. Mechanistic interpretation

### 8.1 Mechanistic basis of the model

The SciQSAR software provides over 400 calculated physico–chemical, electrotopological E-state, connectivity and other molecular descriptors. The descriptors selected for the model may indicate modes of action that are obvious for persons with expert knowledge about the endpoint.

### 8.2 A priori or posteriori mechanistic interpretation

A posteriori mechanistic interpretation. The descriptors selected for the model may provide a basis for mechanistic interpretation.

### 8.3 Other information about the mechanistic interpretation

## 9. Miscellaneous information

### 9.1 Comments

The model can be used to predict if a chemical is acute toxic (48h) to *Daphnia magna*. The Danish QSAR Group applies an algorithm on top of the predictions from the model in order to convert the values from -log(µM) to mg/L, which is the normal unit for this endpoint.

### 9.2 Bibliography

OECD guideline 202 (2004) OECD Guidelines for the Testing of Chemicals No. 202: *Daphnia* sp., Acute Immobilisation Test. Organisation for Economic Cooperation and Development; Paris, France. Available online at: http://www.oecd-ilibrary.org/environment/test-no-202-daphnia-sp-acute-immobilisation-test_9789264069947-en

The training set for the model was compiled from the following sources:

AQUIRE (2000) AQUatic Information REtrieval, US EPA database established in 1981 and in 1995 a component of US EPA ECOTOX Database. Data was compiled from the database in 2000. The database is available online at http://www.epa.gov/med/Prods_Pubs/ecotox.htm#aquatic.

Hermens, J., Canton, H., Janssen, P. and de Jong, R. (1984) Quantitative Structure-Activity Relationships and toxicity studies of mixtures of chemicals with anaesthetic potency: Acute lethal and sublethal toxicity to Daphnia magna. *Aquatic Toxicology*, 5, 143-154.

DK- EPA tests (not published in-house data):

- Immobilisation tests of selected organic amines with the crustacean Daphnia magna. VKI, Department of Ecotoxicology, study director Finn Pedersen
- Immobilization Test of Three Trialkylamine Compounds with the Crustacean Daphnia magna. VKI, Department of Ecotoxicology, study director Finn Pedersen.
- Immobilisation Test of Aniline Compounds with the Crustacean Daphnia magna. VKI, Department of Ecotoxicology, study director Finn Pedersen.

Kühn, R., Pattard, M., Pernak, K-D, and Winter, A. (1989) Results of the harmful effects of selected water pollutants (anilines, phenols, aliphatic compounds) to Daphnia magna. *Water Research,* 23:4, 495-499.

PhD thesis (1998) Aquatic Toxicity of Polar Narcotic Pollutants, by E. Urrestarazu Ramos, University of Utrecht, ISBN 90-393-1638-4, 82-85.

Hansen, O.C. (1999) Quantitiative Structure-Activity relationships (QSAR) and Pesticides. Danish EPA report April 1999.

### 9.3 Supporting information