	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Carcinogenicity ISS model (version 1.0.3)**
	Printing Date: **3-06-2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Carcinogenicity ISS model (version 1.0.3)

1.2.Other related models:

The Carcinogenicity ISS model (version 1.0.3) is an implementation in VEGA software of the set of rules for carcinogenicity obtained from a decisional tree implemented in the ToxTree software

1.3.Software coding the model:

VEGA (https://www.vegahub.eu/)

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

emilio.benfenati@marionegri.it

2.General information

2.1.Date of QMRF:

03/06/2022

2.2.QMRF author(s) and contact details:

[1]Alessio Gamba Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy alessio.gamba@marionegri.it https://www.marionegri.it/

[2]Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy emilio.benfenati@marionegri.it https://www.marionegri.it/

[3] Erika Colombo Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy erika.colombo@marionegri.it https://www.marionegri.it/

2.3.Date of QMRF update(s):

NA

2.4.QMRF update(s):

NA

2.5.Model developer(s) and contact details:

[1] Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy alberto.manganaro@marionegri.it https://www.marionegri.it/

[2]Nina Jeliazkova IDEA Consult Joseph II straat 40 B1, 1000 Brussels, Belgiumjeliazkova.nina@gmail.com

[3]Romualdo Benigni Istituto Superiore di Sanità Department of Environment and PrimaryPrevention, Rome romualdo.benigni@iss.it

[4]Cecilia Bossa Istituto Superiore di Sanità Department of Environment and Primary Prevention,Rome cecilia.bossa@iss.it

2.6.Date of model development and/or publication:

2015

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

[1] R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, A. Worth "The Benigni/Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree" (2008) European Commission report EUR23241 EN

[2] R. Benigni, C. Bossa, T. Netzeva, A. Rodomonte, I. Tsakovska "Mechanistic QSAR of aromatic QMRF identifier (JRC Inventory):To be entered by JRC QMRF Title: Carcinogenicity ISS model (version 1.0.2)Printing Date:13 lug 20201.QSAR identifier amines: new models for discriminating between mutagens and non mutagens, and validation of models for carcinogens" (2007) Environ mol mutag 48:754-771

[3] R. Benigni, C. Bossa, O. Tcheremenskaia "Nongenotoxic carcinogenicity of chemicals: mechanisms of action and early recognition through a new set of structural alerts" (2013) Chemical Reviews 113(5), 2940-57

[4] A. Golbamaki, E. Benfenati, N. Golbamaki, A. Manganaro, E. Merdivan, A. Roncaglioni, G. Gini(2016) New clues on carcinogenicity-related substructures derived from mining two large datasets of chemical compounds. JOURNAL OF ENVIRONMENTAL SCIENCE AND HEALTH, PART C, VOL.34, NO. 2, 97-113

[5] Benfenati E, Manganaro A, Gini G. VEGA-QSAR: AI inside a platform for predictive toxicology

Proceedings of the workshop "Popularize Artificial Intelligence 2013", December 5th 2013, Turin, Italy

Published on CEUR Workshop Proceedings Vol-1107

2.8.Availability of information about the model:

The model is non-proprietary and the training set is available.

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

Another QMRF is not available.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Human

3.2.Endpoint:

TOX 7.7 Carcinogenicity (in vivo)

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:

The dependent variable is cancerogenic effect on rat, as binary classification: 0 (non-carcinogen), 1 (carcinogen)

3.6.Experimental protocol:

The details about the rule set compilation are reported on the ISS website

3.7.Endpoint data quality and variability:

NA

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Expert rule-based model

4.2.Explicit algorithm:

Decisional algorithm based on rules of toxicity.

Algorithm is based on a set of rules for carcinogenicity manually determined (expert system). If one of rules is in the target molecule, the prediction is 'Carcinogen', otherwise it is 'NON-Carcinogen'

4.3.Descriptors in the model:

The following 54 fragments are encoded as SA for matching carcinogen compounds:

SA1 Acyl halides

SA2 Alkyl (C<5) or benzyl ester of sulphonic or phosphonic acid

SA3 N-methylol derivativesSA4 Monohaloalkene

SA5 S or N mustard

SA6 Propiolactones and propiosultones

SA7 Epoxides and aziridines

SA8 Aliphatic halogens

SA9 Alkyl nitrite

SA10 alfa, beta unsaturated carbonyls

SA11 Simple aldehyde

SA12 Quinones

SA13 Hydrazine

SA14 Aliphatic azo and azoxy

SA15 Isocyanate and isothiocyanate groups

SA16 Alkyl carbamate and thiocarbamate

SA17 Thiocarbonyl (Nongenotoxic carcinogens)

SA18 Polycyclic Aromatic Hydrocarbons

SA19 Heterocyclic Polycyclic Aromatic Hydrocarbons

SA20 (Poly) Halogenated Cycloalkanes (Nongenotoxic carcinogens)

SA21 Alkyl and aryl N-nitroso groups

SA22 Azide and triazene groups

SA23 Aliphatic N-nitroSA24 alfa,beta unsaturated alkoxy

SA25 Aromatic nitroso groupSA26 Aromatic ring N-oxide

SA27 Nitro aromatic

SA28a Primary aromatic amine, hydroxyl amine and its derived esters (with restrictions)

SA28b Aromatic mono- and dialkylamine

SA28c Aromatic N-acyl amine

SA29 Aromatic diazo

SA30 Coumarins and Furocoumarins

SA31a Halogenated benzene (Nongenotoxic carcinogens)

SA31b Halogenated PAH (naphthalenes, biphenyls, diphenyls) (Nongenotoxic carcinogens)

SA31c Halogenated dibenzodioxins (Nongenotoxic carcinogens)

SA37 Pyrrolizidine Alkaloids

SA38 Alkenylbenzenes

SA40 Substituted phenoxyacid

SA41 Substituted n-alkylcarboxylic acids

SA42 Phthalate diesters and monoesters

SA43 Perfluorooctanoic acid (PFOA)

SA44 Trichloro (or fluoro) ethylene and Tetrachloro (or fluoro) ethylene

SA45 Indole-3-carbinol

SA46 PentachlorophenolSA47 O-phenylphenol

SA48 Quercetin-type flavonoids

SA49 Imidazole and benzimidazole

SA50 Dicarboximide

SA51 Dimethylpyridine

SA52 Metals, oxidative stress

SA53 Benzensulfonic ethers

SA54 1,3-Benzodioxoles

SA55 Phenoxy herbicides

SA56 Alkyl halides

4.4.Descriptor selection:

NA

4.5.Algorithm and descriptor generation:

The model has been built as a set of rules, taken from the work of Benigni and Bossa (ISS) as implemented in the software ToxTree version 2.6 (http://toxtree.sourceforge.net). The model implement all the rules related to carcinogenicity and does not implement the full decision tree used by ToxTree. If the given compound matches at least one carcinogen rule, the model gives a 'carcinogen' prediction, otherwise it gives a 'non-carcinogen' prediction. The training set for the model has been extracted from ToxTree, and consists of 797 compounds

4.6.Software name and version for descriptor generation:

NA

4.7.Chemicals/Descriptors ratio:

797/54 = 14.76

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The Applicability Domain (AD) is assessed using the original algorithm implemented within VEGA. An overall AD index is calculated, based on a number of parameters, which relate to the results obtained on similar chemicals within the training and test sets.

If 1 ≥ AD index ≥ 0.9, the predicted substance is into the Applicability Domain of the model. It corresponds to good reliability of prediction.

If 0.9 > AD index ≥ 0.65, the predicted substance could be out of the Applicability Domain of the model. It corresponds to moderate reliability of prediction.

If AD index < 0.65, the predicted substance is out of the Applicability Domain of the model and corresponds to low reliability of prediction.

5.2.Method used to assess the applicability domain:

The Applicability Domain and chemical similarity is measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [2]. The AD also evaluates the correctness of the prediction on similar compounds (accuracy), the consistency between the predicted value for the target compound and the experimental values of the similar compounds, the range of the descriptors, and the presence of unusual fragments, using atom centred fragments.

Similar molecules with known experimental value:

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:

If 1 ≥ index > 0.80, strongly similar compounds with known experimental value in the training set have been found.

If 0.80 ≥ index > 0.6, only moderately similar compounds with known experimental value in the training set have been found.

If index ≤ 0.6 , no similar compounds with known experimental value in the training set have been found.

Accuracy (average error) 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 fall 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:

If 1 ≥ index > 0.90, accuracy of prediction for similar molecules found in the training set is good

If 0.9 ≥ index > 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≤ 0.5, accuracy of prediction for similar molecules found in the training set is not adequate

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:

If 1 ≥ index > 0.90, molecules found in the training set have experimental values that agree with the target compound predicted value

If 0.9 ≥ index > 0.5,, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≤ 0.5, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

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:

If index = 1, all atom centered fragment of the compound have been found in the compounds of the training set

If 1 > index ≥ 0.7, some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments

If 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

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. Defined intervals are:

Index = TRUE, descriptors for this compound have values inside the descriptor range of the compounds of the training set

Index = FALSE, descriptors for this compound have values outside the descriptor range of the compounds of the training set

5.3.Software name and version for applicability domain assessment:

VEGA

Included in the VEGA software and automatically displayed when running the model

emilio.benfenati@marionegri.it

https://www.vegahub.eu/

5.4.Limits of applicability:

The model is not applicable to inorganic chemicals and substances containing unusual elements (i.e., different from C, O, N, S, P, Cl, Br, F, I). Salts can be predicted only if converted to the neutralized form.

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

NanoMaterial: 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:

Training set n 797, Positive 603 (76%), negative = 194 (24%)

6.6.Pre-processing of data before modelling:

NA

6.7.Statistics for goodness-of-fit:

TP=488 (true positive)

FP=81 (false positive)

TN=113 (true negative)

FN=115 (false negative)

Accuracy=0.75 Specificity=0.58 Sensitivity=0.81 MCC=0.37

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

NA

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

NA

6.10.Robustness - Statistics obtained by Y-scrambling:

NA

6.11.Robustness - Statistics obtained by bootstrap:

NA

6.12.Robustness - Statistics obtained by other methods:

NA

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

No

7.2.Available information for the external validation set:

NA

7.3.Data for each descriptor variable for the external validation set:

No

7.4.Data for the dependent variable for the external validation set:

No

7.5.Other information about the external validation set:

External validation set is not present

7.6.Experimental design of test set:

NA

7.7.Predictivity - Statistics obtained by external validation:

NA

7.8.Predictivity - Assessment of the external validation set:

NA

7.9.Comments on the external validation of the model:

NA

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model provides a qualitative prediction of carcinogenicity according to specific requirements of Chemical regulation. It is implemented in the VEGA software (https://www.vegahub.eu/). The model has been built as a set of rules, taken from the work of Benigni and Bossa (ISS) as implemented in the ToxTree software version 2.6 (http://toxtree.sourceforge.net). The model implement all the rules related to carcinogenicity and does not implement the full decision tree used by ToxTree. If the given compound matches at least one carcinogen rule, the model gives a 'carcinogen' prediction, otherwise it gives a 'non-carcinogen' prediction

8.2.A priori or a posteriori mechanistic interpretation:

NA

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1] R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, A. Worth "The Benigni/Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree" (2008) European Commission report EUR23241 EN

[2] Floris, M., Manganaro, A., Nicolotti, O. et al. A generalizable definition of chemical similarity for read-across. J Cheminform 6, 39 (2014). https://doi.org/10.1186/s13321-014-0039-1

9.3.Supporting information:

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

All available dataset are present in the model inside the VEGA software.

10.Summary (JRC QSAR Model Database)

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