MultiCASE CASE Ultra model for carcinogenicity exclusively in rodent liver *in vivo*

1. QSAR identifier

1.1 QSAR identifier (title)

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

1.2 Other related models

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

SciMatics SciQSAR model for carcinogenicity exclusively in rodent liver *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

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

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 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

CASE unit, 45 for positives and 10 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

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 ≥ 0.6 and Q2 ≥ 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 = 0.68 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

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

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 38.3% (612/(5*320)) of the predictions which were within the applicability domain:

- Sensitivity (true positives / (true positives + false negatives)): 34.2%
- Specificity (true negatives / (true negatives + false positives)): 87.4%
- 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

Yes

7.2 Available information for the external validation set

CAS

SMILES

7.3 Data for each descriptor variable for the external validation set

No

7.4 Data for the dependent variable for the external validation set

All

7.5 Other information about the validation set

The 306 negatives which were not included in the training set were originally (in 2003) used for an external validation of the MC4PC model. This external validation has not been repeated (yet) in this new CASE Ultra version of the model.

7.6 Experimental design of test set

See 3.6.

7.7 Predictivity – Statistics obtained by external validation

Of the 306 negatives, which had been randomly removed from the CPDB data set prior to modelling, 182 (59%) were predicted in the MC4PC 2003 model version within the defined AD. 157 were correctly predicted negative giving a specificity of 86.3%.

An external validation for sensitivity using the 2003 version of CPDB was not feasible as only one additional exclusive hepatocarcinogen had been added during the course of four years.

7.8 Predictivity – Assessment of the external validation set

7.9 Comments on the external validation of the model

Only the specificity for the model could be estimated as no positive test chemicals were available to estimate the models sensitivity.

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 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