1. QSAR identifier

1.1. QSAR identifier (title):
Skin Sensitization model (IRFMN/JRC) 1.0.0

1.2. Other related models:
Skin sensitization model (CAESAR) version 2.1.6

1.3. Software coding the model:
Skin sensitization model (IRFMN/JRC) version 1.0.0

2. General information

2.1. Date of QMRF:
05/06/2019

2.2. QMRF author(s) and contact details:
Gianluca Selvestrel Istituto di Ricerche Farmacologiche Mario Negri - IRCCS
gianluca.selvestrel@marionegri.it

2.3. Date of QMRF update(s):
06/02/2020

2.4. QMRF update(s):
Modification of section 9.3

2.5. Model developer(s) and contact details:
[1] Serena Manganelli
[2] David Asturiol

2.6. Date of model development and/or publication:
November 2017

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

2.8. Availability of information about the model:
Available through the VEGA software; free download of software is possible on VEGA website (https://www.vegahub.eu).
Training and Test sets are also available (see 9.3)

2.9. Availability of another QMRF for exactly the same model:
Other QMRF for this model are not available.

3. Defining the endpoint - OECD Principle 1

3.1. Species:
CBA mice

3.2. Endpoint:
Skin sensitization Skin sensitization on mouse (Local Lymph Node Assay)

3.3. Comment on endpoint:
Skin sensitizers are substances able to elicit an allergic response following contact with the skin, termed allergic contact dermatitis
(ACD) in humans. The LLNA assay evaluates the induction phase of the allergenic response. Molecules have been classified as "sensitizers" or "non sensitizers".

3.4. Endpoint units:
Adimensional

3.5. Dependent variable:

3.6. Experimental protocol:
OECD 429 Test

3.7. Endpoint data quality and variability:
See training and test sets in section 9.3

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:
Decision trees (rpart R package) based on 2D DRAGON descriptors: nDB, IC1, SIC2, GATS8s, nRCHO, CATS2D_01_NL, F04[C-O], MLOGP

4.2. Explicit algorithm:
Decision tree created using rpart R packages, as encoded in VEGA

4.3. Descriptors in the model:

1. nDB number of double bonds
2. IC1 Information Content index (neighborhood symmetry of 1-order)
3. SIC2 Structural Information Content index (neighborhood symmetry of 2-order)
4. GATS8s Geary autocorrelation of lag 8 weighted by I-state
5. nRCHO number of aldehydes (aliphatic)
6. CATS2D_01_NL CATS2D Negative-Lipophilic at lag 01
7. F04[C-O] Frequency of C - O at topological distance 4
8. MLOGP Moriguchi octanol-water partition coeff. (logP)

4.4. Descriptor selection:
An in-house tool developed in the R statistical platform has been used to select the best descriptors set and size to be employed for the final model. The approach was based on a forward selection technique: starting from the descriptor most correlated with the experimental data, at each iteration the descriptor leading to the best model (among all the available descriptors) was added, until the size of 25 descriptors. Models have been built, with a linear discriminant analysis modelling, applied with a bootstrap cross-validation approach (n = 100). For each model, the fitness function has been calculated as the mean value between specificity and sensitivity obtained from the models built in each bootstrap iteration. This fitness function has been used to select the best descriptor to be added to proceed to the next iteration. From this procedure, the set of descriptors with the best cross-validation values has been chosen as the set to be used for the final model. The "best" values have been considered taking into account their trend: by progressively adding descriptors to the model, the cross-validation performances increase until a plateau (and, following, a decrease), which means that the optimal number of descriptors have been reached, and adding
4.5. **Algorithm and descriptor generation:**
VEGA platform, where the descriptors have been implemented following the definition available in the Dragon software.

4.6. **Software name and version for descriptor generation:**
DRAGON

Dragon is an application for the calculation of molecular descriptors (MDs) and fingerprints (FPs). These can be used to evaluate molecular structure-activity or structure-property relationships, as well as for similarity analysis and high-throughput screening of molecule databases.

https://chm.kode-solutions.net

VEGA
http://www.vegahub.eu (Emilio Benfenati, Alberto Manganaro)
http://www.vegahub.eu

4.7. **Chemicals/Descriptors ratio:**
264/8 = 33

5. **Defining the applicability domain - OECD Principle 3**

5.1. **Description of the applicability domain of the model:**
The applicability domain of the model implemented in VEGA v. 1.1.5 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). For each index, including the final ADI, three intervals for its values are defined, such that the first interval corresponds to a positive evaluation, the second one corresponds to a suspicious evaluation and the last one corresponds to a negative evaluation. Following, all applicability domain components are reported along with their explanation.

Similar molecules with known experimental value. This index takes into account how similar are the first two 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.

Accuracy (average error) of prediction for similar molecules. This index takes into account the error in prediction for the two most similar compounds found. Values near 0 mean that the predicted compounds falls in an area of the model's space where the model gives reliable predictions, otherwise the greater is the value, the worse the model
behaves.
Concordance with similar molecules (average difference between target compound prediction and experimental values of similar molecules). This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made agrees with the experimental values found in the model's space, thus the prediction is reliable.
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, 0,6 if a fragments is found, 0,4 if more than 1 fragments is found. Then, the final index is given as the product RARE NOT FOUND.
Model descriptors range check. This index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range.
Global AD Index. The final global index takes into account all the previous indices, in order to give a general global assessment on the applicability domain for the predicted compound.

5.2. Method used to assess the applicability domain:
The applicability domain may be assessed through the VEGA v 1.1.5 - Skin Sensitization model (IRFMN/JRC) v. 1.0.0

5.3. Software name and version for applicability domain assessment:
VEGA v 1.1.5
http://www.vegahub.eu

5.4. Limits of applicability:
The model was built only for organic chemicals

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:
Unknown

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

6.5. Other information about the training set:
N/A

6.6. Pre-processing of data before modelling:

6.7. Statistics for goodness-of-fit:

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

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:
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:
Unknown

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

7.5. Other information about the external validation set:

7.6. Experimental design of test set:
The dataset were randomly split into training and test set with respectively the 80% and the 20% of the compounds.
Training set= 264
Test set= 68

7.7. Predictivity - Statistics obtained by external validation:
Accuracy: 0.71
Specificity: 0.82
Sensitivity: 0.65
MCC: 0.44

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:
Some reasons of activity related to chemical moieties have been identified a posteriori.

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:
9.3. Supporting information:
   Training set(s)
   | Training set Skin Sensitization IRFMN JRC.txt | file:///C:\Users\GSelvestrel\Desktop\QMRF QPRF\Training set Skin Sensitization IRFMN JRC.txt |
   Test set(s)
   | Test set Skin Sensitization IRFMN JRC.txt | file:///C:\Users\GSelvestrel\Desktop\QMRF QPRF\Test set Skin Sensitization IRFMN JRC.txt |

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