SciMatics SciQSAR model for Syrian Hamster Embryo (SHE) Cell Transformation *in vitro*

1. QSAR identifier

1.1 QSAR identifier (title)

SciMatics SciQSAR model for Syrian Hamster Embryo (SHE) Cell Transformation *in vitro*, Danish QSAR Group at DTU Food.

1.2 Other related models

MultiCASE CASE Ultra model for Syrian Hamster Embryo (SHE) Cell Transformation *in vitro*, Danish QSAR Group at DTU Food.

Leadscope Enterprise model for Syrian Hamster Embryo (SHE) Cell Transformation *in vitro*, 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

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

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 data were compiled from Isfort *et al.* (1996), Kerckaert *et al.* (1996), Gibson *et al.* (1997), Kerckaert *et al.* (1998) and Park *et al.* (2002) (see 9.2). In addition, 39 physiological chemicals from Grant *et al.* (2000), which are assumed to have a low probability of activity in this assay, were added as negatives to balance the training set against overrepresentation of positive test results. 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

Syrian hamster (embryo cells).

3.2 Endpoint

QMRF 4. Human Health Effects

QMRF 4.10. Mutagenicity

3.3 Comment on endpoint

Syrian hamster embryo (SHE) cells are genetically stable, diploid, metabolically and p53-competent primary cells, that have the ability to biotransform a wide range of xenobiotics. SHE cells have been used since the mid-1960ies to study the transforming ability of a variety of chemicals and physical agents. Exposure of the SHE cells to mutagenic chemicals results in an increase of morphologically transformed (MT) colonies, which are characterised by disorganised growth patterns and mimicking an early stage in carcinogenesis. It has been shown that SHE cells can be morphologically transformed by treatment with both genotoxic and non-genotoxic carcinogens. The exact molecular mechanisms involved in cell transformations are only partially understood. The transformation of these primary, diploid SHE cells is considered a model of the multistep process of carcinogenesis, as it appears to follow a staged process. The transformants are thought to be stem cells with blockages in their differentiation pathways. The transformed phenotype is characterized as a neoplastic progression-predisposing state that permits further steps toward acquisition of immortality, tumourigenicity and, finally, full malignancy. Upon further passages *in vitro*, transformed colonies clonally isolated from treated cultures, frequently generate cells with an infinite cellular lifespan or an ability to form tumours in syngenic (i.e. genetically identical) hosts. Untransformed clones on the other hand become senescent. The cell transformation results from structural alterations and changes in the expression of genes involved in cell cycle control, genomic stability, proliferation and differentiation. Genetic changes affecting these processes may result from direct genotoxic mechanisms or from non-genotoxic disturbance of gene expression and genomic stability through hyper- or hypomethylation of DNA, histone modifications and nucleosomal remodelling. In morphologically transformed SHE cell lines, cell cycle checkpoint control (G2) is often compromised. (OECD Guideline Draft 2013)

The SHE cell transformation assay is a short-term *in vitro* assay that predicts rodent carcinogenicity of chemicals by detecting the earliest identifiable stage in carcinogenesis; morphological transformation (MT) of cell colonies induced by chemicals. In contrast to most other short-term *in vitro* assays, both genotoxic and non-genotoxic carcinogens are identified.
A high correlation between results from the SHE cell transformation assay and rodent carcinogenicity data has been shown (Isfort *et al.* 1996). It was also shown that this assay was better at predicting rodent carcinogens compared to the Salmonella mutagenicity test (i.e. the Ames test). This is probably due to the fact the Salmonella mutation test only identifies genotoxic carcinogens.

3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

3.5 Dependent variable

Syrian Hamster Embryo (SHE) cell transformation *in vitro*, positive or negative.

3.6 Experimental protocol

The experimental protocol for the SHE cell transformation *in vitro* assay is described in an OECD Guideline Draft (2013). Briefly, SHE cells are seeded at clonal density onto a feeder layer of x-ray-irradiated SHE cells in culture conditions allowing for the development of colonies. After plating, the cells are exposed to the test substance for 7 days. Then cells are washed, fixed and stained, and the colonies are scored for their morphological phenotype by stereomicroscopy. Cytotoxicity is evaluated by inhibition of cloning efficiency and reduction in size or density of the colonies. The number of morphologically transformed (MT) colonies relative to the total number of scorable colonies is calculated for each concentration tested. The frequency of MT colonies relative to total number of colonies in the substance-treated test groups is compared to the frequency of MT colonies in the solvent-treated control group.

3.7 Endpoint data quality and variability

Data for the training set originates from multiple sources and therefore some degree of variability in data is expected. Difference in the use of physiological pH (approx. 7.4) or reduced pH (6.7) in the experimental protocol has been shown not to affect the final results significantly (OECD Guideline Draft 2013) and therefore data from assays using either of the pH values are useful. The assay has been thoroughly validated since it was first introduced in the 1960ies and data are in general of high quality (Gibson *et al.* 1997, Kerckaert *et al.* 1998).

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.

65 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 quadratic 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 65 descriptors were used. The training set consists of 352 compounds. The descriptor/chemical ratio is 1:5.4 (65:352).

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

352 compounds are in the training set: 176 positives and 176 negatives.

6.6 Pre-processing of data 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

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)): 96.0%
- Specificity (true negatives / (true negatives + false positives)): 85.8%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 90.9%

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)): 76.1%
- Specificity (true negatives / (true negatives + false positives)): 66.5%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 71.3%

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 applied to predict a result for the Syrian Hamster Embryo (SHE) cell transformation *in vitro* assay.

9.2 Bibliography

Gibson, D.P., Brauninger, R., Shaffi, H.S., Kerckaert, G.A., LeBoeuf, R.A., Isfort, R.J., and Aardema, M.J. (1997) Induction of micronuclei in Syrian hamster embryo cells: comparison to results in the SHE cell transformation assay for national toxicology program test chemicals. *Mutation Research*, 392, 61-90.

Grant, S.G., Zhang, Y.P., Klopman, G., and Rosenkranz, H.S. (2000) Modeling the mouse lymphoma forward mutational assay: The Gene-Tox program database. *Mutation Research* 465, 201-229.

Isfort, R.J., Kerckaert, G.A., and LeBoeuf, R.A. (1996) Comparison of the standard and reduced pH Syrian Hamster Embryo (SHE) cell in vitro transformation assays in predicting the carcinogenic potential of chemical. *Mutation Research*, 356, 11-63.

Kerckaert, G.A., Isfort, R.J., Carr, G.J., Aardema, M.J., and LeBoeuf, R.A. (1996) A comprehensive protocol for conducting the Syrian hamster cell transformation assay at pH 6.70. *Mutation Research*, 356, 65-84.

Kerckaert, G.A., LeBoeuf, R.A., and Isfort, R.J. (1998) Assessing the Predictiveness of the Syrian Hamster Embryo Cell Transformation Assay for Determining the Rodent Carcinogenic Potential of Single Ring Aromatic/Nitroaromatic Amine Compounds. *Toxicological Sciences*, 189-197.

Park, J., Kamendulis, L.M., and Klaunig, J.E. (2002) Mechanisms of 2-Butoxyethanol Carcinogenicity: Studies on Syrian Hamster Embryo (SHE) Cell Transformation. *Toxicological Sciences*, 68, 43-50.

OECD Guideline Draft (2013) *In Vitro* Carcinogenicity: Syrian Hamster Embryo (SHE) Cell Transformation Assay. Available online at: http://www.oecd.org/env/ehs/testing/CTA%20TG_Feb2013.pdf

9.3 Supporting information