	QMRF identifier (JRC Inventory): *To be entered by JRC*
	QMRF Title: *Nonlinear classification ANN QSAR Model for prenatal developmental toxicity study*
	Printing Date: *8.03.2011*

1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear classification ANN QSAR Model for prenatal developmental toxicity study

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:

12.05.2010

2.2.QMRF author(s) and contact details:

Dimitar Dobchev, Tarmo Tamm, Gunnar Karelson, Indrek Tulp, Kaido Tämm, Jaak Jänes, Eneli Härk, 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):

8.3.2011

2.4.QMRF update(s):

This is an update of previously developed ANN model from 12.05.2010. The current model has been rebuilt and improved using the same data source.

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]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.
[3]Statistica 7 www.statsoft.com

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:

See 2.4

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Animal, human

3.2.Endpoint:

4.Human health effects 4.16.In vivo pre-natal-developmental toxicity

3.3.Comment on endpoint:

Determination of evidence of prenatal development toxicity

The prenatal developmental toxicity was determined using the OECD test guideline 414 . This guideline for developmental toxicity testing is designed to provide general information concerning the effects of prenatal exposure on the pregnant test animal and on the developing organism; this may include assessment of maternal effects as well as death, structural abnormalities, or altered growth in the foetus [1].

The experiments are costly and time-consuming. An alternative to animal experiments is the derivation of structure–activity relationships (SARs) which are computational models that relate the biological activity of chemical agents to chemical structure. The measures of chemical structure are generally physical and chemical attributes or structural features; the biological activity is typically the laboratory experiment, usually in animals, which indicates the presence or absence of toxicity.

A database for computational analysis of developmental toxicity was created by combining subsets of information from the Teratogen Information System (TERIS) and the Food and Drug Administration (FDA) guidelines. Both sources are evaluations of the existing human and animal data on potentially teratogenic chemicals, which physicians used for reference. The TERIS compilation is skewed toward a complete evaluation of the animal data whereas the FDA discussion emphasizes human studies or case reports, with reference to related animal studies.

The database was constructed by researchers in the Department of Environmental and Occupational Health at the University of Pittsburgh and it contains information about 292 chemicals and their disposition as to development toxicity. Forty-one percent, 116 chemicals were deemed to be developmentally active substances, while the remaining 176 chemicals showed no evidence of developmental toxicity [2-5].

3.4.Endpoint units:

n/a

3.5.Dependent variable:

Presence or absence of toxicity (indicated as Tox Index +1 and -1)

3.6.Experimental protocol:

The test substance is administered daily from implantation to the day prior to scheduled caesarean section. Dose levels are selected taking into account any existing toxicity data as well as additional information on metabolism and toxicokinetics of the test substance or related materials. The females are mated with males of the same species and strain. Immediately after termination or as soon as possible after death, the uteri are removed and the pregnancy status of the animals ascertained. The evaluation of study includes the following information:

Maternal toxic response data by dose

Developmental endpoints by dose for litters with implants, including:

- number of corpora lutea;
- number of implantations, number and percent of live and dead foetuses and resorptions;
- number and percent of pre- and post-implantation losses.

Developmental endpoints by dose for litters with live foetuses, including:

- number and percent of live offspring;
- sexratio;
- foetal body weight, preferably by sex and with sexes combined;
- external, soft tissue, and skeletal malformations and other relevant alterations; - criteria for categorisation if appropriate; - total number and percent of foetuses and litters with any external, soft tissue, or skeletal alteration, as well as the types and incidences of individual anomalies and other relevant alterations.

3.7.Endpoint data quality and variability:

Experimental data is a compilation of different databases.

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 neural network predictor with structure 8-8-1
Standard Backpropagation Neural Network (Multilayer Perceptron) classification

4.3.Descriptors in the model:

[1]Highest coulombic interaction (AM1) [eV]
[2]PPSA3 Atomic charge weighted PPSA (Zefirov) [Å2*au]
[3]Charged (Zefirov) Surface Area of H atoms [Å2]
[4]Gravitation index (all bonds) (AM1) [amu2/Å2]
[5]HA dependent HDCA-2/SQRT(TMSA) (AM1) [Å2]
[6]Kier flexibility index [unitless]
[7]Molecular weight [g/mol]
[8]Relative number of N atoms [unitless]

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 first highest F (low p) descriptors (8) were selected from the whole (~1000) descriptors. These 8 descriptors were used as inputs to the network. 26 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 150 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 linear(inputs) and hyperbolic(hidden) adn logistic(output) 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:

29

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 (including ethers, esters, amides, amines,halides, aromatic, aliphatic functional groups) 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
Desc 1 2 3 4 5 6 7 8
Min 6.11 3.24 3.24 224.18 0.00 1.13 46.07 0.00
Max 13.80 28.87 28.83 4313.03 0.42 15.21 627.94 0.31

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 augmented by ±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 5.2

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: 181

6.6.Pre-processing of data before modelling:

6.7.Statistics for goodness-of-fit:

Training negatives Training positives Selection negatives Selection positives Test negatives Test positives
Total105.0076.0025.0015.0023.0017.00
Correct100.0068.0020.0012.0016.0012.00
Wrong5.008.005.003.007.005.00
Correct(%)95.2489.4780.0080.0069.5770.59
Wrong(%)4.7610.5320.0020.0030.4329.41

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)= 0.28
, RMS(Selection)= 0.44
, RMS(Test) = 0.54

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

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 over fit it) and the test set (used to test the external prediction of the net after training) are quite high according to the classification matrix, see 6.7.

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 nature of the ANN(artificial neural network). In addition, it cannot be given easy interpretation from mechanistic viewpoint because of the for the classification of the property in two categories. However, it can be concluded that model descriptors are related to structural, electrostatic and hydrogen donor/acceptor ability of the compounds. Moreover, one of the most significant descriptor in

the model

Highest coulombic interaction (AM1) indicates that the compounds with minimal values are in most cases toxic. An opposite trend can aslo be observed for PPSA3 Atomic charge weighted PPSA (Zefirov) descriptor i.e.

larger values would lead to more toxic compounds. The charged surface area of H atoms roughly indicates that lower values lead to non toxic compounds. The same trend holds for HA dependent HDCA-2/SQRT(TMSA) (AM1)

descriptor. However, such interpretations should always be considered together with thhe remaining descriptors.

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)

9.2.Bibliography:

[1]Prenatal developmental toxicity study, OECD TG 414, 2001.

[2]Arena V. C., Sussman N. B., Mazumdar S., Yu S., Macina O. T. The Utility of Structure-Activity Relationship (SAR) Models for Prediction and Covariate Selection in Developmental Toxicity: Comparative Analysis of Logistic Regression and Decision Tree Models. SAR and QSAR in Environmental Research, 15(1), 2004, 1-18.

[3]Ghanooni, M., Mattiston, D.R., Zhang, Y.P., Macina, O.T., Rosenkranz, H.S. and Klopman, G. Structural determinants associated with risk of human developmental toxicity. Am. J. Obstet. Gynecol. 176, 1997, 799–806.

[4]Briggs, G.G., Freeman, R.K. and Yaffe, S.J. Drugs in Pregnancy and Lactation, 3rd Edn. Williams and Wilkens, Baltimore, MD, 1990, p. 537.

[5]Shepard, T.H. Catalog of Teratologic Agents, 5th Edn. Johns Hopkins University Press, Baltimore, MD, 1992, p. 534.

[6]V. C. Arena; N. B. Sussman; S. Mazumdar; S. Yu; O. T. Macina, SAR and QSAR in Environmental Research, Vol. 15 (1), 2004, pp. 1–18 http://dx.doi.org/10.1080/1062936032000169633

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