	***QMRF identifier (JRC Inventory):*** **To be entered by JRC**
	QMRF Title: **Mutagenicity (Ames test) model (KNN/Read-Across) - v. 1.0.1.**
	Printing Date: Feb 18, 2022

1.QSAR identifier

1.1.QSAR identifier (title):

Mutagenicity (Ames test) model (KNN/Read-Across) - v. 1.0.1

1.2.Other related models:

NA

1.3.Software coding the model:

istKNN v. 0.9

The read-across model has been built with the istKNN application (developed by Kode srl, http://chm.kode-solutions.net) and it is based on the similarity index developed inside the VEGA platform

http://chm.kode-solutions.net

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:

February 2022

2.2.QMRF author(s) and contact details:

[1] Azadi Golbamaki Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy azadi.golbamaki@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/

2.3.Date of QMRF update(s):

NA

2.4.QMRF update(s):

Updates made by Giuseppa Raitano giuseppa.raitano@marionegri.it.

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] 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] Giuseppa Raitano Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2,20156 Milano, Italy giuseppa.raitano@marionegri.it https://www.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] Emilio Benfenati, Serena Manganelli, Sabrina Giordano, Giuseppa Raitano & Alberto Manganaro (2015) Hierarchical Rules for Read-Across and In Silico Models of Mutagenicity, Journal of Environmental Science and Health, Part C, 33:4, 385-403, DOI: 10.1080/10590501.2015.1096881

[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

[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 invitro 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 cancerogenic effect on rat, as binary classification: mutagenic and non-mutagenic

3.6.Experimental protocol:

Ames test is an in vitro model of chemical mutagenicity and consists of a range of bacterial strains that together are sensitive to a large array of DNA-damaging agents

3.7.Endpoint data quality and variability:

The estimated inter-laboratory reproducibility rate of S. typhimurium test data is 85% [ref.1, sect.9.2]. The dataset used to develop and validate the model includes 5770 compounds collected from a large set of compounds [ref.2, sect.9.2] and data produced within the Ames QSAR project organized by National Institute of Health Sciences of Japan [ref.3, sect.9.2]. The Ames assays were conducted under GLP according to Industrial Safety and Health Act in Japan

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The model performs a read-across and provides a qualitative prediction of mutagenicity on Salmonella typhimurium (Ames test). It is implemented inside the VEGA online platform, accessible at: http://www.vegahub.eu

4.2.Explicit algorithm:

The model was built using the istKNN software v0.9, which performs k-NN predictions based on the k most similar compounds retrieved with the similarity index developed in VEGA. A complete description of this k-NN approach and of its implementation in the istKNN application has been provided by Manganaro et al (2015) (Manganaro, A., Pizzo, F., Lombardo, A., Pogliaghi, A. and Benfenati, E. (2015). Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN) algorithm. Chemosphere. 144, 1624-1630). This approach assesses the predictive power by calculating the predictions on each molecule from the training set itself in a leave-one-out (LOO) approach. The algorithm for the prediction involves the following steps:

1. The first k molecules with the closest similarity to the target compound are extracted.

2. Molecules with a similarity index lower than a selected threshold S1 are excluded.

3. If no molecules are left, no prediction is provided (missing value).

4. If only one molecule is left, it is used as prediction only if its similarity value is equal to or higher than a given threshold S2, otherwise no prediction is provided (missing value).

5. In all other cases, the prediction is calculated as a weighted consensus of the experimental values among the remaining molecules.

A score for each class is calculated as the sum of the weights of compounds experimentally belonging to the class itself. Finally, the class with the highest score is chosen as the prediction to be provided. The weights (similarity values) can be raised to the power of a given value E, called the enhance factor, as for integers larger than 1 the result is to enhance the role of molecules with higher similarity values in the prediction. In general it is recommended that the user try several integers, which have the effect to slightly modulate the algorithm such that an optimal result is achieved.

4.3.Descriptors in the model:

The model is based on the similarity index developed inside the VEGA platform taking into account different chemical features of the molecule (size, presence/absence of certain heteroatoms, functional groups etc.). It was described by Floris et al., 2014 (Floris, M., Manganaro, A., Nicolotti, R.Medda, G. F. Mangiatordi, E. Benfenati (2014). A generalizable definition of chemical similarity for read-across, J. Cheminform., 6, 39)

4.4.Descriptor selection:

NA

4.5.Algorithm and descriptor generation:

NA

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 = the number of k compounds used in the KNN model for the prediction 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 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 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 [4]. 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.

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:

No

6.4.Data for the dependent variable for the training set:

All

6.5.Other information about the training set:

The dataset includes 5770 (3254 mutagenic, 2516 NON-Mutagenic) mono-constituent organic compounds collected from:

- Freely available Benchmark dataset for in silico prediction of Ames mutagenicity [2]
- Japan's Health Ministry (data produced within the Ames QSAR project organized by National Institute of Health Sciences of Japan) [3][5]

6.6.Pre-processing of data before modelling:

All chemical structures have been checked manually.

6.7.Statistics for goodness-of-fit:

NA

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

The statistics were calculated on the entire dataset using the leave-one-out cross-validation.

n = 5764; Accuracy = 0.80; Specificity = 0.76; Sensitivity = 0.83 Non predicted compounds: n = 6

TP 2706; TN 1906; FP 606, FN 546

On the basis of this structural similarity index, the four compounds from the dataset resulting most similar to the chemical to be predicted are taken into account; compounds with a similarity value lower than 0.7 are discarded. If no compounds fall under these conditions, no prediction is provided.

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 [5][6].

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:

The external validation set is composed of 12593 substances, 1798 experimentally positive and 10795 experimentally negative on Ames test.

7.6.Experimental design of test set:

NA

7.7.Predictivity - Statistics obtained by external validation:

71 compounds were not predicted (molecule error: unable to normalize SMILES string and unable to perform Applicability Domain check), then the available predictions for the statistical assessment were 12522.

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

The predictions of 2437 substances are in AD. AD index >=0.9.

Sensitivity	Specificity	Accuracy	MCC
0,62	0,84	0,81	0,40

TP 238, TN1731, FP 323, FN 145

The predictions of 7487 substances could be out of the AD. 0.9> AD index >= 0.65

Sensitivity	Specificity	Accuracy	MCC
0,53	0,69	0,67	0,16

TP 544, TN4460, FP 1999, FN 484

The predictions of 2598 substances are out of the AD. AD index <0.65

Sensitivity	Specificity	Accuracy	MCC
0,49	0,64	0,61	0,09

TP 188, TN1409, FP 807, FN 194

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 86% of the compounds is non mutagenic experimentally.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

Not possible: the model adopts the read-across approach based on chemical similarity

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] Piegorsch WW & Zeiger E (1991) Measuring intra-assay agreement for the Ames salmonellaassay. In Statistical Methods in Toxicology, Lecture Notes in Medical Informatics. Edited by HotornL. Springer-Verlag, 35-41 https://www.springer.com/gp/book/9783540536215

[2] Hansen, K., Mika, S., Schroeter, T., Sutter, A., Ter Laak, A., Steger-Hartmann, T., Heinrich N.,and Muller, K.-R. (2009). Benchmark data set for in silico prediction of Ames mutagenicity. J. Chem.Inf. Model. 49, 2077-2081

[3] Giuseppa Raitano, Alessandra Roncaglioni, Alberto Manganaro, Masamitsu Honma, Laurent Sousselier, Quoc Tuan Do, Eric Paya, Emilio Benfenati, Integrating in silico models for the prediction of mutagenicity (Ames test) of botanical ingredients of cosmetics, Computational Toxicology, Volume 12, 2019,

https://doi.org/10.1016/j.comtox.2019.100108.

[4] 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

[5] 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.

[6] 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:

The available dataset is 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