	QMRF identifier (JRC Inventory): **To be entered by ECB**
	QMRF Title: **CAESAR model for carcinogenicity based on Counter Propagation Neural Network.**
	Printing Date: **1-feb-2018**

1.QSAR identifier

1.1.QSAR identifier (title):

CAESAR model for carcinogenicity based on Counter Propagation Neural Network.

1.2.Other related models:

n/a

1.3.Software coding the model:

CAESAR Application.

The model has been developed within a dedicated framework, freely available on-line, together with the others CAESAR models. The model has been developed in JAVA and its code is open-source.
http://www.caesar-project.eu/software/

Vega (Virtual Models for evaluating the properties of chemicals within a global architecture, version 1.1.4, 2017)

VEGA provides tens of QSAR models to predict tox, ecotox, environ, and phys-chem properties of chemical substances.

https://www.vegahub.eu/contacts/

https://www.vegahub.eu/

2.General information

2.1.Date of QMRF:

February 2018

2.2.QMRF author(s) and contact details:

[1]Manuela Pavan S-In Soluzioni Informatiche Soluzioni Informatiche Srl Via Ferrari 14, I-36100 Vicenza manuela.pavan@s-in.it http://www.s-in.it

[2]Simona Kovarich S-In Soluzioni Informatiche Soluzioni Informatiche Srl Via Ferrari 14, I-36100 Vicenza simona.kovarich@s-in.it http://www.s-in.it

2.3.Date of QMRF update(s):

n/a

2.4.QMRF update(s):

n/a

2.5.Model developer(s) and contact details:

[1]Natalja Fjodorova National Institute of Chemistry, Hajdrihova 19, SI-1001 Ljubljana, Slovenia

[2]Marjan Vrako National Institute of Chemistry, Hajdrihova 19, SI-1001 Ljubljana, Slovenia

[3]Marjana Novi National Institute of Chemistry, Hajdrihova 19, SI-1001 Ljubljana, Slovenia

2.6.Date of model development and/or publication:

The model was published in 2010 (see Fjodorova et al. in bibliography 9.2)

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

Natalja Fjodorova, Marjan Vrako, Marjana Novi, Alessandra Roncaglioni and Emilio Benfenati. New public QSAR model for carcinogenicity. Chemistry Central Journal 2010, 4(Suppl 1):S3doi:10.1186/1752-153X-4-S1-S3 http://www.journal.chemistrycentral.com/content/4/S1/S3

2.8.Availability of information about the model:

The model has been released open source and is available through the VEGA HUB portal. The training and test set are available.

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

Not to date

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Rats.

3.2.Endpoint:

QMRF 4.12. Carcinogenicity OECD 451 Carcinogenicity Studies

3.3.Comment on endpoint:

Carcinogenicity is a very complex biochemical phenomenon involving processes at the cellular level. The carcinogenicity of a substance depends on its molecular structure and a certain number of phenomena which are only partially known. Typically, the carcinogenic process involves one or more processes, showing a relationship with the mutagenic potential of a substance, but other processes are possible for carcinogens which are non mutagenic.

3.4.Endpoint units:

adimensional

3.5.Dependent variable:

carcinogenic/non carcinogenic, based on the rat carcinogenicity

3.6.Experimental protocol:

Experimental values for carcinogenicity were taken from http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html.

3.7.Endpoint data quality and variability:

The DSSTox database refers to the L.Gold database and is considered of good quality.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR model based on CP ANN

4.2.Explicit algorithm:

Counter Propagation Artificial Neural Network (CP ANN)

CP ANN consists of two layers of neurons arranged in a two-dimensional rectangular matrix. For more information see Fjodorova et al. 2010 [1]

4.3.Descriptors in the model:

[1]PW5 path/walk 5 - Randic shape index

[2]D/Dr06 distance/detour ring index of order 6

[3]MATS2p Moran autocorrelation - lag 2 / weighted by atomic polarizabilities

[4]EEig10x Eigenvalue 10 from edge adj. matrix weighted by edge degrees

[5]ESpm11 Spectral moment 11 from edge adj. matrix weighted by edge degrees

[6]ESpm09 Spectral moment 09 from edge adj. matrix weighted by dipole moments

[7]GGI2 topological charge index of order 2

[8]JGI6 mean topological charge index of order6

[9]nRNNOx number of N-nitroso groups (aliphatic)

[10]nPO4 number of phosphates / thiophosphates

[11]N-067 number of Al2-NH atom centered fragments

[12]N-078 number of Ar-N=X / X-N=X atom centered fragments

4.4.Descriptor selection:

Descritpor selection was performed using cross correlation matrix,multicolinearity and fisher ratio techniques. As a result descriptors space was reduced from 835 to 12 descriptors listed in 4.3.

4.5.Algorithm and descriptor generation:

The descriptors were calculated, in the original model, by means of dragonX software and are now entirely calculated by an in-house software module in which they are implemented as described in Todeschini&Consonni (2009) [2].

4.6.Software name and version for descriptor generation:

4.7.Descriptors/Chemicals ratio:

12 descriptors / 645 training chemicals = 0.019

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The model is applicable to heterogeneous organic chemicals. In the software implementation of the model several pieces of information are given to evaluate if a prediction is reliable (chemical falling in the Applicability Domain or not). The information about Applicability Domain (AD) is combined into a unique index called Global Applicability Domain Index (ADI). Global AD Index values range between 0 and 1. ADI > 0.8 means that the compound is in the AD of the model, ADI < 0.6 means that the compound is out of the AD of the model, a value between 0.6 and 0.8 means that the compound is possibly out of model AD.

5.2.Method used to assess the applicability domain:

Within Vega, the Applicability Domain of the model is defined by considering several parameters as described below:

1. Similar molecules with known experimental values: this index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation. Defined intervals are: 0-0.6 (no similar compounds with known experimental value in the training set have been found), 0.6-0.8 (only moderately similar compounds with known experimental value in the training set have been found, 0.8-1 (strongly similar compounds with known experimental value in the training set have been found).
2. Accuracy of prediction for similar molecules: this index takes into account the classification accuracy in prediction for the two most

similar compounds found. Values near 1 mean that the predicted compounds falls in an area of the model's space where the model gives reliable predictions (no misclassifications), otherwise the lower is the value, the worse the model behaves. Defined intervals are: 0-0.5 (prediction accuracy not adequate), 0.5-0.9 (prediction accuracy not optimal), 0.9-1 (good prediction accuracy).

3. Concordance for similar molecules: this index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are: 0-0.5 (similar molecules found in the training set have experimental values that disagree with the predicted value), 05-0.9 (some similar molecules found in the training set have experimental values that disagree with the predicted value), 0.9-1 (similar molecules found in the training set have experimental values that agree with the predicted value).
4. Atom Centered Fragments similarity check: this index takes into account the presence of one or more fragments that aren't found in the training set, or that are rare fragments. First order atom centered fragments from all molecules in the training set are calculated, then compared with the first order atom centered fragments from the predicted compound; then the index is calculated as following: a first index RARE takes into account rare fragments (those who occur less than three times in the training set), having value of 1 if no such fragments are found, 0.85 if up to 2 fragments are found, 0.7 if more than 2 fragments are found; a second index NOTFOUND takes into account not found fragments, having value of 1 if no such fragments are found, 0.6 if a fragments is found, 0.4 if more than 1 fragment is found. Then, the final index is given as the product RARE * NOTFOUND. Defined intervals are: index<0.7 (a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments), 0.7<= index < 1 (some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments), index = 1 (all atom centered fragment of the compound have been found in the compounds of the training set).
5. Model assignment reliability: this index checks if the two neural network output values (positive and non-positive) lead to an unreliable prediction; when the difference between these two values is lower than 0.1, the neuron where the predicted compound falls can not provide a good classification, thus the index is set to 0. Otherwise the index is set to 1.
6. Neural map neurons concordance: this index checks the concordance of the predicted compound with the experimental values of the other compounds that falls in the same neuron. The index is built considering two sub-indices: Population (the number of compounds found in the neuron) and Concordance (the number of compounds in the neuron

that have experimental value matching with current prediction divided by the number of compounds in the neuron). Low values mean that the predicted compound falls in a zone of the neural network that has no experimental compounds, or that has experimental compounds with heterogeneous experimental values, thus leading to a low reliability of the prediction. Index values: 0.5 (predicted substance falls into a neuron that is populated by no compounds of the training set), 0.75 (predicted value disagrees with experimental values of training set compounds laying in the same neuron), 1 (predicted value agrees with experimental values of training set compounds laying in the same neuron).7. Model descriptors range check: this index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range. Index values: True (descriptors for this compound have values inside the descriptor range of the compounds of the training set), False (descriptors for this compound have values outside the descriptor range of the compounds of the training set). Descriptors range: PW5: min = 0, max = 0,17; D/Dr06: min = 0, max =1114,64; MATS2p: min = -1, max = 1; EEig10x: min = -1, max = 3,93; ESpm11x: min = 0,69, max = 19,86; ESpm09d: min = 0, max = 15,48; GGI2: min = 0, max = 15,11; JGI6: min = 0, max = 0,05; nRNNOx: min= 0, max = 2; nPO4: min = 0, max = 2; N-067: min = 0, max = 2; N-078: min = 0, max = 4.

5.3.Software name and version for applicability domain assessment:

Vega (Virtual Models for evaluating the properties of chemicals within a global architecture, version 1.1.4, 2017)

The model has been developed within a dedicated framework, freely available on-line, together with the others CAESAR models. The model has been developed in JAVA and its code is open-source. https://www.vegahub.eu/

5.4.Limits of applicability:

Gobal AD Index intervals:

ADI < 0.6:predicted substance is out of the the Applicability Domain of the model

0.6<= ADI < 0.8: predicted substance could be out of the Applicability Domain of the model

AD >=0.8: predicted substance is into the Applicability Domain of the model

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: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

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:

The whole training set, which consists of 645 chemicals, is provided in supporting information (Training set.xls).

The training set structures are provided in supporting information (Training set.smi)

6.6.Pre-processing of data before modelling:

6.7.Statistics for goodness-of-fit:

Accuracy = 0.87; Specificity = 0.86; Sensitivity = 0.89

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

n/a

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

n/a

6.10.Robustness - Statistics obtained by Y-scrambling:

n/a

6.11.Robustness - Statistics obtained by bootstrap:

n/a

6.12.Robustness - Statistics obtained by other methods:

n/a

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: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

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 test set, which consists of 161 chemicals, is provided in supporting information (Test set.xls).

The test set structures are provided in supporting information (Test set.smi).

7.6.Experimental design of test set:

n/a

7.7.Predictivity - Statistics obtained by external validation:

Accuracy = 0.67; Specificity = 0.61; Sensitivity = 0.72

7.8.Predictivity - Assessment of the external validation set:

n/a

7.9.Comments on the external validation of the model:

n/a

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

n/a

8.2.A priori or a posteriori mechanistic interpretation:

n/a

8.3.Other information about the mechanistic interpretation:

n/a

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Natalja Fjodorova, Marjan Vrako, Marjana Novi, Alessandra Roncaglioni and Emilio Benfenati. New public QSAR model for carcinogenicity. Chemistry Central Journal 2010, 4(Suppl 1):S3doi:10.1186/1752-153X-4-S1-S3 http://www.journal.chemistrycentral.com/content/4/S1/S3

[2]R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009.

9.3.Supporting information:

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

10.Summary (ECB Inventory)

10.1.QMRF number:

To be entered by ECB

10.2.Publication date:

To be entered by ECB

10.3.Keywords:

To be entered by ECB

10.4.Comments:

To be entered by ECB