	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Chromosomal aberration model (CORAL) - v 1.0.1**
	Printing Date: **June 1, 2022**

1.QSAR identifier

1.1. QSAR identifier (title):

Chromosomal aberration model (CORAL) - v 1.0.1

1.2. Other related models:

No

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:

June 1, 2022

2.2. QMRF author(s) and contact details:

[1] Andrey A. Toropov, Istituto di Ricerche Farmacologiche Mario Negri IRCSS Via Mario Negri 2, 20156 Milano, Italy (andrey.toropov@marionegri.it) https://www.marionegri.it/

[2] Alla P. Toropova, Istituto di Ricerche Farmacologiche Mario Negri IRCSS Via Mario Negri 2, 20156 Milano, Italy (alla.toropova@marionegri.it) https://www.marionegri.it/

[3] Giuseppa Raitano, Istituto di Ricerche Farmacologiche Mario Negri IRCSS Via Mario Negri 2, 20156 Milano, Italy (giuseppa.raitano @marionegri.it) https://www.marionegri.it/

[4] Erika Colombo , Istituto di Ricerche Farmacologiche Mario Negri IRCSS Via Mario Negri 2, 20156 Milano, Italy (erika.colombo@marionegri.it) https://www.marionegri.it/

2.3. Date of QMRF update(s):

No update

2.4. QMRF update(s):

No update

2.5. Model developer(s) and contact details:

[1] Andrey A. Toropov, Istituto di Ricerche Farmacologiche Mario Negri IRCSS Via Mario Negri 2, 20156 Milano, Italy (andrey.toropov@marionegri.it) https://www.marionegri.it/

[2] Alla P. Toropova, Istituto di Ricerche Farmacologiche Mario Negri IRCSS Via Mario Negri 2, 20156 Milano, Italy (alla.toropova@marionegri.it) https://www.marionegri.it/

[3] Giuseppa Raitano, Istituto di Ricerche Farmacologiche Mario Negri IRCSS Via Mario Negri 2, 20156 Milano, Italy (giuseppa.raitano @marionegri.it) https://www.marionegri.it/

2.6. Date of model development and/or publication:

May 9, 2018

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

[1] A.A. Toropov, A.P. Toropova, G. Raitano, E. Benfenati, CORAL: building up QSAR models for the chromosome aberration test. Saudi Journal of Biological Sciences, 26 (2019), 1101–1106. https://doi.org/10.1016/j.sjbs.2018.05.013

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:

Data for chromosomal aberrations determined by in vitro test using Chinese hamster lung (CHL) and ovary (CHO) cells, with and without metabolic activation (metabolic system S9).

3.2. Endpoint:

Chromosomal aberrations

3.3. Comment on endpoint:

Structural aberrations may be of two types, chromosome or chromatid. Polyploidy (including endoreduplication) could arise in chromosome aberration assays in vitro. While aneugens can induce polyploidy, polyploidy alone does not indicate aneugenic potential and can simply indicate cell cycle perturbation or cytotoxicity. Chromosome aberrations, including breakage, rearrangement, and change in the number of chromosomes in the metaphase, have been associated with carcinogenicity and developmental effects.

3.4. Endpoint units:

The model provides a quantitative prediction for chromosomal aberrations in a continuous number between +1 (active) and -1 (inactive). The threshold to discriminate between active and inactive is 0 (< 0 is inactive; > 0 is active).

3.5. Dependent variable:

One-variable model based on optimal descriptor.

3.6. Experimental protocol:

According to OECD 473 guideline, the purpose of the in vitro chromosomal aberration test is to identify substances that cause structural chromosomal aberrations in cultured mammalian cells. This test is not designed to measure aneuploidy. The analysis of chromosomal aberration induction should be done using cells in metaphase. It is thus essential that cells should reach mitosis both in treated and in untreated cultures. Cell cultures of human or other mammalian origin are exposed to the test chemical both with and without an exogenous source of metabolic activation unless cells with an adequate metabolizing capability are used. At an appropriate predetermined interval after the start of exposure of cell cultures to the test chemical, they are treated with a metaphase-arresting substance (e.g. Colcemid® or colchicine), harvested, stained and metaphase cells are analyzed microscopically for the presence of chromatid-type and chromosome-type aberrations.

3.7. Endpoint data quality and variability:

Experimental data for this work were taken from the Genotoxicity OASIS Database (http://oasis-lmc.org/products/databases/rat-liver-metabolism-extended.aspx) and the Toxicity Japan MHLW (http://dra4.nihs.go.jp/mhlw_data/jsp/SearchPageENG.jsp) that include data for chromosomal aberrations determined by in vitro test using Chinese hamster lung (CHL) and ovary (CHO) cells, with and without metabolic activation (metabolic system S9). Data were selected on the basis of their compliance with OECD 473 guideline.
After removing duplicates we collected a set of 477 organic compounds: 223 are classified as active and 254 are classified as inactive in chromosomal aberrations test. For each compound, CAS number, simplified molecular input-line entry system (SMILES) and experimental data expressed as active (+1) or inactive (−1) are represented. Finally, SMILES have been normalized by the VEGA platform (www.vega-qsar.eu/). These compounds were randomly split into the training (80%), calibration (10%), and validation (10%) sets (five splits are examined).
After the implementation in VEGA, the dataset was split in training (442 chemicals) and test (35 chemicals)

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

One-variable model based on SMILES-derived descriptors.

4.2. Explicit algorithm:

CORAL (http://www.insilico.eu/coral/), which is based on the Monte Carlo method.

Chromosomal aberrations = -0.604 (± 0.002) + 0.0614 (± 0.0001) * DCW(1,13)

$$Category = \begin{cases} active & if, y \geq 0.5 \\ inactive & if, y < 0.5 \end{cases}$$

4.3. Descriptors in the model:

SMILES-derived optimal descriptor DCW(T*,N*)

4.4. Descriptor selection:

SMILES-derived optimal descriptor, according to the equation:

$$DCW(T*, N*) = \sum CW(S_k) + \sum CW(SS_k) + CW(HARD)$$

The “descriptors”, SMILES attributes, are based on SMILES, and refer to short sequences of characters (up to three). The model identifies these SMILES attributes which are statistically significant. The other are excluded. In this way these SMILES attributes are selected, through the Monte Carlo method.

4.5. Algorithm and descriptor generation:

The Monte Carlo method

4.6. Software name and version for descriptor generation:

CORAL-2017 (http://www.insilico.eu/coral/)

4.7. Chemicals/Descriptors ratio:

One-variable model

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 ≥ AD index > 0.8, the predicted substance is into the Applicability Domain of the model. It corresponds to good reliability of prediction.

If 0.8 ≥ AD index > 0.6, the predicted substance could be out of the Applicability Domain of the model. It corresponds to moderate reliability of prediction.

If AD index ≤ 0.6, the predicted substance is out of the Applicability Domain of the model and corresponds to low reliability of prediction.

5.2. Method used to assess the applicability domain:

The Applicability Domain and 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 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 centred 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 > 0.80, strongly similar compounds with known experimental value in the training set have been found.

If 0.80 ≥ 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: 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:

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

5.3. Software name and version for applicability domain assessment:

CORAL-2017 (http://www.insilico.eu/coral/)

5.4. Limits of applicability:

Qualitative

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: No

INChI: No

MOL file: No

NanoMaterial: No

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

No (there are not descriptors, the system uses the SMILES).

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

Yes

6.5. Other information about the training set:

The initial dataset was divided five times into sets (training, calibration, and validation) using a random distribution.

6.6. Pre-processing of data before modelling:

The SMILES have been normalized by the VEGA platform (www.vegahub.eu/).

6.7. Statistics for goodness-of-fit:

Sensitivity, specificity, accuracy, MCC, as in the table.

Split	Set	n	Sensitivity	Specificity	Accuracy	MCC
1	Training	399	0.7592	0.7981	0.7794	0.5578
	Calibration	39	0.8333	0.8667	0.8462	0.6868
2	Training	407	0.7016	0.8009	0.7543	0.5059
	Calibration	35	0.9375	0.9471	0.9429	0.8849
3	Training	380	0.7348	0.7889	0.7632	0.5248
	Calibration	49	0.9333	0.8235	0.8571	0.7097
4	Training	398	0.7513	0.7707	0.7613	0.5221
	Calibration	40	0.9412	0.9565	0.9500	0.8977
5	Training	399	0.6742	0.8326	0.7619	0.5156
	Calibration	39	0.7600	1.000	0.8462	0.7294

After the implementation in VEGA:

Training set: n = 442, Balanced Accuracy 0.77, Sensitivity 0.72, Specificity 0.81, MCC 0.54. TP 149, TN 191, FP 44, FN 58.

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

Non-available

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

Non-available

6.10. Robustness - Statistics obtained by Y-scrambling:

Non-available

6.11. Robustness - Statistics obtained by bootstrap:

Non-available

6.12. Robustness - Statistics obtained by other methods:

Non-available

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: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

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

No (there are not descriptors, the system uses the SMILES).

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

Yes

7.5. Other information about the external validation set:

Non-available

7.6. Experimental design of test set:

Available

7.7. Predictivity - Statistics obtained by external validation:

Available, as below.

Split	n	Sensitivity	Specificity	Accuracy	MCC
1	39	0.8750	0.8387	0.8462	0.6244
2	35	0.8750	1.000	0.9429	0.8898
3	48	0.8148	1.000	0.8958	0.8112
4	39	1.000	0.6923	0.7949	0.6574
5	39	0.8500	0.9474	0.8974	0.7995

After the implementation in VEGA:
Test set: n = 35, Balanced Accuracy 0.94, Sensitivity 0.88, Specificity 1.00, MCC 0.89. TP 14, TN 19, FP 0, FN 2

Test set in AD: n = 14, Balance Accuracy 1, Sensitivity 1, Specificity 1, MCC 1, TP 8, TN 6

Test set could be out of AD: n = 17, Balanced accuracy 0.83, Sensitivity 0.67, Specificity 1, MCC 0.75. TP 4, TN 11, FP 0, FN 2

Test set out of AD: n = 4, Balanced accuracy 1, Sensitivity 1, Specificity 1, MCC 1, TP 2, TN 2

7.8. Predictivity - Assessment of the external validation set:

Yes

7.9. Comments on the external validation of the model:

It has been obtained through a random selection; the data on validation set is unavailable during modelling process.

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

Analysis of results on several runs of the Monte Carlo optimization, thus statistical, not mechanistic, basis.

8.2. A priori or a posteriori mechanistic interpretation:

A posteriori only

8.3. Other information about the mechanistic interpretation:

List of molecular features which are promoters of increase or decrease for endpoint.

9.Miscellaneous information

9.1. Comments:

Manual for the CORAL software available on the Internet.

9.2. Bibliography:

[1] A.A. Toropov, A.P. Toropova, G. Raitano, E. Benfenati, CORAL: building up QSAR models for the chromosome aberration test. Saudi Journal of Biological Sciences, 26 (2019) 1101–1106. https://doi.org/10.1016/j.sjbs.2018.05.013

[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