MultiCASE CASE Ultra model for mutations in the hypoxanthine-guanine phosphoribosol transferase (HGPRT) locus in Chinese Hamster Ovary (CHO) cells *in vitro*

1. QSAR identifier

1.1 QSAR identifier (title)

MultiCASE CASE Ultra model for mutations in the hypoxanthine-guanine phosphoribosol transferase (HGPRT) locus in Chinese Hamster Ovary (CHO) cells *in vitro*, Danish QSAR Group at DTU Food.

1.2 Other related models

Leadscope Enterprise model for mutations in the hypoxanthine-guanine phosphoribosol transferase (HGPRT) locus in Chinese Hamster Ovary (CHO) cells *in vitro*, Danish QSAR Group at DTU Food.

SciMatics SciQSAR model for mutations in the hypoxanthine-guanine phosphoribosol transferase (HGPRT) locus in Chinese Hamster Ovary (CHO) cells *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

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 consists of data from Li *et al.* (1988) plus data from Chemical Carcinogenesis Research Information System (CCRIS), Hazardous Substances Data Bank (HSDB®) and Environmental Mutagen Information Center (EMIC) extracted from CHEM-BANK (2002). 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

Chinese hamster (Chinese Hamster Ovary (CHO) cells).

3.2 Endpoint

QMRF 4.10. Mutagenicity

OECD 476 In Vitro Mammalian Cell Gene Mutation Test

3.3 Comment on endpoint

Mammals differ from prokaryotes in their level of organization and repair of DNA, mechanisms of metabolism, and other related reactions; thus studies of mutagenesis in prokaryotes (e.g. the Ames test) may not reveal some fundamental mechanisms of mutagenesis in mammals. For these reasons mammalian-cell mutational assay systems, utilizing mammalian cells in culture (e.g. CHO cells), have been developed. Such systems are valuable in assessing the genetic hazard of environmental agents to the human population. The Chinese Hamster Ovary cell/hypoxanthine guanine phosphoribosyl transferase (CHO/HGPRT) assay is an example of such a system.

Hypohanthine-guanine phosphoribosol transferase (HGPRT) is an enzyme that is involved in the conversion of the nucleoside guanine to the mononucleotide guanosine monophosphate (GMP) used in the formation of DNA. The conversion of the guanine analog, 6-thioguanin (6-TG) to 6-thioguanine-containing mononucleotide is also catalyzed by HGPRT. The 6-thioguanine-containing mononucleotide cannot, in oppose to GMP, be used in the synthesis of DNA. Therefore the presence of 6-TG will give rise to absence of cell proliferation.
Damage such as forward mutation(s) (i.e. mutation from wild type to mutant) in the X-linked HGPRT locus coding for the HGPRT enzyme may result in inability of the HGPRT enzyme to catalyse the conversion of purines to mononucleotides. Fortunately, GMP can be produced from guanine by other metabolic pathways and the cells will survive despite a dysfunctional HGPRT enzyme.
Cells with functional HGPRT enzyme (no mutation in the HGPRT locus) will in medium containing 6-TG, the selective agent, undergo cell death because of missing DNA replication. Therefore no colonies will appear in the selective medium (except for few colonies because of background mutations). If on the other hand a test substance is causing a mutation in the HGPRT locus, the cells become resistant to 6-TG as it cannot be conversed to the 6-TG containing mononucleotide. The cells will therefore, because of GMP derived from other pathways, be able to proliferate in the selective 6-TG medium and form colonies.

Part of the training set data were compiled by Li *et al.* (1988), who reviewed and evaluated literature containing CHO/HGPRT assay results published from mid-1979 through June 1986. A careful evaluation of data quality was done by Li and coworkers, who put up criteria for data inclusion such as requirement of negative controls, cytotoxicity testing, metabolic activation systems etc. (see section II in Li *et al.* 1988 for more details). In addition to this data, similar data were compiled from various databases (CHEM-BANK 2002) and included in the training set.

3.4 Endpoint units

CASE units, 39 or 59 for positives and 10 for negatives.

3.5 Dependent variable

Forward mutations in the hypoxanthine-guanine phosphoribosol transferase (HGPRT) locus in Chinese Hamster Ovary (CHO) cells *in vitro*, positive or negative.

3.6 Experimental protocol

The experimental protocol has been described in OECD guideline 476 (1997). Briefly, cells in suspension or monolayer culture are exposed to the test substance, both with and without metabolic activation, for a suitable period of time and subcultured to determine cytotoxicity and to allow phenotypic expression prior to mutant selection. Mutant frequency is determined by seeding known numbers of cells in medium containing a selective agent (6-TG) to detect mutant cells, and in medium without selective agent to determine the cloning efficiency (viability). After a suitable incubation time, colonies are counted.
The CHO/HGPRT assay is an appropriate *in vitro* assay system for use in the screening of chemicals for genotoxicity (Li *et al.* 1988).

3.7 Endpoint data quality and variability

As data originates from multiple sources a certain degree of variability in data is expected. Because of the strict criteria for data acceptance for the HGPRT data from Li and co-workers this data are expected to be of low variability and high quality.

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 = 0.57 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

239 compounds are in the training set: 146 positives and 93 negatives.

6.6 Pre-processing of data before modelling

Only structures acceptable for CASE Ultra 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. No further structures accepted by the software were eliminated (i.e. outliers).

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 47.7% (570/(5*239)) of the predictions which were within the applicability domain:

- Sensitivity (true positives / (true positives + false negatives)): 75.4%
- Specificity (true negatives / (true negatives + false positives)): 84.5%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 78.9%

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 training set

The test set consists of 150 negative compounds from Grant *et al.* (2000).

7.6 Experimental design of test set

Not performed. The test set consists of physiological compounds normally present in mammalian cells and for this reason they are assumed to be negative in the CHO/HGPRT assay.

7.7 Predictivity – Statistics obtained by external validation

External validation for Specificity:

Of the 150 test set compounds 58 compounds (38.7%) were within the applicability domain of the model, and of these 55 were predicted to be negative for mutations in the hypoxanthine-guanine phosphoribosol transferase (HGPRT) locus in Chinese Hamster Ovary (CHO) cells:

- Specificity (true negatives / (true negatives + false positives)): 55/(55+3) = 94.8%

7.8 Predictivity – Assessment of the external validation set

The compounds in the test set were not expected to cause mutations in the hypoxanthine-guanine phosphoribosol transferase (HGPRT) locus in Chinese Hamster Ovary (CHO) cells because of their normal presence in mammalian cells.

7.9 Comments on the external validation of the model

The external validation described was on a former version of this model made in the software MC4PC, a predecessor to the Case Ultra software, and using the same training set. It is therefore not a true external validation of this CASE Ultra model. Because this model is based on very similar methodologies and the same training set as the previous model it is likely that the results for the external validation on the former model to a high degree reflect external validations with the same test sets of 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 if forward mutations will occur in the hypoxanthine-guanine phosphoribosol transferase (HGPRT) locus in Chinese Hamster Ovary (CHO) cells *in vitro* upon exposure to a chemical.

9.2 Bibliography

CHEM-BANK (2002) Databanks of potentially hazardous chemicals: RTECS, OHMTADS, CHRIS, HSDB, IRIS, TSCA, NPG and ERG2000. USA. CHEM-BANK™, CD-ROM, SilverPlatter International N.V., August 2002.

Grant, S.G., Zhang, Y.P., Klopman, G., and Rosenkranz, H.S. (2000) Modeling the mouse lymphoma forward mutational assay: the Gene-Tox program database. Mutation Research, 465, 201–229.

Li, A.P., Gupta, R.S., Heflich, R.H., and Wassom, J.S. (1988) A review and analysis of the Chinese hamster ovary/hypoxanthine guanine phosphoribosyl transferase assay to determine the mutagenicity of chemical agents. A report of phase III of the U.S. Environmental Protection Agency Gene-Tox Program. *Mutation Research,* 196, 17-36.

OECD guideline 476 (1997) OECD Guidelines for the Testing of Chemicals No. 476: *In Vitro* Mammalian Cell Gene 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