SciMatics SciQSAR model for carcinogenicity exclusively in rodent liver *in vivo*

## 1. QSAR identifier

### 1.1 QSAR identifier (title)

SciMatics SciQSAR model for carcinogenicity exclusively in rodent liver *in vivo*, Danish QSAR Group at DTU Food.

### 1.2 Other related models

MultiCASE CASE Ultra model for carcinogenicity exclusively in rodent liver *in vivo*, Danish QSAR Group at DTU Food.

Leadscope Enterprise model for carcinogenicity exclusively in rodent liver *in vivo*, Danish QSAR Group at DTU Food.

### 1.3. Software coding the model

SciQSAR version 3.1.00.

## 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

Sine Abildgaard Rosenberg;

National Food Institute at the Technical University of Denmark;

Trine Klein Reffstrup;

National Food Institute at the Technical University of Denmark;

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.

### 2.3 Date of QMRF update(s)

### 2.4 QMRF update(s)

2.5 Model developer(s) and contact details

Jay Russel Niemelä;

National Food Institute at the Technical University of Denmark;

Eva Bay Wedebye;

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 in 2003 from the Cancer Potency Database (CPDB 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

Rodent (rat and mouse).

### 3.2 Endpoint

QMRF 4. Human Health Effects

QMRF 4.12. Carcinogenicity

### 3.3 Comment on endpoint

Data compiled from the Cancer Potency Database (CPDB 1999) were used to train this model. The CPDB is a unique and widely used international resource currently holding the results of 6540 chronic, long-term animal cancer tests on 1547 chemicals. The CPDB provides easy access to the bioassay literature, with qualitative and quantitative analyses of both positive and negative experiments that have been published over the past 50 years in the general literature through 2001 and by the National Cancer Institute/National Toxicology Program (NCI/NTP) through 2004. The CPDB standardizes the diverse literature of cancer bioassays that vary widely in protocol, histopathological examination and nomenclature, and in the publishing author's choices of what information to provide in their papers. Results are reported in the CPDB for tests in rats, mice, hamsters, dogs, and nonhuman primates (CPDB 1999).

From the CPDB data for substances with organ specific tumour information from rodent (rat and/or mouse) *in vivo* experiments were compiled. Chemicals causing tumours exclusively in the liver of rodents were defined as positives. Chemicals that caused cancer not only in the liver but also in others of the investigated organs were defined as negatives. Due to differences in liver metabolism, rat and mice are more sensitive to certain mechanisms associated with cell proliferation compared to humans. This model is intended to identify substances which are acting by these mechanisms and therefore may possibly not give the same effects in humans.

### 3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

### 3.5 Dependent variable

Carcinogenicity in rodent liver (exclusively), positive or negative.

### 3.6 Experimental protocol

For data to be included in the CPDB, experiments should meet a set of standard inclusion criteria. These inclusion rules can be seen online at: http://toxnet.nlm.nih.gov/cpdb/methods.html#sources. These inclusion criteria for the CPDB were designed to identify reasonably thorough, chronic, long-term tests of single chemical agents (whether positive or negative). The two sources of data are the bioassays of the NCI/NTP and the general published literature. For NCI/NTP bioassay data the standard protocol from the 1970s is described in Sontag *et al.*(1976) and recommends that tests be conducted in two species of rodents (rats and mice) with both sexes tested individually at the maximally tolerated dose (MTD) and half that dose, using a control group and a vehicle control where appropriate. In the early 1990s the standard number of dose groups was increased to 3, and the standard range of doses tested was 4-10 folds.

In order for experiments from the general literature to be included in the database a set of standard inclusion criteria should be met.

For the data in CPDB the following should be noted: For any single chemical, the number of experiments in the database may vary. Some chemicals have only one test in one sex of one species, while others have multiple tests including both sexes of a few strains of rats and mice, possibly using quite different protocols.

### 3.7 Endpoint data quality and variability

Data for the training set originated from multiple sources and therefore some degree of variability is expected. The inclusion rules (see 3.6) for CPDB reduces some of this variability in data.

## 4. Defining the algorithm

### 4.1 Type of model

This is a categorical (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 categorical (Q)SAR model made by use of parametric discriminant analysis to create a linear discriminant function (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

The initial descriptor set is manually chosen by the model developer from the total set of built-in descriptors. Furthermore, the set of descriptors applied in the modelling by the program is on top of this selection determined by thresholds for descriptor variance and number of nonzero values likewise defined by the model developer.

53 descriptors were selected from the initial pool of descriptors by the system and used to build the model.

### 4.5 Algorithm and descriptor generation

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 cross-validation procedures (see 6.).

For this model, the linear method was used.

4.6 Software name and version for descriptor generation

SciQSAR version 3.1.00.

4.7 Descriptors/chemicals ratio

In this model 53 descriptors were used. The training set consists of 320 compounds. The descriptor/chemical ratio is 1:6.0 (53:320).

## 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

The Danish QSAR group has applied a stricter definition of applicability domain for its SciQSAR models. In addition to the applicability domain definition made by SciQSAR a second criterion has been applied for predictions generated from (Q)SAR models with a binary endpoint. For each prediction SciQSAR calculates the probability (p) for the test compound's membership in one of the two outcome classes (positive or negative). The probability of membership in a class is a measure of how well training set knowledge is able to discriminate a positive prediction from a negative prediction within the nearest space of the subject compound-vector. The probability of membership value is also a measure of the degree of confidence of a prediction. The Danish QSAR group uses this probability for a prediction to further define the model's applicability domain. Only positive predictions with a probability equal to or greater than 0.7 and negative predictions with a probability equal to or less than 0.3 are accepted. Positive predictions with a probability between 0.5 and 0.7 as well as negative predictions with a probability between 0.3 and 0.5 are considered outside the model's applicability domain. When these predictions are wed out the accuracy of the model in general increases at the expense of reduced model coverage. Furthermore, as SciQSAR does not define a structural domain, only predictions which were within either Leadscope structural domain (defined as at least one training set chemical within a Tanimoto distance of 0.7) or CASE Ultra structural domain (no unknown fragments for negatives and maximum 1 unknown fragment for positives) were defined as being inside the SciQSAR applicability domain.

### 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 positive predictions with probability equal to or greater than 0.7 and negative predictions with probability equal to or less than 0.3 were accepted.

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

SciQSAR version 3.1.00.

### 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

320 compounds are in the training set: 109 positives and 211 negatives.

6.6 Pre-processing of data before modelling

From (CPDB 1999) substances with organ specific tumour information from rodent in vivo experiments were compiled. For 626 structure information in the form of SMILES were available. Of these, 109 chemicals exclusively caused tumours in the liver and these were defined as positives. Chemicals causing tumours in the liver as well as in other organs were defined as 'negative'. To balance the model so that the ratio of negatives to positives was not too high a random selection was made among them giving a total of 211 negatives. The total number of substances in the training set was therefore 320. The remaining 306 negatives were originally used in an external validation which has not yet been repeated for this new version of the model (originally in MC4PC: 182 of the negatives were within AD giving a specificity of 86.3%).

6.7 Statistics for goodness-of-fit

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

- Sensitivity (true positives / (true positives + false negatives)): 52.3%
- Specificity (true negatives / (true negatives + false positives)): 91.0%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 77.8%

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

Not performed.

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

SciQSAR's own internal 10-fold cross-validation (10*10% out) procedure was used for predictions within the applicability domain as defined by SciQSAR (i.e. the first criterion described in 5.1). As the probability domain was not applied (i.e. the second criterion described in 5.2) the accuracy of the predictions when applying this domain can be expected to be higher than reflected in these cross-validation results. This gave the following Cooper's statistics:

- Sensitivity (true positives / (true positives + false negatives)): 38.5%
- Specificity (true negatives / (true negatives + false positives)): 84.8%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 69.1%

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 validation 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 not 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 applied to predict if a chemical has the potential to cause liver tumours exclusively in rodents (rat and/or mouse). A negative result does not mean that the predicted chemical is not a carcinogen but that the chemical is not exclusively causing tumours in the rodent liver.

### 9.2 Bibliography

CPDB (1999) The Carcinogenic Potency Database (CPDB) [online]. By Lois Swirsky Gold. Last updated September 2011. Available at http://toxnet.nlm.nih.gov/cpdb/

Sontag, J.M., Page, N.P. and Saffiotti, U. (1976) Guidelines for carcinogen bioassay in small rodents. DHHS Publication (National Institutes of Health) 76-801, National Cancer Institute, Bethesda, Maryland.

### 9.3 Supporting information