1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear QSAR: aritficial neural network for mouse carcinogenicity

1.2.Other related models:

1.3.Software coding the model:

[1]QSARModel 3.3.8 Turu 2, Tartu, 51014, Estonia http://www.molcode.com

[2]Statistica 7 StatSoft Ltd http://www.statsoft.com

2.General information

2.1.Date of QMRF:

21.04.2010

2.2.QMRF author(s) and contact details:

[1]Dimitar Dobchev Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[2]Tarmo Tamm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[3]Gunnar Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[4]Indrek Tulp Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[5]Dana Martin Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[6]Kaido Tämm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[7]Deniss Savchenko Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[8]Jaak Jänes Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[9]Eneli Härk Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[10]Andres Kreegipuu Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[11]Mati Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[12]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:

12.04.2010

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

Statistica 7 www.statsoft.com

2.8.Availability of information about the model:

Training, selection and test sets available. Algorithm available.

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

None to date.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Mouse

3.2.Endpoint:

4.Human health effects 4.12.Carcinogenicity

3.3.Comment on endpoint:

Carcinogenicity was determined using the OECD Test Guideline 451 (EU Test Guideline B.32). This method describes the administration of test substance normally seven days per week, by an appropriate route, to several groups of experimental animals, one dose per group, for a major portion of their lifespan, and the daily observation of experimental animals for detection of signs of toxicity, particularly the development of tumours. Chemical carcinogens have been categorized as either genotoxic or non-genotoxic. The former are DNA reactive and the latter act by a variety of other mechanisms. The toxicological property of interest was the carcinogenic potency, expressed as TD50 value. The TD50 value for a given target site (s) in the absence of tumors in control animals, was taken to be the chronic dose (in mg/kg body wt/day) which induced tumors in half of the test animals at the end of a standard lifespan for the species. The TD50 value used for each compound was selected by taking into account the lowest carcinogenic potency value reported for each chemical in all the positive reports for that chemical.

3.4.Endpoint units:

TD50 [mmol/kg]

3.5.Dependent variable:

Log (TD50)

3.6.Experimental protocol:

3.7.Endpoint data quality and variability:

Experimental data from different sources have been validated as reliable (ref.5)

A data set of 340 compounds was collected from the database of Chemical carcinogens: structures and experimental data (ISSCAN) which contains information on chemical compounds tested with the long-term carcinogenicity bioassay on rodents (rat, mouse). Beside being a repository of data, it has been specifically designed as an expert decision support tool. Historically, this database originates from the experience of researchers of the Environment and Primary Prevention Department in the field of structure-activity relationships, aimed at developing models which theoretically predict the carcinogenicity of chemicals. The use of experimental carcinogenicity data for structure-activity relationship studies amplifies their informative value, and contributes to the reduction and replacement of animal experimentation. This database does not contain neither epidemiological data nor regulatory classifications of the carcinogens, but only the experimental results from the carcinogenicity bioassay. The structure of this database is inspired by that of the Distributed Structure-Searchable Toxicity (DSSTox) Network of the US Enviromental Protection Agency (EPA) (http://www.epa.gov/nheerl/dsstox/). Similarly to the DSSTox spirit this project wants to contribute to the free diffusion of scientific data in a standardized, easy to read format [2]. Source of carcinogenicity data: Carcinogenic Potency Database (CPDB) established by Gold and Zeiger (1997) [3,4], TOXNET CCRIS, National Toxicology Program (NTP), International Agency for Research on Cancer (IARC), Survey of Compounds which have been tested for Carcinogenic Activity (SOC), European Inventory of Existing Commercial Chemical Substances (EINECS).References1. Gold et al (1999) (ref 4/ sect 9.2.)2. DSSTox (ref 5)3. Gold L.S. and Zeiger E., 1997 (ref 6)4. Carcinogenic Potency Database (CPDB) (ref 7)

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Neural network

4.2.Explicit algorithm:

Neural network
Nonlinear QSAR: Backpropagation Neural Network (Multilayer Perceptron) regression
The algorithm is based on neural network predictor with structure 9-9-8-1.
The algorithm is given in th eANN.snn file. In order to be used the user must have Statistica 7 or higher with ANN modules to make predictions.

4.3.Descriptors in the model:

[1]Highest e-n attraction (AM1) for N - N bonds
[2]Highest e-e repulsion (AM1) for N - O bonds
[3]Lowest e-n attraction (AM1) for N - O bonds
[4]Max Sigma-Pi bond order (AM1)
[5]Polarity parameter (AM1)
[6]Tot molecular 1-center E-E repulsion (AM1)
[7]DPSA2 Difference in CPSAs (PPSA2-PNSA2) (Zefirov)
[8]Max electrophilic reactivity index (AM1) for C atoms
[9]Kier shape index (order 2)

4.4.Descriptor selection:

Initial pool of ~1000 descriptors. Stepwise descriptor selection based on a set of statistical selection rules as F statistic and p. The highest F (low p) descriptors (9) were selected from the whole (~1000) descriptors taking into account also their value distribution. These 7 descriptors were used as inputs to the network. 21 networks with different structures were tested in order to find the best ANN with lowest RMS (root-mean-squared error). Then 152 epochs were used to train the final network with architecture depicted in 4.2. Optimization of the weights was performed with Levenberg-Marquardt algorithm using logistic activation function.

4.5.Algorithm and descriptor generation:

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

4.6.Software name and version for descriptor generation:

QSARModel
http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

37.7 (377 chemicals/ 10 discriptors)

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain based on training set:

By descriptor value range (between min and max values): The model is suitable for compounds that have the descriptors in the following range augmented with the confidence in 5.2:

Desc ID (See 4.3)
1 2 3 4 5 6 7 8 9
Min -441.347 0.0000 -546.667 0.000000 0.038500 286.705 19.680 0.00000 0.00000

Max 0.000 267.1622 0.000 0.096778 4.041800 6136.148 2534.643 0.06020 25.55286

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%.

If for any compound whose descriptors falls in the interval [|min|-0.3|min|; |max|+0.3|max|], then the ANN model is applicable and the prediction is reliable. Where min and max are the values in 5.1.

5.3.Software name and version for applicability domain assessment:

QSARModel 3.3.8
http://www.molcode.com

5.4.Limits of applicability:

See 5.2

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: 260

6.6.Pre-processing of data before modelling:

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

6.7.Statistics for goodness-of-fit:

	Training log(TD50)	Selection log(TD50)	Test log(TD50)
Data Mean	-1.334	-1.487	-0.961
Data S.D.	1.411	1.410	1.355
Error Mean	0.022	-0.187	-0.029
Error S.D.	0.986	1.074	1.144
Abs E. Mean	0.728	0.791	0.926
S.D. Ratio	0.699	0.762	0.844
Correlation	0.716	0.657	0.550

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

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:

RMS (Training) = 0.108, RMS (Selection) = 0.119, RMS (Test) = 0.125, See 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 randonly selected validation sets – selection (40) and test(40)

7.6.Experimental design of test set:

Randomly selected 40 and 40 data points for selection and test set, respectively

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, this set has not been used in the training set, it can be also considered as external set) 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.Ratio) and RMS error, see 6.7 and 6.12

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The mechanistic picture of the model is complicated due to the mathematical nature of the ANN(artificial neural network). However, it is known that carcinogenicity is greatly related to N-containing compounds. In our case the descriptors Highest e-n attraction (AM1) for N - N bonds, Highest e-e repulsion (AM1) for N - O bonds and Lowest e-n attraction (AM1) for N - O bonds reflect this fact. For instance the higher is the Highest e-n attraction (AM1) for N - N bonds the larger is Log DT50. In other words lower attraction between N-N atoms makes easier the donation of elections or formation of hydrogen bonds which will increase the carcinogenicity. Also other variabilities of these interactions are reflected by O-N bonds. In addition to the above charged surface areas of the compounds and their structural characteristics contribute to the phenomenon under investigation

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation, consistent with published scientific interpretations of experiments.

8.3.Other information about the mechanistic interpretation:

The model interpretation is consistent with some published results [e.g. Morales et al, 2006].

9.Miscellaneous information

9.1.Comments:

Supporting information for: Training set(s), Selection set(s), Test set(s), ANN.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.

The methodology and software (QSARModel) used to create the present model were applied also to obtain the results published in these papers: Katritzky et al. (2009), Karelson et al (2006)

9.2.Bibliography:

[1]Katritzky AR, Dobchev DA, Fara, C, Hur E, Tämm K, Kurunczi L, Karelson M, Varnek A & Solov'ev VP (2006). Skin Permeation Rate as a Function of Chemical Structure. Journal of Medicinal Chemistry 49, 3305-3314.

[2]Karelson M, Dobchev DA, Kulshyn OV & Katritzky A (2006). Neural Networks Convergence Using Physicochemical Data. Journal of Chemical Information and Modeling 46, 1891-1897.

[3]OECD Test Guideline 451 (EU Test Guideline B.32)

http://oberon.sourceoecd.org/vl=3544705/cl=34/nw=1/rpsv/ij/oecdjournals/1607310x/v1n4/s57/p1

[4]Gold LS, Manley NB, Slone TH & Rohrbach L (1999). Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature in 1993–1994 and by the National Toxicology Program in 1995–1996. Environmental Health Perspectives, Suppl. 107, (Suppl. 4) 527–602.

[5]Distributed Structure-Searchable Toxicity (DSSTox) database of the US Enviromental Protection Agency (EPA) http://www.epa.gov/ncct/dsstox/sdf_isscan_external.html

[6]Gold LS & Zeiger E (1997). Handbook of Carcinogenic Potency and Genotoxicity Databases. CRC Press, Boca Roca, FL.

[7]Carcinogenic Potency Database (CPDB) http://potency.berkeley.edu/cpdb.html

[8]Morales AH, Pérez MAC, Combes RD & Pérez González M (2006). Quantitative structure activity relationship for the computational prediction of nitrocompounds carcinogenicity. Toxicology 220, 51-62.

9.3.Supporting information:

Training set(s)

Mouse_carc_20100429_trainingset	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf225_Mouse_carc_20100429_trainingset.sdf

Test set(s)

Mouse_carc_20100429_testset	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf225_Mouse_carc_20100429_testset.sdf
Mouse_carc_20100429_selectionse	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf225_Mouse_carc_20100429_selectionset.sdf

10.Summary (JRC Inventory)

10.1.QMRF number:

Q17-10-1-225

10.2.Publication date:

2010/07/16

10.3.Keywords:

Molcode, Nonlinear QSAR, aritficial neural network, mouse carcinogenicity, Mouse

10.4.Comments: