Leadscope Enterprise model for sister chromatid exchange (SCE) in mouse bone marrow cells *in vivo*

1. QSAR identifier

1.1 QSAR identifier (title)

Leadscope Enterprise model for sister chromatid exchange (SCE) in mouse bone marrow cells *in vivo*, Danish QSAR Group at DTU Food.

1.2 Other related models

MultiCASE CASE Ultra model for sister chromatid exchange (SCE) in mouse bone marrow cells *in vivo*, Danish QSAR Group at DTU Food.

SciMatics SciQSAR model for sister chromatid exchange (SCE) in mouse bone marrow cells *in vivo*, 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

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Jay Russel Niemelä;

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 consists of data from Labbauf *et al.* (1997) using Gene-Tox data from Tucker *et al.* (1993), plus newer data from Environmental Mutation Information Center (EMIC) (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

Mouse (bone marrow cells).

3.2 Endpoint

QMRF 4. Human Health Effects

QMRF 4.10. Mutagenicity

3.3 Comment on endpoint

During the S phase of the cell cycle, DNA is replicated, and each chromosome becomes duplicated into two closely associated sister chromatids, which are joined tightly at the centromere. Sister chromatids are distinctly visible cytologically in late prophase and early metaphase of mitosis before segregating to the resulting daughter cells. Sister chromatid exchange (SCE) is the reciprocal interchange of genetic material between two identical sister chromatids and this event is known to occur as a normal feature of cell division in mammalian tissues. The simplest pathway by which SCE likely occurs is through homologous recombination (HR)-mediated restart of a broken DNA replication fork when it encounters a nick or gap in one parental strand. Therefore, DNA lesions introduced by mutagens induce collapse of the replication forks and the broken strands are subsequently subject to repair mainly by HR, which results in increased SCE frequency. The process is considered conservative and error-free, since no information is generally altered during reciprocal interchange by HR.
The *in vivo* SCE assay detects the ability of a test substance to enhance the exchange of DNA between two sister chromatids of a duplicating chromosome. The test is considered to be a sensitive method for evaluating mutagenicity and may be an indicator of carcinogenicity.

3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

3.5 Dependent variable

Sister chromatid exchange (SCE) in mouse bone marrow cells *in vivo*, positive or negative.

3.6 Experimental protocol

Animals (mice) are exposed to test substance by appropriate routes followed by administration of bromodeoxyuridine (BrdU). A spindle inhibitor is administered prior to sacrifice to arrest cells in a metaphase-like stage of mitosis (c-metaphase), a point in cell division where chromosomes are most dense and therefore visible in the light microscope. After sacrifice, chromosome preparations from bone marrow cells are made, stained (fluorescein plus Giemsa method) and scored for SCE. (US EPA guideline 1996)

It should be noted that BrdUrd itself induces SCE and is responsible for most of the base line SCE observed in absence of test substance (Latt *et al.* 1981).

3.7 Endpoint data quality and variability

The data set from Labbauf *et al.* (1997) was originally compiled by Tucker *et al.* (1993) who reviewed and evaluated published literature from 1979 to 1986. The remaining data in the training set was compiled from EMIC (CHEM-BANK 2002). As data originates from several different sources it cannot be excluded that different experimental procedures have been used to obtain the data and therefore a certain degree of variability in the data is expected (species, strain, treatment protocol etc.).

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 positives (103 chemicals) in each of these and different sub-sets of the negatives (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. 103 positives vs. 162 negatives) 2 submodels for smaller individual training sets consisting of the 103 positives and an equal number of negatives selected by random among the 162 negatives 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

265 compounds are in the training set: 103 positives and 162 negatives.

6.6 Pre-processing of data before modelling

Only structures acceptable for Leadscope 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”, 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 68.1% (902/(5*265)) of the predictions which were within the applicability domains of the respective submodels:

- Sensitivity (true positives / (true positives + false negatives)): 92.6%
- Specificity (true negatives / (true negatives + false positives)): 95.0%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 94.3%

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

7.2 Available information for the external validation set

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

7.6 Experimental design of test set

7.7 Predictivity – Statistics obtained by external validation

7.8 Predictivity – Assessment of the external validation set

7.9 Comments on the external validation of the model

External validation not performed.

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

The model can be used to predict results for the sister chromatid exchange (SCE) in mouse bone marrow cells *in vivo* test.

9.2 Bibliography

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

Labbauf, A., Klopman, G., and Rosenkranz, H.S. (1997) Dichotomous relationship between DNA reactivity and the induction of sister chromatid exchanges in vivo and in vitro. *Mutation Research*, 377, 37–52.

Latt, S.A., Allen, J., Bloom, S.E., Carrano, A., Falke, E., Kram, D., Schneider, E., Schreck, R. Tice, R., Whitfield, B., and Wolff, S. (1981) Sister-chromatid exchanges: A report of the Gene-Tox program. *Mutation Research,* 87, 17—62.

Tucker, J.D., Auletta, A., Cimino, M.C., Dearfield, K.L., Jacobson-Kram, D., Tice, R.R., and Carrano, A.V. (1993) Sister-chromatid exchange: Second report of the Gene-Tox program. *Mutation Research,* 297, 101-180.

US EPA guideline (1996) OPPTS 870.5915 In Vivo Sister Chromatid Exchange Assay. US EPA Health Effect Test Guidelines, June 1996. Available on: http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OPPT-2009-0156-0039

9.3 Supporting information