SciMatics SciQSAR model for sex-linked recessive lethal test in *Drosophila melanogaster in vivo*

1. QSAR identifier

1.1 QSAR identifier (title)

SciMatics SciQSAR model for sex-linked recessive lethal test in *Drosophila melanogaster in vivo*, Danish QSAR Group at DTU Food.

1.2 Other related models

MultiCASE CASE Ultra model for sex-linked recessive lethal test in *Drosophila melanogaster in vivo*, Danish QSAR Group at DTU Food.

Leadscope Enterprise model for sex-linked recessive lethal test in *Drosophila melanogaster 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

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 consists of Gene-Tox data compiled by Lee and co-workers (1983) plus data from EMIC (US Environmental Mutagen Information Center), IARC (International Agency for Research on Cancer) and NTP (US National Toxicology Program) etc. 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

Drosophila melanogaster (germ cells).

3.2 Endpoint

QMRF 4.10. Mutagenicity

EC B.20. Sex-Linked recessive Lethal Test in Drosophila Melanogaster

3.3 Comment on endpoint

Drosophila melanogaster, generally known as the common fruit fly, is the test organism most often used to detect transmissible mutations in germ cells of eukaryotes. The short generation times of 10 days, low cost of culture media, and a large number of well-defined genetic tests for mutations are the principal advantages of using *D. melanogaster* as compared with using the mouse or rat, which are the only other well-developed systems for testing mutagenesis in germ cells. The sex-linked recessive lethal (SLRL) test using *D. melanogaster* detects the occurrence of chromosome aberrations and mutations, both point mutations and small deletions, in different stages of germ cell development of the insect. It is capable of detecting both direct-acting mutagens and promutagens, i.e. compounds that require activation to become mutagenic. The test is therefore not specific for any one class of chemicals.

This SLRL test is a forward mutation assay capable of screening for mutations in around 800 loci on the X-chromosome. This represents about 80 % of all X-chromosomal loci. The X-chromosome represents approximately one-fifth of the entire genome. Therefore the test gives a good estimate of mutation frequency in the entire genome. A lethal mutation is a change in the genome which, when expressed, causes death to the carrier. A recessive mutation is a change in the genome which is expressed in the homozygous or hemizygous condition. Sex-linked genes are present on the sex (X or Y) chromosomes. Mutations in the X-chromosome of *D. melanogaster* are phenotypically expressed in males carrying the mutant gene. When the mutation is lethal in the hemizygous condition, its presence is inferred from the absence of one class of male offspring out of the two that are normally produced by a heterozygous female.
Positive results from the SLRL-test in *D. melanogaster* indicate that a substance induces mutations in the germ line of the insect. Negative results indicate that, under the test conditions, the test substance does not induce mutations in the germ line of the insect. A high correlation between mutagenesis in the SLRL test and carcinogenesis has been found. (Lee *et al.*1983)

3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

3.5 Dependent variable

Sex-linked recessive lethal (SLRL) test in *D. melanogaster in vivo,* positive or negative.

3.6 Experimental protocol

Data has been generated using similar experimental protocols similar to that described in OECD guideline 477 (1984). Briefly, 3 to 5 days old wild-type males are treated with the test substance and mated individually to an excess of virgin females. The females are replaced with fresh virgins every 2 to 3 days to cover the entire germ cell cycle. The offspring of these females are scored for lethal effects corresponding to the effects on mature sperm, mid or late-stage spermatids, early spermatids, spermatocytes and spermatogonia at the time of treatment.
Heterozygous F1 females from the above crosses are mated individually with their brothers. In the F2 generation each separate cross is scored for the absence of phenotypically wild-type males. If a culture appears to have arisen from a F1 female carrying a lethal mutation in the parental X-chromosome (i.e. no males with the treated chromosome are observed), a daughter of that female with the same genotype should be tested to ascertain whether the lethality is repeated at the next generation.
The assay has a low sensitivity for genotoxins other than direct-acting agents and simple promutagens, but a very high specificity.
From Lee *et al.* (1983): A positive mutagenic response was the demonstration of a difference between the mutation frequencies in a treated and a concurrent control group that was statistically significant at the 5% level. If the investigation of a compound did not have a concurrent control, but the mutation frequency was significantly higher than 0.5%, the compound was accepted as a mutagen. The frequency of 0.5% was selected because the spontaneous frequencies for the standard strains range from 0.1 to 0.3%.
A test was considered negative if both of the following criteria were met: (1) The observed increase in the treated group over control is less than 0.2% and sample size is large enough so that an observed increase of 0.2% would be statistically significant. (2) The second criterion takes into account the possible differential response of different post-meiotic germ cell stages to direct and indirect mutagens as revealed by a mating pattern analysis. If none of the mating's analysed gives a positive result, the data must indicate a statistically negative response in at least 2 mating's, preferably representing mature sperm and early spermatids.

3.7 Endpoint data quality and variability

The data set from Lee *et al.* (1983) consists of data compiled from publications in the EMIC (Environmental Mutagen Information Center) file for the period 1968 to 1978. The publications were reviewed thoroughly and only data that meet the criteria defined by the Working Group were included. As training set data originates from different sources a certain degree of variability in the experimental protocols (strain, mating protocols, route of administration etc.) and data is expected although this variability has been diminished by the criteria for inclusion.

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.

70 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 kernel 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 70 descriptors were used. The training set consists of 370 compounds. The descriptor/chemical ratio is 1:5.3 (70:370).

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

370 compounds are in the training set: 186 positives and 184 negatives.

6.6 Pre-processing of data before modelling

From the original data set from Lee *et al.* (1983) only compounds for which SMILES codes could be found and that had a structure acceptable for the commercial software 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.

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)): 100%
- Specificity (true negatives / (true negatives + false positives)): 98.9%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 99.5%

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)): 74.2%
- Specificity (true negatives / (true negatives + false positives)): 78.3%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 76.2%

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 results for the sex-linked recessive lethal (SLRL) *in vivo* test in *Drosophila melanogaster*.

9.2 Bibliography

Lee, W.R., Abrahamson, S., Valencia, R., von Halle, E.S., Würgler, F.E., and Zimmering, S. (1983) The sex-linked recessive lethal test for mutagenesis in *Drosophila melanogaster*. A report of the U.S. Environmental Protection Agency (EPA) Gene-Tox Program. *Mutation research,* 123, 183-279.

OECD guideline 477 (1984) Genetic Toxicology: Sex-linked Recessive Lethal Test in *Drosophila melanogaster*. OECD guidelines for testing of chemicals. Organisation for Economic Cooperation and Development; Paris, France. Available online at: http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788.

9.3 Supporting information