SciMatics SciQSAR model for acute toxicity to *Fathead minnow* (96h mortality, $LC_{50}$)

1. QSAR identifier

1.1 QSAR identifier (title)

SciMatics SciQSAR model for acute toxicity to *Fathead minnow* (96h mortality, $LC_{50}$), Danish QSAR Group at DTU Food.

1.2 Other related models

Leadscope Enterprise model for acute toxicity to *Fathead minnow* (96h mortality, $LC_{50}$), 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 US EPA MED-Duluth Fathead minnow database (MED-Duluth) in 1999. 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

Fish (Fathead minnow, i.e. *Pimephales promelas)*.

### 3.2 Endpoint

QMRF 3. Ecotoxic effects

QMRF 3. 3. Acute toxicity to fish (lethality)

OECD 203 Fish, Acute Toxicity 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, which has severe impacts on its biotic and abiotic environment. A typical endpoint used in initial effect assessment of a chemical on aquatic organisms is the 96h LC50 value for the fish, fathead minnow. The fathead minnow (*Pimephales promelas*) is a species of temperate freshwater fish belonging to the *Pimephelas* genus.
The training set consists of data for acute toxicity to fathead minnow from the US EPA MED-Duluth Fathead minnow database (MED-Duluth). MED-Duluth tested a series of industrial organic compounds using the fathead minnow for the purpose of developing an expert system to predict the acute mode of toxic action from chemical structure. The entire Duluth fathead minnow database and results related to the acute mode of action are presented in Russom *et al.* (1997), see references 15. to 19. in Russom *et al.* (1997).

### 3.4 Endpoint units

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

### 3.5 Dependent variable

Acute toxicity to fathead minnow (96h lethal concentration): $LC_{50}$, in μM.

### 3.6 Experimental protocol

The experimental protocol is described in OECD guideline 203 (1992). Briefly, the fish are exposed to the different concentrations of the test substance preferably for a period of 96 hours. Mortalities are recorded at 24, 48, 72 and 96 hours and the concentration that kills 50% of the fish (lethal concentration, $LC_{50}$, in mg/L) after 96 hours is estimated.

3.7 Endpoint data quality and variability

The data is of good quality and as all experimental results originate from the same source (MED-Duluth) the variability in data is expected to be low.

## 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)

20 descriptors were selected from the initial pool of descriptors and distributed on 20 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 20 descriptors were used. Dimensionality was reduced by applying 20 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 [-4.34;6]. Secondly, only predictions that falls within the response variable $LC_{50}$ interval (µM) [0.0004;707945.78] 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 [-4.34;6] are within the applicability domain. The generated predictions should fall within the response variable interval [0.0004;707945.78] of the training set. Any prediction outside this interval is set to the closest response variable limit (0.0004 or 707945).

### 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

565 compounds are in the training set.

6.6 Pre-processing of data before modelling

The training set $LC_{50}$ (96h) results were given in mg/L and were converted to -log(µM) before modelling.

Only structures acceptable for SciQSAR 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: 0.64

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: 0.60

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 (96h) to *Fathead minnow* (fish). 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

MED-Duluth: Geiger, D.L., Call, D.J., and Brooke, L.T. (1988) Acute Toxicities of Organic Chemicals to Fathead Minnows (*Pimephales promelas*). Vol. 1-4, Center for Lake Superior Environmental Studies, University of Wisconsin-Superior, USA. The database is available online at http://www.epa.gov/med/Prods_Pubs/fathead_minnow.htm

OECD guideline 203 (1992) OECD Guidelines for the Testing of Chemicals No. 203: Fish, Acute Toxicity Test. Organisation for Economic Cooperation and Development; Paris, France. Available online at: http://www.oecd-ilibrary.org/environment/test-no-203-fish-acute-toxicity-test_9789264069961-en

Russom, C.L., Bradbury, S.P., Broderius, S.J., Hammermeister, D.E., and Drummond, R.A. (1997) Predicting modes of action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*). *Environmental Toxicology and Chemistry*, 16:5, 948-967.

### 9.3 Supporting information