QMRF	*QMRF identifier (JRC Inventory):To be entered by JRC* *QMRF Title:* *Nonlinear ANN QSAR Model for Skin sensitization (GPMT)* *Printing Date:4.04.2011*	QMRF

1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear ANN QSAR Model for Skin sensitization (GPMT)

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:

10.10.2010

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):

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:

[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:

No other QMRF available for the same model

3.Defining the endpoint - OECD Principle 1

3.1.Species:

guinea pigs

3.2.Endpoint:

4.Human health effects 4.6.Skin sensitisation

3.3.Comment on endpoint:

The skin senzitation effects of chemicals can be assessed by ranking procedure using guinea pig maximization test (GPMT) index. Preliminary tests are performed to identify the maximum non-irritating dose.

3.4.Endpoint units:

-

3.5.Dependent variable:

GPMT Score ss

3.6.Experimental protocol:

The Guinea-Pig Maximazation Test (GPMT) was determined using the EU Test Guideline B.6 (OECD TG 406).

The GPMT is a widely used adjuvant-type test. Although several other methods can be used to detect the potential of a substance to provoke skin sensitisation reaction, the GPMT is considered to be the preferred adjuvant technique. The test animals are initially exposed to the test substance by intradermal injections and/or epidermal application (induction exposure). Following a rest period of 10 to 14 days (induction period), during which an immune response may develop, the animals are exposed to a challenge dose. The extent and degree of skin reaction to the challenge exposure in the test animals is compared with that demonstrated by control animals which undergo sham treatment during induction and receive the challenge exposure.

The GPMT data were compiled from Cronin and Basketter [1] and Devillers [2]. A database containing experimental GPMT data for 258 chemical compounds was used for modeling skin sensitization. The molecules were categorized as being non-sensitizers, weak, moderate, and strong sensitizers, and toxicity scores of 0, 0.33, 0.66, and 1 were assigned, respectively. Molecules obtained from Unilever that are classified as being strong were given a score of 0.83 (average of moderate and strong) because the database classifies both moderate and strong sensitizers as being strong

3.7.Endpoint data quality and variability:

Source data is a compilation of several experimental datasets, based on GLP experiments following the OECD 406 guideline [1,2]

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 7-7-7-1.

4.3.Descriptors in the model:

[1]HOMO - LUMO energy gap (AM1) [eV]
[2]Lowest resonance energy (AM1) for C - H bonds [eV]
[3]Highest n-n repulsion (AM1) for C - H bonds [eV]
[4]Number of rings [unitless]
[5]Lowest exchange energy (AM1) for C - C bonds [eV]
[6]Highest coulombic interaction (AM1) for C - H bonds [eV]
[7]Max nucleophilic reactivity index (AM1) for O atoms [1/eV]

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 (7) were selected from the whole (976) descriptors. These 7 descriptors were used as inputs to the network. 18 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 791 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:

22.6

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

By descriptor value range (between min and max values): The model is suitable for compounds (including ethers, esters, amides, halides, aromatic, aliphatic functional groups etc) that have the descriptors in the following range augmented with the confidence in 5.2:

Desc ID	1	2	3	4	5	6	7
Min	7.037	-11.246	39.220	0.000	-10.356	3.575	0.000
Max	14.701	-10.496	40.405	5.000	-4.686	4.282	0.051

5.2.Method used to assess the applicability domain:

presence of functional groups in structures see also 5.1

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: 158

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:

	Training GPMT	Selection GPMT	Test GPMT
Data Mean	0.509	0.542	0.512
Data S.D.	0.408	0.424	0.416
Error Mean	0.000	-0.091	0.073
Error S.D.	0.134	0.701	0.526
Abs E. Mean	0.097	0.504	0.368
S.D. Ratio	0.328	1.653	1.266
Correlation	0.945	0.313	0.631

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:

RMS (Training)= 0.133706, RMS(Selection)= 0.706883, RMS(Test) = 0.531382,

In this ANN were used 2 sets randomly chosen (50) 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 (50) and test (50)

7.6.Experimental design of test set:

Randomly selected 50 and 50 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 RMS error and the standard deviation ratio (S.D.Ration), see 6.7 and 6.12

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The complex nature of the ANN model does not allow direct interpretation of the descriptors addressed to the property. Skin sensitization is believed to be underpinned by mechanisms based on chemical reactivity (with the chemical behaving as an electrophile), in most cases binding covalently to a skin protein leading it to becoming antigenic. It has been agreed that the key to predicting likely

sensitization potential is being able to predict electrophilic reactivity and proelectrophilicity. The present model includes a number of chemical reactivity descriptors accounting for these effects.However, it can be noted that descriptors HOMO - LUMO energy gap (AM1) and Lowest resonance energy (AM1) for C - H bonds lead to decreased GPMT index with increasing to their values.

8.2.A priori or a posteriori mechanistic interpretation:

a priori/a posteriori

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)
7-7-7-1.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]Cronin M.T.D., Basketter D.A. Multivariate QSAR analysis of a skin sensitization database. SAR and QSAR in Environmental Research 2, 159–179, 1994.
[2]Devillers J. A neural network SAR model for allergic contact dermatitis. Toxicology Methods 10,181–193, 2000.

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