QMRF	*QMRF identifier (JRC Inventory):* To be entered by JRC *QMRF Title:* Nonlinear Classification ANN QSAR Model for mutagenicity (Salmonella typhimurium strains) *Printing Date:* 30.03.2011	QMRF

1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear Classification ANN QSAR Model for mutagenicity (Salmonella typhimurium strains)

1.2.Other related models:

1.3.Software coding the model:

QSARModel 3.3.8; Statistica 7, StatSoft Ltd. Turu 2, Tartu, 51014, Estonia, http://www.molcode.com

2.General information

2.1.Date of QMRF:

13.12.2010

2.2.QMRF author(s) and contact details:

Dimitar Dobchev, Tarmo Tamm, Gunnar Karelson, Indrek Tulp, Dana Martin, Kaido Tämm, Deniss Savchenko, Jaak Jänes, Eneli Härk, Andres Kreegipuu, Mati Karelson, Molcode model development team Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Molcode model development team Molcode Ltd Molcode Ltd Turu 2, Tartu, 51014, Estonia models@molcode.com www.molcode.com

2.6.Date of model development and/or publication:

12.04.2010The methodology and software (QSARModel) used to create the present model were applied also to obtain the results published in these papers.

1)Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Hur, E.; Tämm, K.; Kurunczi, L.; Karelson, M.; Varnek, A.; Solov'ev, V. P. (2006). Skin Permeation Rate as a Function of Chemical Structure . Journal of Medicinal Chemistry, 49(11), 3305 - 3314.

2)Karelson, M.; Dobchev, D. A.; Kulshyn, O. V.; Katritzky, A. (2006). Neural Networks Convergence Using Physicochemical Data. Journal of Chemical Information and Modeling, 46, 1891 - 1897.

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

[1]1)Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Hur, E.; Tämm, K.; Kurunczi, L.; Karelson, M.; Varnek, A.; Solov'ev, V. P. (2006). Skin

Permeation Rate as a Function of Chemical Structure . Journal of Medicinal Chemistry, 49(11), 3305 - 3314.
[2]2)Karelson, M.; Dobchev, D. A.; Kulshyn, O. V.; Katritzky, A. (2006). Neural Networks Convergence Using Physicochemical Data. Journal of Chemical Information and Modeling, 46, 1891 - 1897.
[3]Statistica 7 www.statsoft.com
[4]

2.8.Availability of information about the model:

All information in full detail is available

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

No other QMRF available for the same model

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Salmonella typhimurium strains TA98, TA100, TA1535, TA1537, TA97, TA102 and TA1538

3.2.Endpoint:

Mutagenicity

3.3.Comment on endpoint:

Determination of evidence of mutagenicity

Mutagenicity: reverse mutation test using bacteria was carried out according to the OECD 471 (EU B.13/14) test guideline [1]. The bacterial reversed mutation assay (Ames Test) is used to detect point mutations, which involve substitution, addition or deletion of one or a few DNA base pairs.

3.4.Endpoint units:

3.5.Dependent variable:

Mutagenicity Index (+ presence, - absence) - AMES

3.6.Experimental protocol:

The dataset comprises 220 compounds with α,β-unsaturated carbonyl moiety derived from the Ames test classification for mutagenicity [2]. The data were collected from the Chemical Carcinogenicity Research Information System (CCRIS) database, which contains scientifically evaluated Ames test data for approximately 7000 compounds and mixtures, which are identified with a CAS registry number and/or chemical name(s).

The additional data were obtained from other public toxicity databases, which also contain data from Ames tests that were performed before strict regulatory requirements were imposed for the authorization of new chemicals. The molecular structures of these compounds were either retrieved from the National Cancer Institute's Developmental Therapeutics Program database and via Beilstein by means of their CAS registry number or constructed from their chemical name(s). Inorganic

compounds, organometallic compounds, and additional occurrences of enantiomers and diastereoisomers were then removed from this dataset.

For the construction of a consistent mutagenicity dataset the following criteria were applied. First, to diminish data heterogeneity and avoid data pollution by nonstandard Ames tests, the analysis was restricted to standard Ames test data of Salmonella Typhimurium strains TA98, TA100, TA1535 and either TA1537 or TA97, which are required for regulatory evaluation of drug approval. In addition, strains TA102 and TA1538 were also selected, since they are applied in cases where results of other strains are equivocal or difficult to interpret.

Further, 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. 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].

3.7.Endpoint data quality and variability:

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Nonlinear QSAR: Backpropagation Neural Network (Multilayer Perceptron) regression

4.2.Explicit algorithm:

The algorithm is based on regression neural network predictor with structure 8-8-7-1

4.3.Descriptors in the model:

[1]Partial Charged (AM1) Surface Area of H atoms
[2]Square root of Partial Charged (AM1) Surface Area of H atoms
[3]Highest atomic state energy (AM1) for O atoms
[4]RPCG Relative positive charge (QMPOS/QTPLUS) (AM1)
[5]Max atomic orbital electronic population (AM1)
[6]HASA-1/TMSA (AM1) (all)
[7]Max net atomic charge (Zefirov) for O atoms
[8]Lowest e-e repulsion (1-center) (AM1) for O atoms
[9]

4.4.Descriptor selection:

Initial pool of ~909 descriptors. Stepwise descriptor selection based on a set of statistical selection rules as F statistic and p. The first highest F (low p) descriptors (8) were selected from the total number of descriptors. These 8 descriptors were used as inputs to the network. 9 networks with different structures were tested in order to find the best ANN with lowest RMS (root-mean-squared error) and highest correct predictions (for training, selection and test sets). Then 311 epochs were used to train the final network with architecture depicted in 4.2. Optimization of the weights was performed with Levenberg-Marquardt algorithm encoded in the backpropagation scheme using linear and hyperbolic activation functions.

4.5.Algorithm and descriptor generation:

All descriptors were generated using QSARModel on structure optimized by AM1 semiempirical quantum mechanical model.

4.6.Software name and version for descriptor generation:

QSARModel
http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

17

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain based on training set:

a)functional grups as phenols, aldehydes, nitro, amino, alcohols,halides, aromatic, aliphatic functional groups and other

b)The model is suitable for compounds that
have descriptors values in the followin range;

Desc 1 2 3 4 5 6 7 8
min0.0000.000-308.8560.0541.8840.019-0.185200.970
max0.0940.021-304.4560.5731.9890.954-0.057228.153

5.2.Method used to assess the applicability domain:

Presence of functional groups in structures.
Range of descriptor values in training set with ±30% confidence.
Descriptor values must fall between maximal and minimal descriptor values (see5.1) of training set ±30%.

5.3.Software name and version for applicability domain assessment:

QSARModel 3.3.8
http://www.molcode.com

5.4.Limits of applicability:

See 5.1, 5.2

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:No
Formula:No
INChI:No
MOL file:Yes

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

All

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

All

6.5.Other information about the training set:

data points: 136

6.6.Pre-processing of data before modelling:

Standardization and normalization of the inputs by taking into account the mean and standard deviation

6.7.Statistics for goodness-of-fit:

Notations: T- training set, S - selection set, X- test set

T.AMES.1T.AMES.-1S.AMES.1S.AMES.-1X.AMES.1X.AMES.-1
Total67.000069.0000017.0000023.0000019.0000021.00000
Correct67.000067.000009.0000020.0000017.0000018.00000
Wrong0.00002.000008.000003.000002.000003.00000
Correct(%)100.000097.1014552.9411886.9565289.4736885.71429

Wrong(%)0.00002.8985547.0588213.0434810.5263214.28571

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

See 6.7

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

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

Training Performane=0.985, Selection Performance = 0.725, Test Performance=0.875

In this ANN were used 2 sets randomly chosen (40) to test the network – selection set and test set, see also 6.7

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:No
Formula:No
INChI:No
MOL file:Yes

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

All

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

All

7.5.Other information about the external validation set:

The method used two validation sets – selection (40) and test (40)

7.6.Experimental design of test set:

Randomly selected 40 and 40 data points

7.7.Predictivity - Statistics obtained by external validation:

see 6.7 and 6.12

7.8.Predictivity - Assessment of the external validation set:

The descriptors for the test set are in the limit of applicability, see 6.7 and 6.12

7.9.Comments on the external validation of the model:

Overall predictions for the selection set (used to stop the ANN training and not to overfit it) and the test set (used to test the external prediction of the net after training) are significant according to the standard deviation ratio (S.D.Ration-Performance) and the confusion matrix see 6.7 and 6.12

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

It is difficult to interpret the ANN model for this propery becaus of the ANN mathematical structure. However, some general insights for the

mutagenicity can be drawn based on the descriptors in the next.. The Partial Charged (AM1) Surface Area of H atoms

tend to lead to mutagenic compounds whn its values are higher. The same holds for Square root of Partial Charged (AM1) Surface Area of H atoms. In contrast the Max net atomic charge (Zefirov) for O atoms leads to mutagenic compound when its values are low.

Hydrogen surface acceptor area HASA-1/TMSA (AM1) (all) descriptors follows the first trend as for the hydrogen atoms.It seems that compounds with hydrogen acceptor ability tend to be mutagenic.

8.2.A priori or a posteriori mechanistic interpretation:

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

Supporting information for :Training set(s)
Selection set(s)
Test set(s)
8-8-7-1.snn file -includes the ANN model, in order to be used the user must have statistica 7 or higher with ANN modules to make predictions.

9.2.Bibliography:

[1]OECD TG 471, Bacterial Reverse Mutation Test (1997).
[2]Pérez-Garrido A., Morales Helguera A., Girón Rodríguez F., D.S. Cordeiro M. N. QSAR models to predict mutagenicity of acrylates, methacrylates and α,β-unsaturated carbonyl compounds, Dental materials 2010, 26, 397–415.
[3]Kazius J, McGuire R, Bursi R. Derivation and validation of toxicophores for mutagenicity prediction. J. Med. Chem. 2005, 48, 312–20.
[4]

9.3.Supporting information:

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

10.Summary (JRC Inventory)

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