MultiCASE CASE Ultra commercial model RC2_AF4 for Rodent carcinogenicity in female mouse in vivo, Danish QSAR Group at DTU Food

1. QSAR identifier

1.1 QSAR identifier (title)

MultiCASE CASE Ultra commercial model RC2_AF4 for Rodent carcinogenicity in female mouse *in vivo*, Danish QSAR Group at DTU Food.

1.2 Other related models

Leadscope Enterprise commercial model for Rodent carcinogenicity in female mouse *in vivo*, Danish QSAR Group at DTU Food.

MultiCASE CASE Ultra commercial model RC1_AF1 for Rodent carcinogenicity in male rat *in vivo*, Danish QSAR Group at DTU Food.

MultiCASE CASE Ultra commercial model RC1_AF2 for Rodent carcinogenicity in female rat *in vivo*, Danish QSAR Group at DTU Food.

MultiCASE CASE Ultra commercial model RC2_AF3 for Rodent carcinogenicity in male mouse *in vivo*, Danish QSAR Group at DTU Food.

MultiCASE CASE Ultra commercial model RC1_AFV for Rodent carcinogenicity in rat *in vivo*, Danish QSAR Group at DTU Food.

MultiCASE CASE Ultra commercial model RC2_AFW for Rodent carcinogenicity in mouse *in vivo*, Danish QSAR Group at DTU Food.

MultiCASE CASE Ultra commercial model RC_AFU for Rodent carcinogenicity *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

Trine Klein Reffstrup;

National Food Institute at the Technical University of Denmark;

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Sine Abildgaard Rosenberg;

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

MultiCASE Inc. 23811 Chagrin Blvd, Suite 305, Beachwood, OH 44122, USA. www.multicase.com.

2.6 Date of model development and/or publication

Commercial model updated by MultiCASE/FDA (U.S. Food and Drug Administration) regularly. For the Danish QSAR predictions database the 2013 version was used.

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 consists of non-proprietary studies and is composed of data harvested from FDA approval packages and the published literature (for more details see 6.5). The model algorithm is proprietary from commercial software. The model has been created by FDA and MultiCASE based on FDA data as part of a Research Cooperation Agreement (RCA).

The training set was constructed using the NTP (U.S. National Toxicology Program) Rodent carcinogenicity database, the Lois Gold Carcinogen Potency Database, FDA/CDER (U.S. Food and Drug Administration / Center for Drug Evaluation and Research) archives, and the scientific literature. The model including the training data set is commercially available from MultiCASE Inc.

2.9 Availability of another QMRF for exactly the same model

3. Defining the endpoint

3.1 Species

Mouse, female.

3.2 Endpoint

4. Human health effects

4.12. Carcinogenicity

3.3 Comment on endpoint

The two-year rodent bioassay is considered the regulatory standard for evaluating the carcinogenic potential of a chemical and provides information on the possible health hazards likely to arise from repeated exposure for a period lasting up to the entire lifespan of the species used. The assay is usually performed in both sexes of rats or mice for a period of two years, with the chemical administered at high doses, primarily by the oral route. Extrapolation from the rodent assay to humans is based on two principal assumptions. The first assumption is that a carcinogenic response in rats/mice predicts possible carcinogenicity in humans (interspecies extrapolation). The second assumption is that carcinogenicity detected at a high dose implies carcinogenicity at low doses, although at a lower rate (dose extrapolation).

3.4 Endpoint units

MultiCASE CASE unit, 30-79 for positives, 20-29 for marginals and 10-19 for negatives.

3.5 Dependent variable

Carcinogenicity in female mice *in vivo*, positive or negative.

3.6 Experimental protocol

Female mice (50-70 animals/group) are divided randomly into one or two control groups and three treatment groups. Historically, the highest dose in the studies generally approximates the maximum tolerated dose (MTD) in the test specie. The test substance is normally administered in the feed or by oral gavage for two years. In NTP studies the most often used mouse strain is the hybrid B6C3F1 (C3H x C57B16) mouse while the CD-1 Swiss- Webster derived mouse is the predominant strain in pharmaceutical studies submitted to the FDA. Tumor findings are classified as positive if either benign and/or malignant findings are statistically significant in pair-wise comparison to concurrent controls ($p \leq 0.01$) by Fisher's Exact Test or equivalent statistical analysis. The tumor findings are adjusted for rare (with a spontaneous background incidence rate equal to or less than 1 %) and common events (Contrera *et al*. 2005).

3.7 Endpoint data quality and variability

Data in the training set originates from multiple sources and therefore some variability in the experimental procedures (e.g. strains, concentration ranges) and experimental results is expected.

In an external validation exercise (see 7.) performed, the activity scores for 1028 duplicate compounds between the training set and the external test set were compared. Depending on the endpoint, concordance in activity scores ranged from 86.7% to 90.7% (Stavitskaya *et al.*, 2013).

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 = 0.78 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

The training set is commercially available embedded in software from MultiCASE Inc.

### 6.2 Available information for the training set

No

### 6.3 Data for each descriptor variable for the training set

No

### 6.4 Data for the dependent variable for the training set

No

### 6.5 Other information about the training set

1208 compounds are in the training set: 526 positives, 621 negatives and 61 marginals.

### 6.6 Pre-processing of data before modelling

"Rodent carcinogenicity studies of compounds that have been tested in at least the male and female animals of one rodent species (i.e., 2 cells) were included. Although the MCASE database modules contained compounds tested by nonoral routes of administration (inhalation, intravenous, intramuscular, dermal, intraperitoneal, and subcutaneous), the majority of acceptable studies used an oral route of exposure (feed, gavage, or drinking water). The duration of acceptable carcinogenicity studies was limited to ≥18 months for negative (inactive) compounds. All studies with compound-related tumor findings (positive studies) were acceptable regardless of duration of treatment, with one exception. Positive (active) nonoral studies were included if tumors were induced at other than the site of application." (Matthews and Contrera, 1998). Further information can be found in (Matthews and Contrera, 1998), (Contrera *et al.* 2003) and (Contrera *et al.* 2005).

### 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 47.4% (2864/(5*1208)) of the predictions which were within the applicability domain:

- Sensitivity (true positives / (true positives + false negatives)): 38.7%
- Specificity (true negatives / (true negatives + false positives)): 87.5%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 65.5%

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

No

### 7.2 Available information for the external validation set

No

### 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

No

### 7.5 Other information about the training set

The external data set was comprised of non-proprietary data was harvested by Leadscope from FDA approval packages and the published literature. The entire set contained 2115 compounds, but 1393 of the compounds were removed as they were already part of the training set or were stereo or geometric isomers of structures already in the training set or were duplicates or perceived duplicates within the set. Therefore the final external test set consisted of 722 compounds (34-52 % active and 48-66 % inactive) (Stavitskaya *et al.*, 2013).

In the external validation only the definition of the applicability domain in CASE Ultra was used (point 1 and not point 2 in 5.1).

### 7.6 Experimental design of test set

Not available.

### 7.7 Predictivity – Statistics obtained by external validation

95 % of the 722 compounds in the external test set were in the applicability domain, i.e. 686 compounds. Information is not given about the balance between positives and negatives in this set.

Performance of the model:

Sensitivity: 67%

Specificity: 60%

(Stavitskaya *et al.*, 2013)

### 7.8 Predictivity – Assessment of the external validation set

See section 3.7 for information on concordance between experimental tests in training sets and validation sets.

7.9 Comments on the external validation of the model

The results deviate from the results from the cross-validation reported under section 6, possibly because of the different applicability domains applied. Furthermore, information is not given about how well the 686 represent the full applicability domain of the 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 if a chemical has the potential to cause cancer in female mice.

A version of this model made in MC4PC, the precursor 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

Contrera, J.F., Matthews, J. and Benz, R.D. (2003). Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regulatory Toxicology and Pharmacology*, 38, 243-259.

Contrera, J.F., MacLaughlin, P., Hall, L.H., Kier, L.B. (2005) QSAR Modeling of Carcinogenic Risk Using Discriminant Analysis and Topological Molecular Descriptors. *Current Drug Discovery Technologies*, 2, 55-67.

Matthews, J. and Contrera, J.F. (1998) A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodent using enhanced MCASE (Q)SAR-ES software. *Regulatory Toxicology and Pharmacology*, 28, 242-264.

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/

Stavitskaya, L., Kruhlak, N.L., Cross, K.P., Minnier, B.L., Bower, D.A., Chakravarti, S., Saiakhov R.D., Benz, R.D.. (2013) P217 Development of Improved In Silico Models for Predicting Rodent Carcinogenicity. Poster at the American College of Toxicology Annual Meeting 2012. Poster abstract published in *International Journal of Toxicology*, 32 (1), 72.

### 9.3 Supporting information