| | | |
|---|---|---|
| QMRF | *QMRF identifier (ECB Inventory):* To be entered by JRC<br>*QMRF Title:* Nonlinear Classification ANN QSAR Model for in vitro chromosomal aberration data in mammalian cells<br>*Printing Date:* Jun 7, 2010 | QMRF |

## 1.QSAR identifier

### 1.1.QSAR identifier (title):

Nonlinear Classification ANN QSAR Model for in vitro chromosomal aberration data in mammalian cells

### 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:

4.06.2010

### 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]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:**

Chinese Hamster Lung Cells

**3.2.Endpoint:**

4.Human health effects 4.10.Mutagenicity -IN VITRO MAMMALIAN CHROMOSOME ABERRATION TEST

**3.3.Comment on endpoint:**

Description of the in vitro chromosome aberration test

The test system and its purpose are described in OECD Guideline for the Testing of chemicals, No. 473 (1).

"The purpose of the in vitro chromosome aberration test is to identify agents that cause structural chromosome aberrations in cultured mammalian cells. Structural aberrations may be of two types, chromosome or chromatid. With the majority of chemical mutagens, induced aberrations are of the chromatid type, but chromosome-type aberrations also occur. An increase in polyploidy may indicate that a chemical has the potential to induce numerical aberrations. However, this guideline is not designed to measure numerical aberrations and is not routinely used for that purpose. Chromosome mutations and related events are the cause of many human genetic diseases and there is substantial evidence that chromosome mutations and related events causing alterations in oncogenes and tumour suppressor genes of somatic cells are involved in cancer induction in humans and experimental animals."

**3.4.Endpoint units:**

-

**3.5.Dependent variable:**

Chromosome aberration (CA) values -1, 1( or NEG, POS)

**3.6.Experimental protocol:**

All tests were performed using a Chinese Hamster Lung Cell (CHL) fibroblast cell line, which has been kept as a single cell sub-clone since 1973. This cell line has been used almost exclusively in Japan to test hundreds of chemicals over more than two decades, as opposed to the Chinese Hamster Ovary (CHO) cell lines that are more common in Europe and the United States. Much of the test information has been published in numerous scientific articles during the years over which it has been generated. An example is provided by Ishidate et al. (4).

**3.7.Endpoint data quality and variability:**

The test data used in this model were taken from a single source, the Data Book of Chromosomal Aberration Test In Vitro (2). This book is written in Japanese, but all tables are in English and the authors were provided with English translations for everything except the Introduction. The Introduction is identical to that used in the previous version of the book, published in English by Dr. Motoi Ishidate (3), which was also available to the authors.

Test results for a total of 901 substances are presented in the Data Book (2). The chemicals were chosen for a variety of reasons, including use in foods. A number fall into the class commonly referred to as UVCB's, or chemicals that cannot be represented by a complete structure diagram and specific molecular formula. These were excluded for the obvious reason that it is impossible to model a chemical for which a structure is not available. However, it was found that this is not always a totally unambiguous process, so the authors made the best judgement they could. Inorganic chemicals were also excluded, as the modeling platform used by the authors cannot deal with them. A very small number of chemicals were excluded because the true identity was not clear (inconsistencies between chemical name, CAS number and structure/molecular weight that we were unable to resolve). A few stereo-isomers with conflicting results were also removed as they cannot be distinguished by SMILES notation (a computer code for 2D structures).

A toxicological decision was made to include chemicals as being positive if they were active in inducing either aberrations or polyploidy. While the current test guideline does not specify testing for a length of time, which would allow polyploidy to be assessed, much of the CHL data does and the information was felt to be too valuable to lose (18 chemicals). Chemicals were also retained even if the test had not been performed both in the presence and absence of metabolic activation.

Beyond this, the judgement of the authors was used in their interpretation of the final test result. This included dropping 16 of 18 chemicals that the authors considered inconclusive in repeat tests (two were kept because while they were inconclusive for polyploidy, they were clearly positive for structural aberrations).

Seventy-eight chemicals were excluded because the authors considered them False Positive (only active at dose of more than 10 mM where effects could be due to osmotic pressure).

As the modeling system was not able to handle salts (e.g. sodium salts, hydrochlorides), further interpretation was necessary. In the majority of cases there was no conflict with regard to results of testing ionised or non-ionised forms. However, in certain cases there were. The authors decided that for some simple organic acids that were active but where the salt was clearly inactive, to consider these as being inactive in accordance with the advice, given in the OECD Guidelines and Morita et al. (5), that particularly low pH may lead to false positive predictions. It is not known if this decision is right or wrong in relation to use of results of this in vitro system for predicting in vivo effects, but it will clearly affect the performance of the model.

A few decisions have been done on a basis of additional data from the literature: vitamin B2 (Riboflavin, CAS 83-88-5) tested positive in insoluble form, but was negative in soluble form. The negative result was retained, as the mechanism for the insoluble compound appears to be physical (6). After some consideration, saccharin (CAS 81-07-2) and EDTA (CAS 60-00-4) were entered as negatives, in agreement with Ashby et al. (7), even though there was conflicting information for some of the salts.

Finally, about 40 chemicals having only equivocal results were excluded. This is also an arbitrary decision, but it was felt that equivocal results were not likely to lead to a better training set.

Thus, a total of 513 chemicals remained. Their identities and SMILES notations are available in Training_set.doc. There were 263 positive and 250 negative substances in the training set, giving the nearly 50:50 split considered ideal for modeling purposes.

For external validation, data generated over a six-year period (1991-1996) was used for chromosomal aberration testing of high production volume (HPV) industrial chemicals that had been conducted using Chinese hamster lung (CHL/IU) cells according to the OECD HPV testing program and the national program in Japan (Kusakabe et al., 8)

Of a total of 98 substances, two were removed in the authors' analyses: dicyclopentadiene (CAS 77-73-6), because it was already in the training set, and Pigment Green No. 7 (CAS 14832-145), a copper complex that cannot be modeled in the selected system. The 98 chemicals are available in Validation_set.doc. On further examination of the data set, it was noticed that one substance (4-(1-Methylpropyl)phenol, CAS 99-71-8) was actually a false positive (only active at very high concentration, and ultimately judged inactive following an in vitro micronucleus test). Eight additional chemicals were identified where the chromosomal aberrations are induced under non-physiological culture conditions (pH<6), which could be kept in mind when using the data.

## 4.Defining the algorithm - OECD Principle 2

### 4.1.Type of model:

Nonlinear QSAR: Backpropagation Neural Network (Multilayer Perceptron) classification

### 4.2.Explicit algorithm:

The algorithm is based on neural network predictor with structure 9-9-8-1
Standard Backpropagation Neural Network (Multilayer Perceptron) classification

### 4.3.Descriptors in the model:

[1]Square root of Partial Surface Area of H atoms
[2]Partial Surface Area of H atoms
[3]HOMO - LUMO energy gap (AM1)
[4]No. of occupied electronic levels (AM1) / # atoms
[5]WFOSA Atomic charge (Zefirov) weighted FOSA
[6]Highest exchange energy (AM1) for C - C bonds
[7]Number of H atoms
[8]DPSA1 Difference in CPSAs (PPSA1-PNSA1) (AM1)
[9]Max Sigma-Sigma bond order (AM1)

### 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 (9) were selected from the whole (~1075) descriptors. These 9 descriptors were used as inputs to the network. 12 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 1998 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. The cost function was Entropy function.

**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 3.3.8
Molcode Ltd. Turu 2, Tartu, 51014, Estonia
http://www.molcode.com

**4.7.Chemicals/Descriptors ratio:**

66

## 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, halides, aromatic, aliphatic functional groups etc) that have the descriptors in the following range augmented with the confidence in 5.2:

Desc ID
See 4.3123456789
Min0.0000000.0000001.257470.9714290.00000-10.08720.00000-228.998 0.590701
Max0.2372280.97832514.615912.90000025.565230.000067.00000791.387 0.930916

**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
Molcode Ltd. Turu 2, Tartu, 51014, Estonia
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: 501 (initial set was refined: salts and equivocal exp values were removed)

**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 negativesTraining positivesSelection negativesSelection positivesTest negativesTest positives
Total242.0000259.000019.0000031.0000023.0000027.00000
Correct233.0000252.000013.0000022.0000013.0000018.00000
Wrong9.00007.00006.000009.0000010.000009.00000
Correct(%)96.281097.297368.4210570.9677456.5217466.66667
Wrong(%)3.71902.702731.5789529.0322643.4782633.33333

**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:**

See 6.7 for classification statistics

## 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 (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

**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 very good 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 is difficult to analyze because of the nature of the ANN models. According to the descriptors used as inputs to the network, it can be concluded that the property is mainly related to the charged surfaces that may play important role in defining the property values. For instance, the most significant descriptor (according to F) Square root of Partial Surface Area of H atoms leads to positive index of the chromosomal aberration when its values are lower.

In addition to the charged surfaces, hydrogen abilities of the compounds are also important in conjunction with the energy terms related to HOMO-LUMO and exchange interactions for the C-C bond.

**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:**

## 9.Miscellaneous information

**9.1.Comments:**

Supporting information for :Training set(s)
Selection set(s)
Test set(s)
9-9-8-1.snn file (binary) -includes the ANN model, in order to be used the user must have Statistica 7 or higher with ANN modules

**9.2.Bibliography:**

[1]OECD (1997). OECD Guidelines for the Testing of Chemicals No. 473: Genetic Toxicology: In Vitro Mammalian Cytogenetic Test. Organisation for Economic Cooperation and Development; Paris, France.
[2]Sofuni, T., Ed. (1998). Data Book of Chromosomal Aberration Test In Vitro, Revised Edition.. Life-Science Information Center; Tokyo, Japan.
[3]Ishidate, Motoi Jr., Ed. (1988). Data Book of Chromosomal Aberration Test In Vitro, Revised Edition. Elsevier; Amsterdam, New York, Oxford.
[4]Ishidate, M. Jr., Haronois, M.C. & Sofuni, T. (1988). A Comparative analysis of data on the clastogenicity of 951 chemicals tested in mammalian cell cultures. Mutation Research 195, 151-213.
[5]Morita, T., Nagaki, T., Fukuda, I. & Okumura, K. (1992). Clastogenicity of low pH to various cultures mammalian cells. Mutation Research 268, 297-305.
[6]Kawaguchi, Y., Hayashi, H., Sato, M. & Shindo, Y. (1997). Needle crystals of Vitamin B2 induce polyploidy in Chinese hamster lung (CHL/IU) cells. Mutation Research 373, 1-7.
[7]Ashby, J. & Ishidate, M. Jr. (1986). Clastogenicity in vitro of the Na, K, Ca and Mg. Salts of Saccharin; and of magnesium chloride; consideration of significance. Mutation Research 163, 63-73.
[8]Kusakabe, H., Ymakage, K., Wakuri, S., Sasaki, K., Nakagawa, Y., Watanabe, M., Hayashi, M., Sufuni, T., Ono, H. & Tnanka, N. (2002). Relevance of chemical structure and cytotoxicity to the induction of chromosome aberrations based on testing of 98 high production volume industrial chemicals. Mutation Research 517, 187-198.

### 9.3.Supporting information:

Training set(s)

| | |
|---|---|
| Chromosomal_Aberration_trainingset.sdf | http://qsardb.jrc.ec.europa.eu:80/qmrf/download_attachment.jsp?name=qmrf212_Chromosomal_Aberration_trainingset.sdf |

Test set(s)

| | |
|---|---|
| Chromosomal_Aberration_testset.sdf | http://qsardb.jrc.ec.europa.eu:80/qmrf/download_attachment.jsp?name=qmrf212_Chromosomal_Aberration_testset.sdf |

Supporting information

| | |
|---|---|
| Chromosomal_Aberration_selectionset.sdf | http://qsardb.jrc.ec.europa.eu:80/qmrf/download_attachment.jsp?name=qmrf212_Chromosomal_Aberration_selectionset.sdf |
| 9-9-8-1.snn | http://qsardb.jrc.ec.europa.eu:80/qmrf/download_attachment.jsp?name=qmrf212_9-9-8-1.snn |

## 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