MultiCASE CASE Ultra model for the Bacterial Reverse Mutation Test (Ames test) in *S. typhimurium in vitro*

1. QSAR identifier

1.1 QSAR identifier (title)

MultiCASE CASE Ultra model for the Bacterial Reverse Mutation Test (Ames test) in *S. typhimurium in vitro*, Danish QSAR Group at DTU Food.

1.2 Other related models

Leadscope Enterprise model for the Bacterial Reverse Mutation Test (Ames test) in *S. typhimurium in vitro*, Danish QSAR Group at DTU Food.

SciMatics SciQSAR model for the Bacterial Reverse Mutation Test (Ames test) in *S. typhimurium in vitro*, 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

Gunde Egeskov Jensen;

National Food Institute at the Technical University of Denmark;

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 kindly provided by Kazius *et al.* (2005). 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

Salmonella typhimurium (multiple strains).

3.2 Endpoint

QMRF 4.10. Mutagenicity

OECD 471 Bacterial Reverse Mutation Test

3.3 Comment on endpoint

The bacterial reversed mutation *in vitro* assay using *Salmonella typhimurium* is also referred to as the Ames test. The test is used to evaluate compounds mutagenic properties as it detects point mutations, which involve substitution, addition or deletion of one or a few DNA base pairs. The test uses amino acid-dependent strains of *S. typhimurium*. These strains contain a mutation that makes them unable to synthesize the amino acid histidine. Therefore, in the absence of an external histidine source, the bacteria cannot grow and form colonies. Colony growth is resumed if a reversion of the mutation occurs, allowing the production of histidine to be resumed. Spontaneous reversions occur within each of the strains. If a compound cause an increase in the number of revertant colonies relative to the background level it is said to be positive in the Ames test and therefore assigned mutagenic. Different strains of *S. typhimurium* exist and these have several features that make them more sensitive for the detection of mutations, including responsive DNA sequences at the reversion sites, increased cell permeability to large molecules and elimination of DNA repair systems or enhancement of error-prone DNA repair processes. The specificity of the test strains can provide useful information on the types of point mutations that are induced such as frameshift mutations or base-pair mutations.

Point mutations are the cause of many human genetic diseases and there is substantial evidence that point mutations in oncogenes and tumour suppressor genes of somatic cells are involved in tumour formation in humans and experimental animals. The bacterial reverse mutation test is rapid, inexpensive and relatively easy to perform and for these reasons it has become a useful tool as an initial screen for potential *in vivo* genotoxic activity, and is present the most extensively used *in vitro* short-term test in the screening for point mutation-inducing activity. The test utilises prokaryotic cells, which differ from mammalian cells in such factors as uptake, metabolism, chromosome structure and DNA repair processes. Tests conducted *in vitro* generally require the use of an exogenous source of metabolic activation. *In vitro* metabolic activation systems cannot mimic entirely the mammalian *in vivo* conditions. The test therefore does not provide direct information on the mutagenic and carcinogenic potency of a substance in mammals. It has though been demonstrated that many chemicals that are positive in this test also exhibit mutagenic activity in other tests. For certain classes of chemicals, for example highly bactericidal compounds (e.g. certain antibiotics) and those which are thought (or known) to interfere specifically with the mammalian cell replication system, this test may not be appropriate. Also, there are carcinogens that are not detected by this test because they act through other, non-genotoxic mechanisms or mechanisms absent in bacterial cells.

All chemicals in the training set The categorization of each compound as either a mutagen or a nonmutagen, which was based on the available, occasionally conflicting, Ames test results is described under 3.6. (Kazius *et al.* 2005).

The data used to train this model were compiled by Kazius and co-workers (2005) from multiple sources. All the structures in the data set have experimental results in one or more of the following *S. typhimurium* tester strains: TA98, TA100, TA1535 and either TA1537 or TA97. Strains TA102 and TA1538 were also applied in cases where results of other strains are equivocal or difficult to interpret. The inclusion criteria for the data as well as the categorization of chemicals in to Ames mutagens or non-mutagens is described in Kazius *et al.* (2005): "…Ames tests were only considered if they were performed with the standard plate method or the preincubation method, either with or without a metabolic activation mixture. Second, this study required the categorization of each compound as either a mutagen or a nonmutagen, which was based on the available, occasionally conflicting, Ames test results determined in different laboratories. In this study, a compound was categorized as a mutagen if at least one Ames test result was positive. Consequently, a false positive Ames test result will erroneously rendering a compound mutagenic, irrespective of the number of negative results. In general, the categorization of a compound as nonmutagenic is sufficiently reliable if at least four Ames tests, performed with different strains, give reproducible negative results. In this study, to assemble a large dataset with maximal compound diversity, a compound was categorized as a nonmutagen if exclusively negative Ames test results - one or more - were reported. Further, the robustness of the above mutagenicity categorization of the CCRIS database was tested by applying the same categorization criteria to another set of Ames test results collected from the NTP (National Toxicology Program). The results obtained for approximately 1500 compounds present in both the NTP and the CCRIS databases showed contradicting categorizations in 11% of the cases. Because this error was smaller than 15%, which is the average interlaboratory reproducibility error of Ames tests, the categorization applied in this study was considered satisfactory. To further increase the consistency of the dataset, compounds whose CCRIS data showed contradicting categorizations with the NTP data were removed from the dataset. In conclusion, a dataset of 4337 compounds with corresponding molecular structures and toxicity categorizations (2401 mutagens and 1936 nonmutagens) was constructed."

3.4 Endpoint units

CASE units; 45 for positives and 10 for negatives.

3.5 Dependent variable

Mutagenic in the Bacterial Reverse Mutation Test (Ames test) in *Salmonella Typhimurium in vitro*, positive or negative.

3.6 Experimental protocol

The experimental protocol is described in OECD guideline 471 (1997). Briefly, suspensions of bacterial cells are exposed to the test substance in the presence and in the absence of an exogenous metabolic activation system. The most commonly used system is a cofactor supplemented post-mitochondrial fraction (called S9) prepared from the livers of rodents. In the plate incorporation method, these suspensions are mixed with an overlay agar and plated immediately onto minimal medium. In the preincubation method, the treatment mixture is incubated and then mixed with an overlay agar before plating onto minimal medium. For both techniques, after two or three days of incubation, revertant colonies are counted and compared to the number of spontaneous revertant colonies on solvent control plates (OECD guideline 471, 1997).

3.7 Endpoint data quality and variability

As data originates from multiple sources and consist of a combination of results from different *S. typhimurium* tester strains some degree of variability in the data is expected. Further, as described by Kazius *et al.* (2005): "The reproducibility of Ames tests is limited by the purity of the tested chemical, inconsistencies in the interpretation of dose-response curves, interference of further toxic side effects (such as cytotoxicity), variations in the methodology employed, and variations in the materials used (bacterial strains and metabolic activation mixtures). Nevertheless, the average interlaboratory reproducibility of a series of Ames test data from the National Toxicology Program (NTP) was determined to be 85%."

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 = 0.83 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

4102 compounds are in the training set: 2299 positives and 1803 negatives.

6.6 Pre-processing of data before modelling

The original data set from Kazius *et al.* (2005) consisted of 4337 molecular structures with corresponding Ames test data. Of these 235 were excluded in the pre-processing due to:

- Only structures acceptable for the commercial software could be processed
- 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 replicates had the same activity and all were removed if they had different activity

4102 chemicals went successfully through the pre-processing and were applied as training set for the model.

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 60.2% (12340/(5*4102)) of the predictions which were within the applicability domain:

- Sensitivity (true positives / (true positives + false negatives)): 86.0%
- Specificity (true negatives / (true negatives + false positives)): 86.0%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 86.0%

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

No

7.2 Available information for the external validation set

The test set is commercial.

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

The test set is commercial and consists of 3,509 compounds. These were not in any way part of model development (i.e. included in the model's training set) but had experimental results comparable to training set data.

7.6 Experimental design of test set

As the test set is commercial no explicit information about the endpoint for the test set data can be given.

7.7 Predictivity – Statistics obtained by external validation

Of the 3,509 test set compounds 1.096 compounds (31%) were within the applicability domain of the model. For the predictions within the applicability domain the following statistics were obtained:

- Sensitivity (true positives / (true positives + false negatives)): 191/(191+32)= 85.7%

- Specificity (true negatives / (true negatives + false positives)): 773/(773+100)= 88.5%

- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): (191+773)/1096= 88.0%

7.8 Predictivity – Assessment of the external validation set

7.9 Comments on the external validation of the model

The external validation of the model is in good compliance with the cross-validation results described under 6.9.

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

This model can be used to predict if a chemical is an Ames mutagen or non-mutagen according to the categorization made by Kazius and co-workers (2005).

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

Kazius, J., McGuire, R., and Burs, R. (2005) Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.,* 48, 312-320.

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 471 (1997) OECD Guidelines for the Testing of Chemicals No. 471, Bacterial Reverse Mutation Test. 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