	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Mutagenicity (Ames test) model (KNN/Read-Across) - v. 1.0.0.**
	Printing Date: **18-feb-2020**

1.QSAR identifier

1.1.QSAR identifier (title):

Mutagenicity (Ames test) model (KNN/Read-Across) - v. 1.0.0.

1.2.Other related models:

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.vega-qsar.eu/

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

2.General information

2.1.Date of QMRF:

29 March 2017

2.2.QMRF author(s) and contact details:

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

2.3.Date of QMRF update(s):

17/02/2020

2.4.QMRF update(s):

Updates made by Giuseppa Raitano, email: giuseppa.raitano@marionegri.it

Updates in the fields:

-1.1
-1.2
-1.3
-2.2
-2.5
-2.7
-2.8
-3.1
-3.2
-3.3
-3.5
-3.6-3.7-5.2
-5.3-5.4

-6.3

-6.4-6.5

-6.6-8.1

2.5.Model developer(s) and contact details:

[1]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]Alberto Manganaro Kode srl info@kode-solutions.net www.kode-solutions.net

[3]Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy emilio.benfenati@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:

A generalizable definition of chemical similarity for read-across

2.8.Availability of information about the model:

Complete documentation, comprising training and test information, about the model is available on the guideline of the model, inside the VEGA application (see section 1.2)

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

Other QMRF for this model are not available. This is an updated version of the original one.

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.

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

Binary classification as: 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:

k-Nearest Neighbor (k-NN) model

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:

4.5.Algorithm and descriptor generation:

See section 4.3

4.6.Software name and version for descriptor generation:

istKNN

4.6. Software name and version for descriptor generation: The model was built with istKNN v0.9 application (developed by Kode srl, www.kode-solutions.net) and it is based on the similarity index developed inside the VEGA platform; the index takes into account several structural aspects of the compounds, such as their fingerprint, the number of atoms, of cycles, of heteroatoms, of halogen atoms, and of particular fragments (such as nitro groups). The index value ranges from 1 (maximum similarity) to 0.

4.7.Chemicals/Descriptors ratio:

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of predictions is assessed using an Applicability Domain Index (ADI) that has values from 0 (worst case) to 1 (best case). The ADI is calculated by grouping several other indices, each one
taking into account a particular issue of the applicability domain. Most of the indices are based on the calculation of the most similar compounds found in the training and test set of the model, calculated by a similarity index that consider molecule's fingerprint and structural aspects (count of atoms, rings and relevant fragments).
Following, all applicability domain components are reported along with their explanation:
- Similar molecules with known experimental value. This index considers how similar are the first three 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. The interval defined for this model are:
- 1 >= index > 0.8 strongly similar compounds with known experimental value in the training set have been found
- 0.8 >= index > 0.6 only moderately similar compounds with known experimental value in the training set have been found
- index <= 0.6 no similar compounds with known experimental value in the training set have been found.
- Accuracy of prediction for similar molecules. This index considers the classification accuracy in prediction for the three 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. The interval defined for this model are:
- 1 >= index > 0.9 accuracy of prediction for similar molecules found in the training set is good
- 0.9 >= index > 0.5 accuracy of prediction for similar molecules found in the training set is not optimal
- index <= 0.5 accuracy of prediction for similar molecules found in the training set is not adequate.
- Concordance for similar molecules. This index considers the difference between the predicted value and the experimental values of the three 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. The interval defined for this model are:
- 1 >= index > 0.9 similar molecules found in the training set have experimental values that agree with the predicted value

- 0.9 >= index > 0.5 some similar molecules found in the training set
have experimental values that disagree with the predicted value- index <= 0.5 similar molecules found in the training set have
experimental values that disagree with the predicted value.- Atom Centered Fragments similarity check. This index considers the
presence of one or more fragments that aren't found in the training set,
or that are rare fragment. The interval defined for this model are:
- index = 1 all atom centered fragment of the compound have been found
in the compounds of the training set- 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
- 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.

5.2.Method used to assess the applicability domain:

Global Applicability Domain Index (ADI) values goes between 0 and 1.
If the ADI value ranges from 0.9 to 1 this means that the predicted
substance is into the Applicability Domain of the model. If 0.9 > ADI >=
0.65 it means that the predicted substance could be out of the
Applicability Domain of the model index. Finally, if ADI < 0.65 it means
that the predicted substance is out of the Applicability Domain of the
model.

5.3.Software name and version for applicability domain assessment:

VEGA
AD included in the VEGA software and automatically displayed when running the model
https://www.vegahub.eu/

5.4.Limits of applicability:

See 2.8

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 include 5770 compounds collected from:

1. Freely available Benchmark dataset for in silico prediction of Ames mutagenicity (Hansen, K., Mika, S., Schroeter, T., Sutter, A., Ter Laak, A., Steger-Hartmann, T., Heinrich N., and Mu?ller, K.-R. (2009). Benchmark data set for in silico prediction of Ames mutagenicity. J. Chem. Inf. Model. 49, 2077-2081)

2. Japan's Health Ministry (data produced within the Ames QSAR project organized by National Institute of Health Sciences of Japan).

6.6.Pre-processing of data before modelling:

All chemical structures have been checked manually.

6.7.Statistics for goodness-of-fit:

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

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:

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

No

7.2.Available information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

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:

7.6.Experimental design of test set:

7.7.Predictivity - Statistics obtained by external validation:

7.8.Predictivity - Assessment of the external validation set:

7.9.Comments on the external validation of the model:

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:

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Piegorsch WW & Zeiger E (1991) Measuring intra-assay agreement for the Ames salmonella assay. In Statistical Methods in Toxicology, Lecture Notes in Medical Informatics. Edited by Hotorn L. 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]Integrating in silico models for the prediction of mutagenicity (Ames test) of botanical ingredients of cosmetics

9.3.Supporting information:

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

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