*QMRF identifier (JRC Inventory):***To be entered by ECB**

*QMRF Title:***IRFMN/ISSCAN-CGX expert rule-based model for carcinogenicity**

*Printing Date:***24-gen-2020**

1.QSAR identifier

1.1.QSAR identifier (title):

IRFMN/ISSCAN-CGX expert rule-based model for carcinogenicity

1.2.Other related models:

n/a

1.3.Software coding the model:

VEGA (Virtual Models for evaluating the properties of chemicals within a global architecture, version 1.1.4, 2017)

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

https://www.vegahub.eu/contacts/

https://www.vegahub.eu/

2.General information

2.1.Date of QMRF:

February 2018

2.2.QMRF author(s) and contact details:

Simona Kovarich S-In Soluzioni Informatiche Soluzioni Informatiche Srl Via Ferrari 14, I-36100 Vicenza simona.kovarich@s-in.it http://www.s-in.it

2.3.Date of QMRF update(s):

n/a

2.4.QMRF update(s):

n/a

2.5.Model developer(s) and contact details:

Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri (IRFMN) Istituto di Ricerche Farmacologiche Mario Negri. Via La Masa, 19 - 20156 Milano emilio.benfenati@marionegri.it http://www.marionegri.it

2.6.Date of model development and/or publication:

2016

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

[1]VEGA (Virtual Models for evaluating the properties of chemicals within a global architecture, version 1.1.4, 2017) https://www.vegahub.eu/

[2]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 http://dx.doi.org/10.1080/10590501.2016.1166879

2.8.Availability of information about the model:

The model has been released open source and is available through the VEGA HUB portal. The datasets used for the exctraction of the rules (structural alerts), i.e., ISSCAN database and CGX dataset, are available and included as Supporting Information.

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

Not to date

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Rodents

3.2.Endpoint:

QMRF 4.12. Carcinogenicity OECD 451 Carcinogenicity Studies

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:

Qualitative prediction of carcinogenicity (expert assessment based on carcinogenic effects in diffferent species). Prediction call: "Carcinogen"/"Possible NON-Carcinogen"

3.6.Experimental protocol:

The rules (structural alerts) have been extracted with SARpy software from a dataset obtained from the union of the ISS carcinogenicity (ISSCAN) database from Istituto Superiore della Sanità (available at: http://www.iss.it/meca/index.php?lang=1&anno=2013&tipo=25 - last access 14/02/2018) [1], and of the Carcinogenicity Genotoxicity eXperience (CGX) dataset (available at: https://eurl-ecvam.jrc.ec.europa.eu/databases/genotoxicity-carcinogenicity-db/genotoxicity-carcinogenicity-db - last access 14/02/2018) [2].

3.7.Endpoint data quality and variability:

Carcinogenicity data on rodents have been processed by human experts (from ISS and JRC). In more detail:

Experimental carcinogenicity data and chemical structures included in the ISSCAN database have been curated by ISS scientists (toxicological data assessed and critically selected).

The EURL ECVAM Genotoxicity & Carcinogenicity Consolidated database is a structured master database that compiles available genotoxicity and carcinogenicity data for Ames positive chemicals originating from different sources, including regulatory agencies, industry and literature databases covering different sectors (e.g., US-NTP, EFSA, SCCS, Cosmetic Europe, BASF, ECHA, ISSTox, …). Only chemicals with a known chemical identity (structure, purity, molecular weight, CAS number) and valid in vitro and in vivo results for the genotoxicity endpoints and/or for carcinogenicity were included. "Overall Calls" were defined for each genotoxicity assay in vitro and in vivo and carcinogenicity by following defined criteria for the reliability of each study and quality of data for those chemicals appearing in more than one source with different calls. 4 categories were considered (+), (-), (E) and (I). Where information was missing, even for those chemicals with one single data entry, scientific literature was consulted.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Expert rule-based system

4.2.Explicit algorithm:

Expert rule-based system

Set of 43 rules (structural alerts) related to carcinogenic activity. These rules are expressed SMARTS representing molecular fragments (reported in section 4.3). If at least one rule is matching with the target compound, a “Carcinogen” prediction is given. Otherwise, a “Possible NON-Carcinogen” prediction is given.

4.3.Descriptors in the model:

[1]O=NNCC

[2]c1occc1

[3]O=CN(N)C

[4]CCCN(CC)CC

[5]C1CC(=CC)CCC1

[6]Nc1ccc(cc1C)C

[7]NCCCN

[8]O=S(=O)(OC)

[9]c1ccc2OCOc2c1

[10]Nc1ncccc1

[11]N(CCCl)CCCl

[12]c1cn(cnc1)

[13]C=C(C=C)C

[14]O=NNC

[15]O=P(OC)

[16]O(c1ccc(cc1)CC=C)

[17]c1ncn(c1)C

[18]C(CCCC(CC)Cl)Cl

[19]c1ncsc1

[20]C=CCN

[21]O=Cc1ccccc1O

[22]O(c1ccc(cc1N))C

[23]O1CC1C

[24]SN(C)C

[25]C(CCl)Cl

[26]c1c(cc(cc1Cl)Cl)Cl

[27]NNCC

[28]O=CN(N)
[29]C(OC)C(C)C
[30]c1ccc2cc(ccc2c1)
[31]Nc1cccc(c1C)C
[32]NNc1ccccc1
[33]c1cc(ccc1C)Cl
[34]N(CCO)CCO
[35]Nc1ccc(cc1N)
[36]c1ccc(cc1N)C
[37]O(c1ccc(cc1)C)C
[38]C(c1ccccc1)CO
[39]C(=CCC)CC
[40]N(Cc1ccccc1)C
[41]Nc1ccc(cc1)C
[42]Nc1ccccc1
[43]n1cccc(c1)

4.4.Descriptor selection:

The SARpy software has been used with a cross-validated procedure, ending with the extraction of a set of 43 rules (structural alerts) related to carcinogenic activity.

4.5.Algorithm and descriptor generation:

The 43 rules (structural alerts) are expressed as SMARTS representing molecular fragments. The SARpy software was used to extract the rules from the two carcinogenicity datasets.

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 the active fragment. The system uses SMILES in the canonical form. It allows choice in building more conservative or more accurate models.

4.6.Software name and version for descriptor generation:

SARpy software
free tool to develop a model to classify chemicals according to a given property
Giuseppina Gini, Politecnico di Milano (giuseppina.gini@polimi.it); Emilio Benfenati, Istituto di Ricerche Farmacologiche Mario Negri (emilio.benfenati@marionegri.it)
http://sarpy.sourceforge.net

4.7.Descriptors/Chemicals ratio:

not applicable to expert systems

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The model is applicable to heterogeneous organic chemicals. In the software implementation of the model several pieces of information are given to evaluate if a prediction is reliable (chemical falling in the Applicability Domain or not). The information about Applicability Domain

(AD) is combined into a unique index called Global Applicability Domain Index (ADI). Global AD Index values range between 0 and 1. ADI > 0.8 means that the compound is in the AD of the model, ADI < 0.6 means that the compound is out of the AD of the model, a value between 0.6 and 0.8 means that the compound is possibly out of model AD, and further analysis is required.

5.2.Method used to assess the applicability domain:

Within VEGA, the Applicability Domain of the model is defined by considering several parameters as described below:

1. Similar molecules with known experimental values: this index takes into account 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. Defined intervals are: 0-0.6 (no similar compounds with known experimental value in the training set have been found), 0.6-0.8 (only moderately similar compounds with known experimental value in the training set have been found, 0.8-1 (strongly similar compounds with known experimental value in the training set have been found).
2. Accuracy of prediction for similar molecules: this index takes into account the classification accuracy in prediction for the three most similar compounds found. Values near 1 mean that the predicted compounds falls 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: 0-0.6 (prediction accuracy not adequate), 0.6-0.8 (prediction accuracy not optimal), 0.8-1 (good prediction accuracy).
3. Concordance for similar molecules: this index takes into account 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. Defined intervals are: 0-0.6 (similar molecules found in the training set have experimental values that disagree with the predicted value), 06-0.8 (some similar molecules found in the training set have experimental values that disagree with the predicted value), 0.8-1 (similar molecules found in the training set have experimental values that agree with the predicted value).
4. 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: 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), 0.7<= index < 1 (some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments), index = 1 (all atom centered fragment of the compound have been found in the compounds of the training set).

5.3.Software name and version for applicability domain assessment:

VEGA (Virtual Models for evaluating the properties of chemicals within a global architecture, version 1.1.4, 2017)

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

https://www.vegahub.eu/

5.4.Limits of applicability:

Gobal AD Index intervals:

ADI < 0.6:predicted substance is out of the the Applicability Domain of the model.

0.6<= ADI < 0.8: predicted substance could be out of the Applicability Domain of the model.

AD >=0.8: predicted substance is into the Applicability Domain of the model

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: Yes

INChI: No

MOL file: No

NanoMaterial: null

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 two separate carcinogenicity datasets are provided in supporting information. The combined training set used for the extraction of the rules and model validation consists of 986 compounds (734 carcinogens, 252 non-carcinogens).

6.6.Pre-processing of data before modelling:

No information available

6.7.Statistics for goodness-of-fit:

Accuracy = 0.73; Specificity = 0.60; Sensitivity = 0.78

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

n/a

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

n/a

6.10.Robustness - Statistics obtained by Y-scrambling:

n/a

6.11.Robustness - Statistics obtained by bootstrap:

n/a

6.12.Robustness - Statistics obtained by other methods:

Five-fold cross-validation:

Accuracy = 73%; Sensitivity = 77%; Specificity = 41%.

TP = 562/735; TN = 157/254; FP = 95/254; FN = 172/735.

MCC = 0.36

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: null

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 predictability of the model has been evaluated on two external test sets: 1) ANTARES dataset, a collection of chemical rat carcinogenesis data (presence of carcinogenic effects in male or female rats) obtained from the EU-funded project CAESAR dataset and the “FDA 2009 SAR Carcinogenicity—SAR Structures” database (1543 compounds); 2) ECHA dataset, carcinogenicity data collected from the eChemPortal inventory (258 compounds) [3].

The split among training and test appears within VEGA. Each substance identified in the dataset is labelled if it is in the training or test set.

7.6.Experimental design of test set:

n/a

7.7.Predictivity - Statistics obtained by external validation:

1) External validation on ANTARES dataset:
Accuracy = 59%; Sensitivity = 77%; Specificity = 41%; TP = 599/783; TN = 315/760; FP = 445/760; FN = 184/783; MCC = 0.19.

2) External validation on ECHA dataset:
Accuracy = 64%; Sensitivity = 48%; Specificity = 72%; TP = 43/89; TN = 121/169; FP = 48169; FN = 46/89; MCC = 0.20.

7.8.Predictivity - Assessment of the external validation set:

n/a

7.9.Comments on the external validation of the model:

Accuracy, sensitivity, specificity, and the MCC for the external evaluation are determined using SARpy. Although the external evaluation is considered the best mean for the assessment of the predictive ability of a (Q)SAR model, the results of the external evaluation of any model are highly related to the relative similarity of the external evaluation set in relation to the training set.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The extracted 43 rules ("active" fragments, or structural alerts) represent structural fragments associated to carcinogenic acitivity.Inhibiting conditions are identified which prevent the appearance of the effect, even in presence of the active fragment.

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori: rules (i.e., "active" fragments, or structural alerts) are extracted by SARpy software and not defined a priori.

8.3.Other information about the mechanistic interpretation:

Additional information on fragments analysis are provided in the original publication [3].

9.Miscellaneous information

9.1.Comments:

n/a

9.2.Bibliography:

[1]Benigni R, Battistelli CL, Bossa C, Tcheremenskaia O, Crettaz P (2013) New perspectives in toxicological information management, and the role of ISSTOX databases in assessing chemical mutagenicity and carcinogenicity. Mutagenesis 28 (4), 401–409. doi:10.1093/mutage/get016 https://academic.oup.com/mutage/article/28/4/401/2459896

[2]Kirkland D, Zeiger E, Madia F, Corvi R (2014) Can in vitro mammalian cell genotoxicity test results be used tocomplement positive results in the Ames test and help predictcarcinogenic or in vivo genotoxic activity? II. Construction andanalysis of a consolidated database. Mutation Research 775–776, 69–80. dx.doi.org/10.1016/j.mrgentox.2014.10.006

[3]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 http://dx.doi.org/10.1080/10590501.2016.1166879

9.3.Supporting information:

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

10.Summary (ECB Inventory)

10.1.QMRF number:

To be entered by ECB

10.2.Publication date:

To be entered by ECB

10.3.Keywords:

To be entered by ECB

10.4.Comments:

To be entered by ECB