	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **GT4_ML_UNACT (Mouse Lymphoma unactivated), version 1.9.0.8.2528.300**
	Printing Date: **Mar 27, 2024**

1.QSAR identifier

1.1.QSAR identifier (title):

GT4_ML_UNACT (Mouse Lymphoma unactivated), version 1.9.0.8.2528.300

1.2.Other related models:

This model is part of the set of CASE Ultra Genotoxicity models:
GT2_CHROM_CHL(Chromosomal aberration in CHL cell line),
GT2_CHROM_CHO(Chromosomal aberration in CHO cell line), GT3_MNT_MOUSE (Mouse Micronucleus), GT4_ML_ACT (Mouse Lymphoma activated)

1.3.Software coding the model:

CASE Ultra Version 1.9.0.8
QSAR based bioactivity and toxicity prediction software
sales@multicase.com, MultiCASE Inc, 5885 Landerbrook Dr. #210 Mayfield Heights, OH 44124 USA www.multicase.com
http://www.multicase.com/case-ultra

2.General information

2.1.Date of QMRF:

July 2nd, 2019

2.2.QMRF author(s) and contact details:

[1]Dr Roustem D Saiakhov MultiCASE Inc +1-440-565-7221 saiakhov@multicase.com www.multicase.com
[2]Mounika Girireddy MultiCASE Inc +1-440-565-7221 girireddy@multicase.com www.multicase.com

2.3.Date of QMRF update(s):

March 27, 2024

2.4.QMRF update(s):

- Added data from FDA drug labels and echemportal (ECHA)
- Rebuild and revalidated the models

2.5.Model developer(s) and contact details:

Models were constructed under a Research Collaboration Agreement between the US Food and Drug Administration's Center for Drug Evaluation and Research, and MultiCASE Inc. +1-440-565-7221 saiakhov@multicase.com www.multicase.com

2.6.Date of model development and/or publication:

The original MC4PC model was developed in 2005; Published in 2005; Last update in 2014

2.7.Reference(s) to main scientific papers and/or software package:

[1]Edwin J. Matthews, Naomi L. Kruhlak, Michael C. Cimino, R. Daniel Benz, Joseph F. Contrera. An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: I. Identification of carcinogens using surrogate endpoints Regulatory Toxicology and Pharmacology 44 (2006) 83–96
[2]Edwin J. Matthews, Naomi L. Kruhlak, Michael C. Cimino, R. Daniel Benz, Joseph F. Contrera. An

analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: II. IdentiWcation of genotoxicants, reprotoxicants, and carcinogens using in silico methods Regulatory Toxicology and Pharmacology 44 (2006) 97–110
[3]Yoo JW, Minnier BL, Kruhlak NL, Stavitskaya L. Development of improved (Q)SAR models for predicting the outcome of the in vivo micronucleus genetic toxicity assay. Abstracts of Papers, 54th Society of Toxicology Annual meetings, San Diego CA, March 22-26, 2015, poster presentation.
[4]Hewes KP, Stavitskaya L, Minnier BL, Kruhlak NL. Construction and application of (Q)SAR models to predict in vitro chromosome aberrations. Abstracts of Papers, 54th Society of Toxicology Annual meetings, San Diego CA, March 22-26, 2015, poster presentation.
[5]FDA Drug Labels data from: Data provided by the U.S. Food and Drug Administration https://open.fda.gov
[6]European Chemicals Agency http://echa.europa.eu/

2.8.Availability of information about the model:

Model is commercial. Although the training set is not publicly available, Information about the non-proprietary training set chemicals, assay conditions and details, information about the alerts are available through CASE Ultra interface. For any other specific details contact: sales@multicase.com, MultiCASE Inc. 5885 Landerbrook Dr. #210 Mayfield Heights, OH 44124 USA. Phone: +1-440-565-7221.

2.9.Availability of another QMRF for exactly the same model:

None

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Mouse lymphoma, unactivated (without metabolism)

3.2.Endpoint:

Genotoxicity Mouse Lymphoma, unactivated

3.3.Comment on endpoint:

The in vitro mammalian cell gene mutation test can be used to detect gene mutations induced by chemical substances. This TG includes two distinct in vitro mammalian gene mutation assays requiring two specific tk heterozygous cells lines: L5178Y tk+/-3.7.2C cells for the mouse lymphoma assay (MLA) and TK6 tk+/- cells for the TK6 assay. Genetic events detected using the tk locus include both gene mutations and chromosomal events. Cells in suspension or monolayer culture are exposed to, at least four analyzable concentrations of the test substance, both with and without metabolic activation, for a suitable period of time. They are sub-cultured to determine cytotoxicity and to allow phenotypic expression prior to mutant selection. Cytotoxicity is usually determined by measuring the relative cloning efficiency (survival) or relative total growth of the cultures after the treatment period. The treated cultures are maintained in growth medium for a sufficient period of time, characteristic of each selected locus and cell type, to allow near-optimal phenotypic expression of induced mutations. Mutant frequency is determined by seeding known numbers of cells in medium containing the selective agent to detect mutant cells, and in medium without selective agent to determine the cloning

efficiency (viability). After a suitable incubation time, colonies are counted.

3.4.Endpoint units:

Binary Units

3.5.Dependent variable:

Overall Positive (1) or Negative (0). The final calls were determined as a summary of all the strains used in the test.

3.6.Experimental protocol:

Per OECD Guideline 476, OECD Guideline 490

3.7.Endpoint data quality and variability:

High quality curated data. Structures of training chemicals and names were verified. Duplicates were removed. Mixtures components were manually reviewed and used as separate entries if applicable.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Model built using Statistical Machine Learning techniques.
QSAR model with binary classification ability. Consists of a logistic regression model with molecular fragment/substructures as the descriptors. The descriptors cover both potentiating and deactivating/mitigating molecular features for compounds exhibiting or not-exhibiting genotoxic potential via gene mutation. The molecular features related to gene mutation were identified from training data using various machine learning techniques.

4.2.Explicit algorithm:

Logistic regression QSAR
Training: Multiple Logistic Regression model with occurrence of Alerts and Deactivating Features as independent and overall test outcome as dependent variable. Prediction: Application of the logistic regression model using the identification of alerts and modulators in the query compounds

4.3.Descriptors in the model:

Molecular fragment based descriptors. Occurrence of molecular fragment-based Alerts and modulating features as independent and overall test outcome as dependent variable.

4.4.Descriptor selection:

A descriptor selection process was applied on the initial pool of molecular fragments which picks up the fragments with positive and negative contributions so as to give the best predictive ability to the whole model. The final model contains 434 alerts.

4.5.Algorithm and descriptor generation:

Descriptors for this CASE Ultra model are molecular fragments which are generated from splitting the training set compounds systematically and creating a dictionary of unique fragments. After selecting a few most relevant fragments, a statistical logistic regression data-fitting was applied between the X and Y variables to give the final model.

4.6.Software name and version for descriptor generation:

CASE Ultra V 1.9.0.8
QSAR based bioactivity and toxicity prediction software

sales@multicase.com, MultiCASE Inc, 5885 Landerbrook Dr. #210 Mayfield Heights, OH 44124 USA www.multicase.com
http://www.multicase.com/case-ultra

4.7.Chemicals/Descriptors ratio:

Number of Chemicals = 2528 (732 positives/1796 negatives), Number of Descriptors = 434

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of the model is defined by fragment based chemical space defined by the training set chemicals and range in the computed prediction probabilities where the model has weakest differentiability.

5.2.Method used to assess the applicability domain:

The CASE Ultra program evaluates automatically whether a tested molecule is within the domain of applicability of the model it is tested with. A combination of two criteria were used:

1. Checking for 3-atom fragments that are not present in the training chemicals, and
2. Calculated prediction probabilities that fall between 0.20 - 0.40 where the model has weakest differentiability

5.3.Software name and version for applicability domain assessment:

CASE Ultra Version 1.9.0.8
QSAR expert system for in-silico prediction of toxicity and bioactivity of chemicals
sales@multicase.com, MultiCASE Inc, 5885 Landerbrook Dr. #210 Mayfield Heights, OH 44124 USA www.multicase.com
http://www.multicase.com/case-ultra

5.4.Limits of applicability:

Inorganic compounds, mixtures and large biomolecules are not covered.
In addition,

1. Test chemicals with 3-atom fragments that are not present in the training chemicals potentially are out-of-domain, and
2. Test chemicals with computed prediction probabilities between 0.20 - 0.40 are in grey zone

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

CAS RN: Yes
Chemical Name: Yes
Smiles: Yes
Formula: No
INChI: No
MOL file: Yes
NanoMaterial: null

6.3.Data for each descriptor variable for the training set:

Some

6.4.Data for the dependent variable for the training set:

Some

6.5.Other information about the training set:

Within CASE Ultra interface, all the fragment descriptors are supported by the training chemicals. Every descriptor is supported by relevant statistics, e.g. number of positive and negative training chemicals that contain the fragment. Training set chemicals are explained with assay type, assay conditions, scientific publications etc.

6.6.Pre-processing of data before modelling:

1. Extracted new data from FDA drug safety labels and ECHA, also included data from existing models (data obtained through research collaboration with FDA)
2. Verification of chemical structures, registry numbers, CID and names.
3. Duplicates were removed.
4. Mixture components were treated on case by case basis, components if necessary were separated and assigned activity if possible.

6.7.Statistics for goodness-of-fit:

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

not performed

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

Sensitivity 82.8%; Specificity 83.2% Positive predictivity 66.3% Negative predictivity 92.5%, Coverage 67.1% AUC 0.905, 10 iterations, 10% off Classification cut-off 0.30

6.10.Robustness - Statistics obtained by Y-scrambling:

Sensitivity 47.9%; Specificity 54.8% Positive predictivity 33.7% Negative predictivity 68.6%, Coverage 29.3% AUC 0.532, 10 iterations, 10% off Classification cut-off 0.30

6.11.Robustness - Statistics obtained by bootstrap:

Sensitivity 82.2%; Specificity 82.3% Positive predictivity 64.4% Negative predictivity 92.3%, Coverage 68.1% AUC 0.901, 10 iterations, 10% off Classification cut-off 0.30

6.12.Robustness - Statistics obtained by other methods:

Self Validation - Sensitivity 88.6%; Specificity 89.2% Positive predictivity 77.3% Negative predictivity 95.0%, Coverage 77% AUC 0.963, Classification cut-off 0.30

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes

7.2.Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: Yes

NanoMaterial: null

7.3.Data for each descriptor variable for the external validation set:

Some

7.4.Data for the dependent variable for the external validation set:

Some

7.5.Other information about the external validation set:

7.6.Experimental design of test set:

The external set was composed of 140 compounds, 55 positive and 85 negatives. The external compounds were randomly selected from complete dataset (before splitting into training and test sets). Experimental protocol of the external compounds are same as the training set compounds.

7.7.Predictivity - Statistics obtained by external validation:

Sensitivity 78.05%; Specificity 80.60% Positive predictivity 71.11% Negative predictivity 85.71%, Coverage 77.14%, Classification cut-off 0.30

7.8.Predictivity - Assessment of the external validation set:

7.9.Comments on the external validation of the model:

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

CASE Ultra models do not have any predefined knowledge of molecular mechanism that explains the activity of a molecule. However, the way the modules were built, splitting the entire learning set into clusters of molecules with a dedicated QSAR in every cluster, suggests very close links with a mechanistic explanations of activity. Indeed many of the resulting biophores have modes of action that are obvious to persons with expert knowledge for the endpoint in question. For example, the presence of an alert containing N-nitroso fragment in bacterial mutagenicity model will undoubtedly suggest potential mutagenicity activity. Other fragments, which do not have such a clear mechanism of action assigned to them, can support an intelligent guess about possible sets of events causing activity. Either way, it is certain that the results of a MultiCASE analysis can serve as a mechanistic research tool as well as a QSAR builder.

8.2.A priori or a posteriori mechanistic interpretation:

The mechanistic basis of the model was neither determined a priori nor a posteriori. The selected features were mined completely automatically from the training data during the model building process and they agree very well with the known chemical mechanisms of genotoxicity via gene mutation. The training structures were also not selected with any specific mechanism in mind.

8.3.Other information about the mechanistic interpretation:

None

9.Miscellaneous information

9.1.Comments:

This model should be useful in the risk assessment of identifying compounds causing genotoxicity via gene mutation. It will also be helpful in understanding various known and unknown mechanisms of compounds causing genotoxicity via gene mutation. It will be particularly helpful in regulatory submission. When a prediction is made using this model in CASE Ultra program, the identified alerts (if any) are highlighted in the query chemical which is helpful in interpreting the results.

9.2.Bibliography:

[1]Optimizing predictive performance of CASE Ultra expert system models using the applicability domains of individual toxicity alerts; Chakravarti, S.K., Saiakhov, R.D. and Klopman, G., Journal of Chemical Information and Modeling, 2012, 52, 2609-2618. DOI: 10.1021/ci300111r

[2]Effectiveness of CASE Ultra Expert System in Evaluating Adverse Effects of Drugs; Saiakhov, R.D., Chakravarti, S.K. and Klopman, G.; Molecular Informatics, 2012, 32, 87-97.

[3]Computing similarity between structural environments of mutagenicity alerts, Chakravarti, S.K.,Saiakhov, R. D.; Mutagenesis, October 20, 2018,

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC