	***QMRF identifier (JRC Inventory):* To be entered by JRC**
	***QMRF Title: In vitro* Micronucleus activity (IRFMN/VERMEER) - v.1.0.1**
	Printing Date: May 07, 2022

1.QSAR identifier

1.1.QSAR identifier (title):

In vitro Micronucleus activity (IRFMN/VERMEER) - v.1.0.1

1.2.Other related models:

Other genotoxicity models (including models for carcinogenicity and mutagenicity) are implemented inside VEGA online platform, accessible at: https://www.vegahub.eu/An example is the Caesar hybrid model for bacterial reverse mutation (Ames Test) [Q15-410-0008

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:

07/05/2022

2.2.QMRF author(s) and contact details:

[1]Diego Baderna Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS diego.baderna@marionegri.it diego.baderna@marionegri.it https://www.marionegri.it/laboratories/laboratory-of-chemistry-and-environmental-toxicology

[2]Emilio Benfenati Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS emilio.benfenati@marionegri.it emilio.benfenati@marionegri.it https://www.marionegri.it/laboratories/laboratory-of-chemistry-and-environmental-toxicology

2.3.Date of QMRF update(s):

NA

2.4.QMRF update(s):

NA

2.5.Model developer(s) and contact details:

[1]Diego Baderna Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS diego.baderna@marionegri.it https://www.marionegri.it/laboratories/laboratory-of-chemistry-and-environmental-toxicology

[2]Emilio Benfenati Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS emilio.benfenati@marionegri.it https://www.marionegri.it/laboratories/laboratory-of-chemistry-and-environmental-toxicology

[3]Alberto Manganaro Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS alberto.manganaro@marionegri.it https://www.marionegri.it/laboratories/laboratory-of-chemistry-and-environmental-toxicolog

2.6.Date of model development and/or publication:

2019

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

[1] Baderna, D., Gadaleta, D., Lostaglio, E., Selvestrel, G., Raitano, G., Golbamaki, A., Lombardo, A., Benfenati, E., 2020. New in silico models to predict in vitro micronucleus induction as marker of genotoxicity. Journal of Hazardous Materials 385.

[2] T. Ferrari, D. Cattaneo, G. Gini, N.G. Bakhtyari, A. Manganaro, E. Benfenati. Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction SAR and QSAR in Environmental Research, 24 (2013), pp. 365-383 https://www.ncbi.nlm.nih.gov/pubmed/23710765

[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:

NA

3.Defining the endpoint - OECD Principle 1

3.1.Species:

In vitro mammalian cell lines according to the OECD guideline no 487 "*In Vitro* Mammalian Cell Micronucleus Test" (2010 & 2014)

3.2.Endpoint:

TOX 7.6.1. Genetic toxicity in vitro

3.3.Comment on endpoint:

Genotoxicity is the ability of an agent to cause DNA damage as an alteration in the structure or information content of genetic material in cells, including those that are permanently transmissible. Micronuclei are small cytoplasmic bodies originated from chromosome fragments or whole chromosomes that were unable to migrate to the poles during anaphase in cell division. The assay is typically performed with human or rodent cell lines or primary cell cultures.

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

The dependent variable is genotoxic effect (induction of micronuclei invitro), as binary classification: 0 (non-genotoxic), 1 (genotoxic)

3.6.Experimental protocol:

Data were selected according to their adherence to the OECD 487 guideline considering only data related to experiments conducted with human peripheral blood lymphocytes, CHO, V79, CHL/IU, L5178Y, TK6, HT29, Caco-2, HepaRG, HepG2, A549 and primary Syrian Hamster Embryo cells as cell lines listed in the guideline. Regarding the use of post-mitochondrial liver fraction (S9), we included data from experiments performed in accordance with the test guideline with S9 and positive data from studies without S9 metabolic activation

3.7.Endpoint data quality and variability:

The dataset includes 380 mono-constituent organic compounds with experimental data collected from peer-reviewed literature, SCCS and EFSA opinions, ECVAM guidelines and review, and eChemPortal inventory. We carefully revised the sources in order to ensure their quality and reliability and, to our knowledge, most of the selected data can be classified with a Klimisch score of 1 due to the facts that the studies were done with test procedure in accordance with validated standard methods

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The In vitro Micronucleus Activity (IRFMN/VERMEER) model (version 1.0.0) provides a qualitative prediction of genotoxicity as induction of micronucleus in mammalian cells in vitro. It is based on a set of rules extracted from a set of compounds by SARpy software without any 'a priori' knowledge

4.2.Explicit algorithm:

Prediction of the genotoxic/non-genotoxic activity of chemical is based on selected structural alerts, applying the rule "IF contains SA THEN apply activity label" and considering their Likehood Ratio (LR) values[1]. The algorithm generates substructures of arbitrary complexity, and the fragment candidates to become structural alerts (SAs) are automatically selected based on their prediction performance on a training set. Fragmentation is done directly on the SMILES notation of structures. Positive(genotoxic) and negative (non-genotoxic) rules were found. If at least one rule is matching with the target compound, a prediction is given according to the type of rule (positive or negative); if no rules match with the given compound, no prediction is given.

4.3.Descriptors in the model:

[1]Active Structural Alerts (SAs) adimensional 82 genotoxic (active/positive) SAs were selected from a larger list of SAs automatically extracted with SARpy. Only structural alerts with LR higher than 2 were considered for the development of the model to obtain a more accurate model minimizing the occurrence of wrong predictions.

[2]Inactive Structural Alerts adimensional 56 non-genotoxic (inactive/negative) SAs were selected from a larger list of SAs automatically extracted with SARpy. Only structural alerts with LR higher than 2 were considered for the development of the model to obtain a more accurate model minimizing the occurrence of wrong predictions.

4.4.Descriptor selection:

Only structural alerts with LR higher than 2 were considered for the development of the model to obtain a more accurate model minimizing the occurrence of wrong predictions

4.5.Algorithm and descriptor generation:

The structural alerts (SAs) were extracted using the software SARpy (SAR in python, version 1.0), a freely available software in the VEGA HUB platform (www.vegahub.eu). SARpy automatically extracts sets of rules by generating and selecting substructures based on their prediction performance on a training set used as input. The selection is made with a 3-steps process:

1) the fragmentation of chemicals to extract all the substructures within a user-customizable size range;

2) a validation step to derive the predictive power of each fragment through the analysis of the correlation between the occurrence of each molecular substructure and the experimental activity of the compounds containing the fragment.

3) a selection step in which the most predictive fragments are listed in the form of rules "IF contains SA THEN apply activity label".

SARpy was applied to extract SAs for both positive (genotoxicant) and negative(non-genotoxicant) activity. Various runs with different settings for alert precision were done to obtain fragments with high accuracy or with high coverage. Only structural alerts with LR higher than 2 were considered for the development of the model to obtain a more accurate model minimizing the occurrence of wrong predictions.

4.6.Software name and version for descriptor generation:

SARpy (SAR in python) version 1.0.

SARpy breaks the chemical structures of the compounds in the training set into fragments of a desired size, and it identifies fragments related to the target property. It then also shows the fragments related to the effect. Inhibiting conditions are identified which prevent the appearance of the effect, even in presence of

[1] Likehood ratio (LR): in SARpy, it's a parameter that defines the precision of SAs extractions: a high value of LR corresponds to a higher precision of the model.

the active fragment. The system uses SMILES in the canonical form. It allows choice in building more conservative or more accurate models. https://www.vegahub.eu/portfolio-item/sarpy/

4.7. Chemicals/Descriptors ratio:

The Chemicals Descriptors ratio is 293/138 = 2.12 (considering only the training set, used to extract the SAs)

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 ≥ ADI > 0.8, predicted substance is regarded in the Applicability Domain of the model and predictions are characterized by good reliability

If 0.8 ≥ ADI > 0.6 predicted substance could be out of the Applicability Domain of the model and predictions are characterized by moderate reliability

If ADI ≤ 0.6 predicted substance is regarded out of the Applicability Domain of the model and predictions are characterized by low reliability

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

Similarity index (IdxSimilarity) is calculated as:

(∑_k S_k)/k×(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:

(∑_c log (1+S_c))/(∑_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:

(∑_c log (1+S_c))/(∑_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×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×〖IdxAccuracy〗^0.25×〖IdxConcordance〗^0.25)×IdxACF

No ADI threshold was applied to provide performance calculations on validation set

5.2.Method used to assess the applicability domain:

Thechemical similarity and Applicability Domain are measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it. 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: 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 training set includes 293 mono-constituent organic chemicals. The collection, slightly imbalanced toward genotoxicants (about 60% of the set), includes chemicals from all the commercial and production sectors for which genotoxicity testing is required including drugs, plant protection products, industrial reagents, and Personal Care Products (PCPs) ingredients

6.6.Pre-processing of data before modelling:

Experimental data were retrieved from previously published studies and were selected according to their adherence to the OECD 487 guideline, considering only data related to experiments conducted with human peripheral blood lymphocytes, CHO, V79, CHL/IU, L5178Y, TK6, HT29, Caco-2, HepaRG, HepG2, A549 and primary Syrian Hamster Embryo cells as cell lines listed in the guideline. Regarding the use of post-mitochondrial liver fraction (S9), we included data from experiments with S9 and positive data from studies without S9 metabolic activation.

We carefully revised the sources in order to ensure their quality and reliability and, to our knowledge, most of the selected data can be classified with a Klimisch score of 1 due to the facts that the studies were done with test procedure in accordance with validated standard methods.

A published KNIME workflow for chemical data retrieval and quality checking was used to automatically retrieve SMILES of the chemicals. For each chemical, the structural data are retrieved automatically from four sound web-based chemo-informatic sources (National Cancer Institute (NCI) Chemical Identifier Resolver (CIR), the U.S. Environmental Protection Agency (EPA) CompTox Chemistry Dashboard, PubChem and ChemIDPlus) using chemical name and CAS as input. Structural data that are consistent among all the web databases are further cleaned to remove inorganic and organometallic compounds, isomeric mixtures, polymers and data related to mixtures of chemicals. In the end, the neutralized SMILES are converted into a standardized QSAR-ready format using the OpenBabel KNIME implementation to generate Canonical SMILES. The curated structures were then split into training and test sets according to their structural similarity. Briefly, fingerprints of the structures in the dataset were calculated using the RDKit Fingerprint node and Morgan fingerprints as implemented in KNIME (version 4.0.2). The structural similarity was estimated with the Distance Matrix KNIME node using the Tanimoto function. A k-Medoids clustering algorithm was applied on the distance matrix followed by a partitioning node to split the dataset in training and test sets, with an analogue chemical distribution, in an 80:20 ratio

6.7.Statistics for goodness-of-fit:

Tranining set = 293 (171 active, 122 inactive)

Accuracy 0.88; Specificity 0.73; Sensitivity 0.97; Matthews correlation coefficient 0.75; Unpredicted (no rule has been matched with those compounds) compound 50/293;

False positive 26; False negative 4; True positive 141; True Negative 72

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:

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

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:

All

7.5. Other information about the external validation set:

The validation set includes 87 mono-organic constituent chemicals from all the commercial and production sectors for which genotoxicity testing is required including drugs, plant protection products, industrial reagents, and PCPs ingredients.

7.6. Experimental design of test set:

The curated structures were split into training and test sets according to their structural similarity: the fingerprints of the structures in the dataset were calculated using the RDKit Fingerprint node and Morgan fingerprints as implemented in KNIME (version 4.0.2). Then, the structural similarity was estimated with the Distance Matrix KNIME node using the Tanimoto function. A k-Medoids clustering algorithm was applied on the distance matrix followed by a partitioning node to split the dataset in training and test sets, with an analogue chemical distribution, in an 80:20 ratio

7.7. Predictivity - Statistics obtained by external validation:

Test set = 87 (56 active, 31 inactive)

Accuracy 0.83; Specificity 0.63; Sensitivity 0.94; Matthews correlation coefficient 0.62; Unpredicted compound 12/87

False positive 10; False negative 3; True positive 45; True negative 17

Statistics considering ADI thresholds

27 compounds in the AD, AD index > 0.8

Sensitivity= 93%, Specificity= 83%, Accuracy= 89% and MCC= 0.78

TP 14, TN 10, FP 2, FN 1.

30 compounds could be out the AD, 0.8 ≥ AD index > 0.6

Sensitivity= 96%, Specificity= 50%, Accuracy= 87% and MCC=0.54

TP 23, TN 3, FP 3, FN 1.

18 compounds out of the AD, AD index ≤ 0.6

Sensitivity= 89%, Specificity= 44%, Accuracy= 67% and MCC= 0.37

TP 8, TN 4, FP 5, FN 1.

7.8. Predictivity - Assessment of the external validation set:

The validation set was derived from the whole dataset after a splitting procedure based on chemicals structural similarity and applying a 80:20ratio. The analysis of chemical domain by the Checkmol profiler of

the OECDQSAR Toolbox revealed that the presence of aromatic compounds, heterocyclic compounds, amines, carboxylic acid derivatives, and hydroxyl compounds and several other functional groups while the analysis for the application sector highlights that chemicals in the validation set cover all the commercial and production sectors for which genotoxicity testing is required including drugs, plant protection products, industrial reagents, and PCPs ingredients

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 includes SAs to identify both genotoxic and non-genotoxic compounds. The VEGA system provides, in the final PDF report for the prediction, a set built with the most similar compounds found in the training and test set of the model, coupled to the assessment of the target. An expert-based analysis of these compounds like the predicted one, which are provided with their experimental activity, can lead to a further mechanistic interpretation of the results given by the model

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori: the fragments, identified as statistically associated to the genotoxic or nongenotoxi toxic class, can be investigated to explore the mechanistic basis of the mode

8.3.Other information about the mechanistic interpretation:

The genotoxic SAs (82) are associated to DNA intercalation, ROS formation, SN1 nucleophilic attack, AN2 nucleophilic and Michael addictions, Schiff base formation, SN2 Nucleophilic type substitution, and DNA alkylation

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1]Baderna, D., Gadaleta, D., Lostaglio, E., Selvestrel, G., Raitano, G., Golbamaki, A., Lombardo, A., Benfenati, E., 2020. New in silico models to predict in vitro micronucleus induction as marker of genotoxicity. Journal of Hazardous Materials 385.

[2] T. Ferrari, D. Cattaneo, G. Gini, N.G. Bakhtyari, A. Manganaro, E. Benfenati. Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction SAR and QSAR in Environmental Research, 24 (2013), pp. 365-383 https://www.ncbi.nlm.nih.gov/pubmed/23710765

9.3.Supporting information:

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

All available datasets 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