	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Androgen Receptor-mediated effect (IRFMN/COMPARA) (version1.0.1)**
	Printing Date: **Apr 3, 2020**

1.QSAR identifier

1.1.QSAR identifier (title):

Androgen Receptor-mediated effect (IRFMN/COMPARA) (version1.0.1)

1.2.Other related models:

NA

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:

14/02/20

2.2.QMRF author(s) and contact details:

Alessandra Roncaglioni Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy alessandra.roncaglioni@marionegri.it https://www.marionegri.it/

2.3.Date of QMRF update(s):

NA

2.4.QMRF update(s):

NA

2.5.Model developer(s) and contact details:

[1] Alessandra Roncaglioni Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy alessandra.roncaglioni@marionegri.it https://www.marionegri.it/

[2] Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy alberto.manganaro@marionegri.it https://www.marionegri.it/

2.6.Date of model development and/or publication:

19 December 2018.

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

[1] Serena Manganelli, Alessandra Roncaglioni, Kamel Mansouri, Richard S. Judson, Emilio Benfenati, Alberto Manganaro, Patricia Ruiz “Development, validation and integration of in silico models to identify androgen active chemicals”, Chemosphere, Volume 220, 2019, Pages 204-215.

[2] 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:

Homo Sapiens

3.2.Endpoint:

Endocrine disrupting chemicals Nuclear receptor-mediated endocrine disruption Androgen receptor mediated effect

3.3.Comment on endpoint:

Effect of compound on Androgen receptor (classification)

3.4.Endpoint units:

Model is a classification so there are no units, possible results should be Active/NON-Active

3.5.Dependent variable:

NA

3.6.Experimental protocol:

Model are built based on in vitro high-throughput screening (HTS) assays measuring activity of chemicals at multiple points along the androgen receptor (AR) activity pathway

3.7.Endpoint data quality and variability:

1689 curated chemical structures with AR experimental activity were provided by the EPA's National Center for Computational Toxicology as a training set to develop the in silico models. Experimental data were derived from a collection of 11 in vitro HTS assays exploring multiple points in the AR pathway including three receptor binding, two cofactor recruitment, one RNA transcription, three agonist-mode protein production and two antagonist-mode protein production. A chemical was considered as a binder if it was either an active agonist or antagonist. For more information about the ToxCast assays and how they were combined see[1] .

Chemical data were collected from the open free library ToxCast/Tox21(https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data).

From the initial list of 1,689, we removed a duplicate structure and twenty-one chemicals identified by the EPA as potentially false negatives based on a bootstrapping study, obtaining a final list of 1664 compounds.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The model provides a qualitative prediction for Androgen Receptor (AR) effects mediated through the AR pathway. The data were used to generate binary classification models to discriminate active (both agonists and antagonists) compounds from inactive ones. The model provides a qualitative prediction for Androgen Receptor (AR) effects mediated through the AR pathway. The data were used to generate binary classification models to discriminate active (both agonists and antagonists) compounds from inactive ones. The model provides a qualitative prediction for Androgen Receptor (AR) effects mediated through the AR pathway. The data were used to generate binary classification models to discriminate active (both agonists and antagonists) compounds from inactive ones.

4.2.Explicit algorithm:

Structural fragment features

It is a two steps model developed with SARpy. In the first step SARpy was used to model the two classes, identifying a set of 127 rules (17 for active and 110 for inactive). Then, a second set of 22rules identifying active compounds is applied to unpredicted compounds only. Compounds are labelled as inactive if matched ether by an inactive SA in the first step or not matched by the latter step. More precisely: If a

compound matches active class when applying the first ruleset the output is Active; If a compound matches inactive class when applying the first ruleset the output is Inactive; If a compound matches active class when applying the second ruleset the output is Probably Active; If a compound does not match any class when applying the first ruleset nor the second one the output is Probably Inactive

4.3.Descriptors in the model:

The model is a structure-based model and does not make use of descriptors.

4.4.Descriptor selection:

NA

4.5.Algorithm and descriptor generation:

The first set of structural alerts was extracted relative only for the active class. The second set of structural alerts was extracted taking in account both class. Fragment are generated using SARPy software

4.6.Software name and version for descriptor generation:

NA

4.7.Chemicals/Descriptors ratio:

1667 chemicals/ 149 structural alerts

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 and is defined in this way for this QSAR model´s predictions:

If 1 ≥ AD index > 0.9, the predicted substance is regarded in the Applicability Domain of the model

If 0.9 ≥ AD index > 0.65, the predicted substance could be out of the Applicability Domain of the model

If AD index ≤ 0.65, the predicted substance is regarded out of the Applicability Domain of the model

5.2.Method used to assess the applicability domain:

The Applicability Domain and 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 [2]. The VEGA 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.

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. Defined intervals are:

If 1 ≥ index > 0.8, strongly similar compounds with known experimental value in the training set have been found.

If 0.8 ≥ index > 0.6, only moderately similar compounds with known experimental value in the training set have been found.

If index ≤ 0.6 , no similar compounds with known experimental value in the training set have been found.

Accuracy (average error) of prediction for similar molecules:

This index takes into account the classification accuracy in prediction for the two 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. Defined intervals are:

If index < 0.6, accuracy of prediction for similar molecules found in the training set is good

If 0.8 > index ≥ 0.6, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 0.8, accuracy of prediction for similar molecules found in the training set is not adequate

Concordance for 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 disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are:

If index < 0.6, molecules found in the training set have experimental values that agree with the target compound predicted value

If 0.8 > index ≥ 0.6, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 0.8, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

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:

If index = 1, all atom centered fragment of the compound have been found in the compounds of the training set

If 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

If 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

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. Defined intervals are:

Index = TRUE, descriptors for this compound have values inside the descriptor range of the compounds of the training set

Index = FALSE, descriptors for this compound have values outside the descriptor range of the compounds of the training set

5.3.Software name and version for applicability domain assessment:

VEGA (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:

No

6.2.Available information for the training set:

NA

6.3.Data for each descriptor variable for the training set:

All

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

All

6.5.Other information about the training set:

NA

6.6.Pre-processing of data before modelling:

NA

6.7.Statistics for goodness-of-fit:

Active/Inactive prediction

Training set: accuracy 0.94, sensitivity 0.77, specificity 0.96, MCC 0.70. TP 155, TN 1402, FP 65, FN 42

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:

No

7.2.Available information for the external validation set:

NA

7.3.Data for each descriptor variable for the external validation set:

NA

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

NA

7.5.Other information about the external validation set:

NA

7.6.Experimental design of test set:

NA

7.7.Predictivity - Statistics obtained by external validation:

NA

7.8.Predictivity - Assessment of the external validation set:

NA

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:

NA

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] Kleinstreuer NC, Ceger P, Watt ED, Martin M, Houck K, Browne P, Thomas RS, Casey WM, Dix DJ, Allen D and others. 2017. Development and Validation of a Computational Model for Androgen Receptor Activity. Chem Res Toxicol 30(4):946–964

[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

9.3.Supporting information:

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

All available dataset 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