MultiCASE CASE Ultra model for sex-linked recessive lethal test in *Drosophila melanogaster in vivo*

1. QSAR identifier

1.1 QSAR identifier (title)

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

1.2 Other related models

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

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

1.3. Software coding the model

MultiCASE CASE Ultra 1.4.6.6 64-bit.

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

Klopman, G. (1992) MULTICASE 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Relat.,* 11*,* 176 - 184.

Chakravarti, S.K., Saiakhov, R.D., and Klopman, G. (2012) Optimizing Predictive Performance of CASE Ultra Expert System Models Using the Applicability Domains of Individual Toxicity Alerts. *J. Chem. Inf. Model*., 52, 2609 −2618.

Saiakhov, R.D., Chakravarti, S.K., and Klopman, G. (2013) Effectiveness of CASE Ultra Expert System in Evaluating Adverse Effects of Drugs. *Mol. Inf*., 32, 87 – 97.

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

CASE units, 45 for positives and 10 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

A categorical (Q)SAR model based on structural fragments and calculated molecular descriptors.

4.2 Explicit algorithm

This is a categorical (Q)SAR model composed of multiple local (Q)SARs made by use of stepwise regression. The specific implementation is proprietary within the MultiCASE CASE Ultra software.

4.3 Descriptors in the model

Fragment descriptors,

Distance descriptors,

Physical descriptors,

Electronic descriptors,

Quantum mechanical descriptors

4.4 Descriptor selection

Automated hierarchical selection (see 4.5).

4.5 Algorithm and descriptor generation

MultiCASE CASE Ultra is an artificial intelligence (AI) based computer program with the ability to learn from existing data and is the successor to the program MultiCASE MC4PC. The system can handle large and diverse sets of chemical structures to produce so-called global (Q)SAR models, which are in reality series of local (Q)SAR models. Upon prediction of a query structure by a given model one or more of these local models, as well as global relationships if these are identified, can be involved if relevant for the query structure. The CASE Ultra algorithm is mainly built on the MCASE methodology (Klopman 1992) and was released in a first version in 2011 (Chakravarti *et al.* 2012, Saiakhov *et al.* 2013).

CASE Ultra is a fragment-based statistical model system. The methodology involves breaking down the structures of the training set into all possible fragments from 2 to 10 heavy (non-hydrogen) atoms in length. The fragment generation procedure produces simple linear chains of varying lengths and branched fragments as well as complex substructures generated by combining the simple fragments.

A structural fragment is considered as a positive alert if it has a statistical significant association with chemicals in the active category. It is considered a deactivating alert if it has a statistically significant relation with the inactive category.

Once final lists of positive and deactivating alerts are identified, CASE Ultra attempts to build local (Q)SARs for each alert in order to explain the variation in activity within the training set chemicals covered by that alert. The program calculates multiple molecular descriptors from the chemical structure such as molecular orbital energies and two-dimensional distance descriptors. A stepwise regression method is used to build the local (Q)SARs based on these molecular descriptors. For each step a new descriptor (modulator) is added if the addition is statistically significant and increases the cross-validated R2 (the internal performance) of the model. The number of descriptors in each local model is never allowed to exceed one fifth of the number of training set chemicals covered by that alert. If the final regression model for the alert does not satisfy certain criteria ($R2 \geq 0.6$ and $Q2 \geq 0.5$) it is rejected. Therefore, not all alerts will necessarily have a local (Q)SAR.

The collection of positive and deactivating alerts with or without a local (Q)SAR constitutes a global (Q)SAR model for a particular endpoint and can be used for predicting the activity of a test chemical.

More detailed information about the algorithm can be found in Chakravarti *et al.* (2012), Saiakhov *et al.* (2013).

4.6 Software name and version for descriptor generation

MultiCASE CASE Ultra 1.4.6.6 64-bit.

4.7 Descriptors/chemicals ratio

The program primarily uses fragment descriptors specific to a group of structurally related chemicals from the training set. Therefore estimation of the number of descriptors used in a specific model, which is a collection of local models as explained under 4.5, may be difficult. In general, we estimate that the model uses an order of magnitude less descriptors than there are observations. The number of descriptors in each local (Q)SAR model is never allowed to exceed one fifth of the number of training set chemicals covered by that alert (Saiakhov *et al.* 2013).

It should be noted that due to CASE Ultra's complex decision making scheme overfitting is rare compared to simpler linear models. Warnings are issued in case of statistically insufficient overall number of observations to produce a model, which is not the case in the present model.

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 CASE Ultra and the in-house further refinement algorithm on the output from CASE Ultra to reach the final applicability domain call.

1. CASE Ultra

CASE Ultra recognizes unknown structural fragments in test chemicals that are not found in the training data and lists these in the output for a prediction. Fragments this way impose a type of global applicability domain for the overall model. The presence of more than three unknown structural fragments in the test chemical results in an 'out of domain' call in the program. (Chakravarti *et al.*2012, Saiakhov *et al.*2013).

For each structural alert, CASE Ultra uses the concept of so-called domain adherences and statistical significance.

The domain adherence for an alert in a query chemical depends on the similarity of the chemical space around the alert in the query chemical compared to the chemical space (in terms of frequencies of occurrences of statistically relevant fragments) of the training set chemicals used to derive the alert. The domain adherence value (between zero and one) is the ratio of the sum of the squared frequency of occurrence values of the subset of the fragments that are present in the test chemical and sum of the squared frequency of occurrence of all the fragments that constitute the domain of the alert in question. The more fragments of the domain of the alert in the test chemical the closer the domain adherence value is to 1. The value will never be zero as the alert itself is part of the alerts domain.

Furthermore, all alerts come with a measure of its statistical significance, and this depends on the number of chemicals in the training set which contained the alert and the prevalence within these of actives and inactives. (Chakravarti *et al.*2012).

2. In-house refinement algorithm to reach the final applicability domain call

The Danish QSAR group has applied a stricter definition of applicability domain for its MultiCASE CASE Ultra models.

An optimization procedure based on preliminary cross-validation is applied to further restrict the applicability domain for the whole model based on non-linear requirements for domain adherence and statistical significance, giving the following primary thresholds:

Domain adherence = 0.76 and significance = 70%.

Any positive prediction is required to contain at least one valid positive alert, namely an alert with statistical significance and domain adherence exceeding thresholds defined for the specific model.

The positive predictions for chemicals which only contain invalid positive alerts are considered 'out of domain' (in CASE Ultra these chemicals are predicted to be inactive).

Furthermore, only query chemicals with no unknown structural fragments are considered within the applicability domain, except for chemicals predicted 'positive', where one unknown fragment is accepted. Also no significant positive alerts are accepted for an inactive prediction.

5.2 Method used to assess the applicability domain

The applicability domain is assessed in terms of the output from CASE Ultra with the Danish QSAR group's further refinement algorithm on top as described in 5.1.

Because of the complexity of the system (see 5.1), the assessment of whether a test chemical is within the applicability domain of the model requires predicting the chemical with the specific model, and the application of the Danish QSAR group model-specific thresholds for domain adherence and significance.

This applicability domain was also applied when determining the results from the cross-validations (6.9).

5.3 Software name and version for applicability domain assessment

MultiCASE CASE Ultra 1.4.6.6 64-bit.

5.4 Limits of applicability

All structures are run through the DataKurator feature within CASE Ultra to check for compatibility with the program. Furthermore, 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 CASE Ultra. 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

Not performed.

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

Not performed. (It is not a preferred measurement for evaluating large models).

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

A five times two-fold 50 % cross-validation was performed. This was done by randomly removing 50% of the full training set used to make the “mother model”, thereby splitting the full training set into two subsets A and B, each containing the same ratio of positives to negatives as the full training set. A new model (validation sub-model) was created on subset A without using any information from the “mother model” (regarding e.g. descriptor selection etc.). The validation sub-model was applied to predict subset B (within the CASE Ultra applicability domain for the validation sub-model and the in-house further refinement algorithm for the full model). Likewise, a validation sub-model was made on subset B and this model was used to predict subset A (within the CASE Ultra applicability domain for the validation sub-model and the in-house further refinement algorithm for the full model). This procedure was repeated five times.

Predictions within the defined applicability domain for the ten validation sub-models were pooled and Cooper’s statistics calculated. This gave the following results for the 45.8% (848/(5*370)) of the predictions which were within the applicability domain:

- Sensitivity (true positives / (true positives + false negatives)): 77.1%
- Specificity (true negatives / (true negatives + false positives)): 89.3%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 83.0%

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 model identifies statistically relevant substructures (i.e. alerts) and for each set of molecules containing a specific alert it further identifies additional parameters found to modulate the alert (e.g. logP and molecular orbital energies, etc.). Many predictions 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 identified structural features and molecular descriptors may provide 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*.

A version of this model made in MC4PC, the predecessor to CASE Ultra, was applied in the creation of the Advisory list for self-classification of dangerous substances, published by the Danish Environmental Protection Agency (Niemelä *et al.* 2010).

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.

Niemelä, J.R., Wedebye, E.B., Nikolov, N.G., Jensen, G.E., Ringsted, T., Ingerslev, F., Tyle, H., and Ihlemann, C. (2010) The Advisory list for self-classification of dangerous substances. Danish Environmental Protection Agency, Environmental Project No. 1322, 2010; www.mst.dk. Available on: http://www.mst.dk/English/Chemicals/assessment_of_chemicals/The_advisory_list_for_selfclassification/

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