Leadscope Enterprise model for the Bacterial Reverse Mutation Test (Ames test) in *S. typhimurium in vitro*

1. QSAR identifier

1.1 QSAR identifier (title)

Leadscope Enterprise 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

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

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

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

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

Roberts, G., Myatt, G. J., Johnson, W. P., Cross, K. P., and Blower, P. E. J. (2000) LeadScope: Software for Exploring Large Sets of Screening Data. *Chem. Inf. Comput. Sci.*, 40, 1302-1314.

Cross, K.P., Myatt, G., Yang, C., Fligner, M.A., Verducci, J.S., and Blower, P.E. Jr. (2003) Finding Discriminating Structural Features by Reassembling Common Building Blocks. *J. Med. Chem.*, *46,* 4770-4775.

Valerio, L. G., Yang, C., Arvidson, K. B., and Kruhlak, N. L. (2010) A structural feature-based computational approach for toxicology predictions. *Expert Opin. Drug Metab. Toxicol.*, 6:4, 505-518.

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

No units, 1 for positives and 0 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 features and numeric molecular descriptors.

4.2 Explicit algorithm

This is a categorical (Q)SAR model made by use of partial logistic regression (PLR). The model is a composite model consisting of 2 submodels, using all the negatives (1803 chemicals) in each of these and different sub-sets of the positives (see 4.5). The specific implementation is proprietary within the Leadscope software.

4.3 Descriptors in the model

structural features,

aLogP,

polar surface area,

number of hydrogen bond donors,

Lipinski score,

number of rotational bonds,

parent atom count,

parent molecular weight,

number of hydrogen bond acceptors

4.4 Descriptor selection

Leadscope Predictive Data Miner is a software program for systematic sub-structural analysis of a chemical using predefined structural features stored in a template library, training set-dependent generated structural features (scaffolds) and calculated molecular descriptors. The feature library contains approximately 27,000 pre-defined structural features and the structural features chosen for the library are motivated by those typically found in small molecules: aromatics, heterocycles, spacer groups, simple substituents. Leadscope allows for the generation of training set-dependent structural features (scaffold generation), and these features can be added to the pre-defined structural features from the library and be included in the descriptor selection process. It is possible in Leadscope to remove redundant structural

features before the descriptor selection process and only use the remaining features in the descriptor selection process. Besides the structural features Leadscope also calculates eight molecular descriptors for each training set structure: the octanol/water partition coefficient (alogP), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), Lipinski score, atom count, parent compound molecular weight, polar surface area (PSA) and rotatable bonds. These eight molecular descriptors are also included in the descriptor selection process.

Leadscope has a default automatic descriptor selection procedure. This procedure selects the top 30% of the descriptors (structural features and molecular descriptors) according to X^2-test for a binary variable, or the top and bottom 15% descriptors according to *t*-test for a continuous variable. Leadscope treats numeric property data as ordinal categorical data. If the input data is continuous such as IC_{50} or cLogP data, the user can determine how values are assigned to categories: the number of categories and the cut-off values between categories. (Roberts *et al.*2000).

When developing this model, intermediate models with application of different modelling approaches in Leadscope were tried:

1. 'Single model' using only the Leadscope pre-defined structural features, i.e. no scaffolds, and calculated molecular descriptors for descriptor selection.
2. 'Single model' using both the Leadscope pre-defined structural features and the training set dependent features (scaffolds generation) as well as the calculated molecular descriptors in the descriptor selection.
3. 'Single model' using both Leadscope pre-defined structural features and the training set dependent features (scaffolds generation), with subsequent removal of redundant structural features, and calculated molecular descriptors for descriptor selection.
4. 'Composite model' using only the Leadscope pre-defined structural features, i.e. no scaffolds, and calculated molecular descriptors in the descriptor selection.
5. 'Composite model' using both Leadscope pre-defined structural features and the training set dependent features (scaffolds generation) as well as the calculated molecular descriptors in the descriptor selection.

Based on model performance as measured by a preliminary cross-validation the model developed using approach number 4. was chosen:

In this model the descriptors were automatically selected among the pre-defined structural features and the eight molecular descriptors.

4.5 Algorithm and descriptor generation

For descriptor generation see 4.4.

After selection of descriptors the Leadscope Predictive Data Miner program performs partial least squares (PLS) regression for a continuous response variable, or partial logistic regression (PLR) for a binary response variable, to build a predictive model. By default the Predictive Data Miner performs leave-one-out or leave-groups-out (in the latter case, the user can specify any number of repetitions and percentage of structures left out) cross-validation on the training set depending on the size of the training set. In the cross-validation

made by Leadscope the descriptors selected for the 'mother model' are used when building the validation submodels and they therefore have a tendency to be overfittet and give overoptimistic validation results.

In this model because of the categorical outcome in the response variable PLR was used to build the predictive model. Because of the unbalanced training set (i.e. 2299 positives vs. 1803 negatives) 2 submodels for smaller individual training sets consisting of the 1803 negatives and an equal number of positives selected by random among the 2299 positives were made. The descriptors for each of the 2 submodels were automatically selected from the Leadscope feature library based solely on the training set compounds used to build the individual submodel and was not affected by the training set chemicals in the composite model. Therefore, a different number of descriptors (structural features and molecular descriptors) were selected and distributed on varying number of PLS factors for each submodel.

4.6 Software name and version for descriptor generation

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

4.7 Descriptors/chemicals ratio

As this model is a composite model consisting of 2 submodels with varying training set size and using different descriptors and number of PLS factors (see 4.5), an overall descriptor/chemical ratio for this model cannot be calculated.

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 of a structural domain in Leadscope and the in-house further probability refinement algorithm on the output from Leadscope to reach the final applicability domain call.

1. Leadscope

For assessing if a test compound is within the structural applicability domain of a given model Leadscope examines whether the test compound bears enough structural resemblance to the training set compounds used for building the model (i.e. a structural domain analysis). This is done by calculating the distance between the test compound and all compounds in the training set (distance = 1 - similarity). The similarity score is based on the Tanimoto method. The number of neighbours is defined as the number of compounds in the training set that have a distance equal to or smaller than 0.7 with respect to the test compound. The higher the number of neighbours, the more reliable the prediction for the test compound. Statistics of the distances are also calculated. Effectively no predictions are made for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be calculated in Leadscope.

2. The Danish QSAR group

In addition to the general Leadscope structural applicability domain definition the Danish QSAR group has applied a further requirement to the applicability domain of the model. That is only positive predictions with a probability equal to or greater than 0.7 and negative predictions with probability equal to or less than 0.3 are accepted. Predictions within the structural applicability domain but with probability between 0.5 to 0.7 or 0.3 to 0.5 are defined as positives out of applicability domain and negatives out of applicability domain, respectively. When these predictions are wed out the performance of the model in general increases at the expense of reduced model coverage.

5.2 Method used to assess the applicability domain

Leadscope does not generate predictions for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be calculated.

Only positive predictions with probability equal to or greater than 0.7 and negative predictions with probability equal to or less than 0.3 are accepted.

5.3 Software name and version for applicability domain assessment

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

5.4 Limits of applicability

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). Calculation 2D structures (SMILES and/or SDF) are generated by stripping off ions (of the accepted list given above). 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", where the 50% contains the same ratio of positive and negatives as the full training set. A new model (validation submodel) was created on the remaining 50% using the same settings in Leadscope but with no information from the "mother model" regarding descriptor selection etc. The validation submodel was applied to predict the removed 50% (within the defined applicability domain for the submodel). Likewise, a validation submodel was made on the removed 50% of the training set and this model was used to predict the other 50% (within the defined applicability domain for this submodel). This procedure was repeated five times.

Predictions within the defined applicability domain of the ten validation submodels were pooled and Cooper's statistics calculated. This gave the following results for the 73.7% (15115/(5*4102)) of the predictions which were within the applicability domains of the respective submodels:

- Sensitivity (true positives / (true positives + false negatives)): 84.3%
- Specificity (true negatives / (true negatives + false positives)): 85.7%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 84.9%

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

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 2,036 compounds (58%) 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)): 601/(601+173)= 77.6%

- Specificity (true negatives / (true negatives + false positives)): 1059/(1059+203)= 83.9%

- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): (601+1059)/2036= 81.5%

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 global model identifies structural features and molecular descriptors which in the model development was found to be statistically significant associated with effect. Many predictions may indicate modes of action that are obvious for persons with expert knowledge for 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).

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.

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