	QMRF identifier (JRC Inventory): To be entered by JRC
	QMRF Title: Mutagenicity ISS Model - v. 1.0.3
	Printing Date: 30-05-2022

1.QSAR identifier

1.1.QSAR identifier (title):

Mutagenicity ISS Model (version 1.0.3)

1.2.Other related models:

This is the description of the VEGA model that implements the “In vitro mutagenicity (Ames test) alerts by ISS” as present in the software ToxTree v. 2.6.13

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:

30-05-2022

2.2.QMRF author(s) and contact details:

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

[2] Azadi Golbamaki IRCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156 Milano, Italy azadi.golbamaki@marionegri.it https://www.marionegri.it/

[3] Kristijan Vukovic IRCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156 Milano, Italy kristijan.vukovic@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] Romualdo Benigni Istituto Superiore di Sanità, Department of Environment and Primary Prevention Rome, Italy romualdo.benigni@iss.it

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

[3] Nina jeliazkova IDEA Consult Joseph II straat 40 B1, 1000 Brussels, Belgium jeliazkova.nina@gmail.com

[4] Alberto Manganaro RCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156 Milano, Italy alberto.manganaro@marionegri.it

2.6.Date of model development and/or publication:

2015

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

[1] The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree, (2008) by R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, and A. Worth. European Commission report EUR 23241 EN

[2] R. Benigni, C. Bossa, T. Netzeva, A. Rodomonte, and I. Tsakovska (2007) Mechanistic QSAR of aromatic amines: new models for discriminating between mutagens and nonmutagens, and validation of models for carcinogens. Environ mol mutag 48:754-771

[3] 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:

Histidine-dependent strains of *Salmonella typhimurium* (Ames test)

3.2.Endpoint:

TOX 7.6.1. Genetic toxicity in vitro

3.3.Comment on endpoint:

Mutagenic toxicity is the capacity of a substance to cause genetic mutations. This property is of high public concern because it has a close relationship with carcinogenicity and eventually reproductive toxicity: most of the mutagenic substances are suspected carcinogenic substance in case a genotoxic mechanism is considered. The Ames test is the basic in vitro assay to detect mutagens. The relevant test guideline covering this endpoint is OECD TG 471. The training set is based on test results from either the original version of the test guideline from 1983 or a newer version from 1997. The endpoint covers the DNA base-pair substitution and frameshift mutagenic mechanisms that are covered by the Ames tester strains: TA 1535, TA100, TA 98, and TA 1537 or TA97 or TA 97a. A part of the training set data additionally covers cross-linking mutagenic events measured by the inclusion of the E.coli WP2 or E.coli WP2 (pKM101) or TA 102 test strains. The test strains for DNA cross-links were included in the 1997 guideline update. As the training set does not systematically cover DNA cross-links, mutagenic substances acting by this mechanism may be under-predicted.

The endpoint is measured on the parent compound and the metabolites generated in vitro by the employed S9 mix of enzyme-induced rodent liver homogenates. In a few cases, liver homogenates from hamsters may have been used.

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

The dependent variable is mutagenic effect, as binary classification: 0 (non-mutagenic), 1 (mutagenic)

3.6.Experimental protocol:

Based on the OECD 471 test guideline. The details about the dataset compilation are reported on the ISS website. Ames mutagenicity data is described in Benigni R. Bossa C. Richard A. M. Yang C. 2008 A novel approach: chemical relational databases, and the role of the ISSCAN database on assessing chemical carcinogenicity. Ann. Ist. Super. Sanità 44, 48–56

3.7.Endpoint data quality and variability:

The data were cross-checked on different sources of information available; contradictions were solved going back to the original papers, and results based on insufficient protocols were not included. Second, the biological data (Salmonella mutagenicity) were coded in numerical terms that can be used directly for

QSAR analyses. This aspect of being QSAR-ready eliminates the intermediate passage of data transformation that often is problematic for the QSAR practitioner without specific toxicological expertise.

The general structure of the database is inspired by that of the DSSTox.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Structured-based model

4.2.Explicit algorithm:

The Mutagenicity ISS Model is based on In vitro mutagenicity (Ames test) alerts by ISS module of Toxtree: which is an open source application and the algorithm is available on (http://toxtree.sourceforge.net/ames.html)

4.3.Descriptors in the model:

If at least one mutagenicity rule matches with the given compound, it is classified as mutagenic, otherwise, it is non-mutagenic.

The following fragments are encoded as SA for matching mutagenic compounds:

[1] SA1 Acyl halides

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

[3] SA3 N-methylol derivatives

[4] SA4 Monohaloalkene

[5] SA5 S or N mustard

[6] SA6 Propiolactones and propiosultones

[7] SA7 Epoxides and aziridines

[8] SA8 Aliphatic halogens

[9] SA9 Alkyl nitrite

[10] SA10 alfa, beta unsaturated carbonyls

[11] SA11 Simple aldehyde

[12] SA12 Quinones

[13] SA13 Hydrazine

[14] SA14 Aliphatic azo and azoxy

[15] SA15 Isocyanate and isothiocyanate groups

[16] SA16 Alkyl carbamate and thiocarbamate

[18] SA18 Polycyclic Aromatic Hydrocarbons

[19] SA19 Heterocyclic Polycyclic Aromatic Hydrocarbons

[21] SA21 Alkyl and aryl N-nitroso groups

[22] SA22 Azide and triazene groups

[23] SA23 Aliphatic N-nitro

[24] SA24 alfa,beta unsaturated alkoxy

[25] SA25 Aromatic nitroso group

[26] SA26 Aromatic ring N-oxide

[27] SA27 Nitro aromatic

[28] SA28 Primary aromatic amine, hydroxyl amine and its derived esters (with restrictions)

[28] SA28bis Aromatic mono- and dialkylamine

[28] SA28ter Aromatic N-acyl amine

[29] SA29 Aromatic diazo

[30] SA30 Coumarins and Furocoumarins

[37] SA37 Pyrrolizidine Alkaloids

[38] SA38 Alkenylbenzenes

[39] SA39 Steroidal estrogens

[57] SA57 DNA Intercalating Agents with a basic side chain

[58] SA58 Haloalkene cysteine S-conjugates

[59] SA59 Xanthones, Thioxanthones, Acridones

[60] SA60 Flavonoids

[61] SA61 Alkyl hydroperoxides

[62] SA62 N-acyloxy-N –alkoxybenzamides

[63] SA63 N-aryl-Nacetoxyacetamides

[64] SA64 Hydroxamic acid derivatives

[65] SA65 Halofuranones

[66] SA66 Anthrones

[67] SA67 Triphenylimidazole and related

[10] SA68 9,10 – dihydrophenanthrenes

[69] SA69 Fluorinated quinolones

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 implements all the rules related to mutagenicity and does not implement the full decision tree used by ToxTree. If at least one mutagenicity rule is matching with the given compound, a “mutagen” prediction is given; otherwise, a “non-mutagen” prediction is given. The training set for the model has been extracted from ToxTree and consists of 670 mono-organic constituent compounds.

4.6.Software name and version for descriptor generation:

NA

4.7.Chemicals/Descriptors ratio:

NA

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.

Indices are calculated on the first *k* = 2 most similar molecules, each having S_k similarity value with the target molecule.

Similarity index (*IdxSimilarity*) is calculated as:

$$\frac{\sum_k S_k}{k} \times (1 - Diam^2)$$

where *Diam* is the difference in similarity values between the most similar molecule and the *k*-th molecule.

Accuracy index (*IdxAccuracy*) is calculated as:

$$\frac{\sum_c log\,(1 + S_c)}{\sum_k log\,(1 + S_k)}$$

where the molecules with *c* index are the subset of the *k* molecules where the prediction of the model matches with the experimental value of the molecule.

Concordance index (*IdxConcordance*) is calculated as:

$$\frac{\sum_c log\,(1+S_c)}{\sum_k log\,(1+S_k)}$$

where the molecules with *c* index are the subset of the *k* molecules where the experimental value of the molecule matches with the prediction made for the target molecule.

ACF contribution (*IdxACF*) index is calculated as

$$ACF = rare \times missing$$

where: *rare* is calculated on the number of fragments found in the molecule and found in the training set in less than 3 occurences as following: if the number is 0, *rare* is set to 1.0; if the number is 1, *rare* is set to 0.6; otherwise *rare* is set to 0.4

missing is calculated on the number of fragments found in the molecule and never found in the training set as following: if the number is 0, *missing* is set to 1.0; if the number is 1, *missing* is set to 0.6; otherwise *missing* is set to 0.4

AD final index is calculated as following:

$$ADI = (IdxSimilarity^{0.5} \times IdxAccuracy^{0.25} \times IdxConcordance^{0.25}) \times IdxACF$$

If 1 ≥ AD index ≥ 0.9, the predicted substance is regarded in 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 regarded 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 [ref 1 in 9.2]. 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 centered fragments.

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: Yes

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:

Data for each descriptor variable for the training set can be obtained by running the VEGA model on the training set

6.6.Pre-processing of data before modelling:

NA

6.7.Statistics for goodness-of-fit:

This is not a formal QSAR model developed through statistical methods so no specific internal and external validation has been performed. The SA have been tested on the ISSCAN dataset who served as training set of the ToxTree module and the performances were the following:

Training set: n = 670 (331 positive, 339 negative);

Accuracy = 0.79; Specificity = 0.68; Sensitivity = 0.89

TP 294, TN 232, FP 107, FN 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:

The external validation set is composed of a set of data selected from a big dataset comprising public and proprietary data [ref 2 and 3 in 9.2].

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

NA

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

NA

7.5.Other information about the external validation set:

The external validation set is composed of 17804 substances, 4757 experimentally positive and 13047 experimentally negative on Ames test.

7.6.Experimental design of test set:

NA

7.7.Predictivity - Statistics obtained by external validation:

Four compounds were not predicted (molecule error: unable to normalize SMILES string), then the available predictions for the statistical assessment were 17800.

We applied AD index thresholds to perform predictions on the external validation set and the results are:

The predictions of 2404 substances are in AD. AD index >=0.9.

Sensitivity	Specificity	Accuracy	MCC
0.92	0.80	0.85	0.71

TP 964, TN1079, FP 278, FN 83

The predictions of 5132 substances could be out of the AD. 0.9> AD index >= 0.65

Sensitivity	Specificity	Accuracy	MCC
0.78	0.81	0.80	0.54

TP 1047, TN 3051, FP 734, FN 300

The predictions of 10264 substances are out of the AD. AD index <0.65

Sensitivity	Specificity	Accuracy	MCC
0.72	0.65	0.66	0.31

TP 1694, TN 5128, FP 2773, FN 669

7.8.Predictivity - Assessment of the external validation set:

NA

7.9.Comments on the external validation of the model:

The distribution of the external validation dataset is unbalanced: the 73% of the compounds is non mutagenic experimentally.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model includes SAs to identify toxic compounds, according to the mechanistic basis described by the Benigni-Bossa rules.

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] 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

[2] Honma M, Kitazawa A, Cayley A, Williams RV, Barber C, Hanser T, Saiakhov R, Chakravarti S, Myatt GJ, Cross KP, Benfenati E, Raitano G, Mekenyan O, Petkov P, Bossa C, Benigni R, Battistelli CL, Giuliani A, Tcheremenskaia O, DeMeo C, Norinder U, Koga H, Jose C, Jeliazkova N, Kochev N, Paskaleva V, Yang C, Daga PR, Clark RD, Rathman J. Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. Mutagenesis. 2019 Mar 6;34(1):3-16. doi: 10.1093/mutage/gey031. PMID: 30357358; PMCID: PMC6402315.

[3] Cassano, A.; Raitano, G.; Mombelli, E.; Fernández, A.; Cester, J.; Roncaglioni, A.; Benfenati, E. Evaluation of QSAR Models for the Prediction of Ames Genotoxicity: A Retrospective Exercise on the Chemical Substances Registered Under the EU REACH Regulation. J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev. 2014, 32, 273–298. DOI: 10.1080/10590501.2014.938955.

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