


| | |
|-----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|
|  | QMRf identifier (JRC Inventory): To be entered by JRC |
| | QMRf Title: ACD/Percepta (Impurity Profiling) QSAR model for microbial in vitro Salmonella (composite) |
| | Printing Date: 2018-03-26 |
| | |

1. QSAR identifier

1.1. QSAR identifier (title):

ACD/Percepta (Impurity Profiling) QSAR model for microbial in vitro
Salmonella (composite)

1.2. Other related models:

ACD/Percepta Impurity Profiling package, including probabilistic models
for 21 different endpoints related to:

- 1) Genetic toxicity: Mutagenicity (Ames test, Mouse Lymphoma Assay, CHO/CHL all loci composite, and other standard assays), Clastogenicity (Micronucleus test, Chromosomal Aberrations), DNA damage (Unscheduled DNA Synthesis)
- 2) Carcinogenicity (rodent carcinogenicity)
- 3) Reproductive toxicity: Endocrine disruption mechanisms (estrogen receptor binding)

1.3. Software coding the model:

ACD/Labs Percepta (2017 Release) - Impurity Profiling Module

Impurity profiling module is a result of the collaboration between ACD/Labs and FDA Center for Food Safety and Nutrition (CFSAN). This module consists of a battery of probabilistic models, supported by a knowledge-based expert system, for the evaluation of genotoxic and carcinogenic potential of chemicals.

Advanced Chemistry Development, Inc. (ACD/Labs). 8 King Street East, Suite 107, Toronto, Ontario, Canada M5C 1B5. info@acdlabs.com

<http://www.acdlabs.com/>

2. General information

2.1. Date of QMRf:

26 March 2018

2.2. QMRf author(s) and contact details:

[1] Simona Kovarich S-IN Soluzioni Informatiche Via Ferrari 14, I-36100 Vicenza
simona.kovarich@s-in.com <http://www.s-in.it/>

[2] Kiril Lanevskij ACD/Labs, Inc. ACD/Labs, Inc., A. Mickevicius g. 29, LT-08117 Vilnius, Lithuania
kiril.lanevskij@acdlabs.com www.acdlabs.com

2.3. Date of QMRf update(s):

Not Applicable - this is a new QMRf

2.4. QMRf update(s):

Not Applicable

2.5. Model developer(s) and contact details:

[1] Kiril Lanevskij ACD/Labs, Inc. ACD/Labs, Inc., A. Mickevicius g. 29, LT-08117 Vilnius, Lithuania
info@acdlabs.com www.acdlabs.com

[2] Liutauras Juska 1) ACD/Labs, Inc.; 2) Department of Biochemistry and Biophysics, Vilnius University. 1) ACD/Labs, Inc., A. Mickevicius g. 29, LT-08117 Vilnius, Lithuania; 2) Department of

Biochemistry and Biophysics, Vilnius University, M.K.Ciurlionio g. 21/27, LT-03101 Vilnius, Lithuania
info@acdlabs.com www.acdlabs.com

[3]Justas Dapkunas ACD/Labs, Inc. ACD/Labs, Inc., A.Mickevicius g. 29, LT-08117 Vilnius,
Lithuania info@acdlabs.com www.acdlabs.com

[4]Andrius Sazonovas ACD/Labs, Inc. ACD/Labs, Inc., A.Mickevicius g. 29, LT-08117 Vilnius,
Lithuania info@acdlabs.com www.acdlabs.com

[5]Pranas Japertas ACD/Labs, Inc. ACD/Labs, Inc., A.Mickevicius g. 29, LT-08117 Vilnius,
Lithuania info@acdlabs.com www.acdlabs.com

[6]Remigijus Didziapetris ACD/Labs, Inc. ACD/Labs, Inc., A.Mickevicius g. 29, LT-08117 Vilnius,
Lithuania info@acdlabs.com www.acdlabs.com

2.6.Date of model development and/or publication:

2011

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

[1]ACD/Labs Percepta - Impurity Profiling Module

<http://www.acdlabs.com/products/percepta/impurities.php>

[2]Japertas P et al. A comprehensive approach for in silico risk assessment of impurities and
degradants in drug products. Toxicol Lett. 2011, 205, S95.

2.8.Availability of information about the model:

The model is proprietary, training and test set are not available.

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

Not to date

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Salmonella typhimurium

3.2.Endpoint:

QMRF 4.10. Mutagenicity OECD 471 Bacterial Reverse Mutation Test

3.3.Comment on endpoint:

Mutagenicity assessment based on bacterial reverse mutation test using
different strains of Salmonella typhimurium.

3.4.Endpoint units:

Not applicable

3.5.Dependent variable:

Mutagenicity as microbial *in vitro* Salmonella (composite)
gene mutation assay is modelled for study calls, where the positive calls
are trained as binary 1 and negative calls as binary 0. The output of the
probabilistic QSAR model consists of: the probability that a compound will
result in a positive test in the respective assay ("p-value"); an
indication of whether the compound belongs to the model applicability
domain according to the calculated RI value; and a "positive" or
"negative" call if the compound can be reliably classified on the basis of
p and RI values ("Undefined" otherwise)

3.6.Experimental protocol:

Experimental dataset was obtained from FDA. Data was collected from
EPA GENE-TOX database and scientific literature [3].For
modeling purposes experimental results of microbial in vitro Salmonella

(composite) Assay have been transformed into a binary variable, i.e. positive/negative.

3.7. Endpoint data quality and variability:

No information available

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

QSAR

4.2. Explicit algorithm:

Probabilistic Model (based on GALAS Methodology)

GALAS (Global, Adjusted Locally According to Similarity) modeling methodology. The GALAS model consists of two parts: 1) Global (baseline) model, built using binomial PLS method based on fragmental descriptors, that reflects a "cumulative" mutagenicity potential. 2) Local corrections are applied to baseline predictions using a special similarity-based routine, after performing an analysis for the most similar compounds used in the training set. Experimental values for microbial in vitro Salmonella assay are used during the local part of the modeling to yield final GALAS model.

4.3. Descriptors in the model:

404 fragmental descriptors are used for the development of the GALAS model (see 4.4)

4.4. Descriptor selection:

404 fragmental descriptors were used for the development of the GALAS model. The fragmental descriptor set was identified based on general knowledge and considerations regarding all possible chemical structures and include all the fragments, even those that are not detected in the training set molecules at all. The major part of the utilized fragment set was intended for the description of the general chemical constitution of any compound and comprised conventional fragmental descriptors, such as atoms, functional groups, molecular 'shape fragments', etc. This initial set was expanded with a group of more complex fragments, generally called toxicophores, i.e. substructures identified to be responsible for the toxic action of the molecules possessing them.

4.5. Algorithm and descriptor generation:

The GALAS modelling methodology was applied to derive the algorithm (see 4.2); 404 fragmental descriptors were used for the development of the GALAS model (see 4.4). The output of probabilistic model consists of a "p-value" (probability that a compound will result in a positive test in the microbial in vitro Salmonella gene mutation assay), "Coverage" (an indication of whether the compound belongs to the Model Applicability Domain according to the calculated RI value - see section 5), "Call (+ or -)" consisting of a "Positive" or "Negative" prediction if the compound can be reliably classified on the basis of p and RI values ("Undefined" otherwise)

4.6. Software name and version for descriptor generation:

Algorithm Builder 1.8 software (2006)

(Software used for model development)

Advanced Chemistry Development, Inc. (ACD/Labs). 8 King Street East, Suite 107, Toronto, Ontario,

Canada M5C 1B5.

<http://www.acdlabs.com/>

4.7. Chemicals/Descriptors ratio:

N/A

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

The confidence of predictions is evaluated via a Reliability Index (RI) calculated for each prediction. The RI is a value ranging from 0 and 1 (0 – unreliable prediction, 1 – idealistic, fully reliable prediction) and is an indicator of how well a particular compound is represented within the training set of the model. Two criteria are applied for reliability estimation:

- 1) Similarity of the analyzed molecular structure to compounds in the Self-training Library (prediction is considered unreliable if no similar compounds have been found in the training set).
- 2) Consistency of experimental data for similar compounds (inconsistent data for similar molecules lead to lower RI values).

RI can serve as a valuable tool for interpreting prediction results. If a compound obtains RI lower than a certain cut-off value (here set at 0.3), it means that this compound falls outside the applicability domain of the model and the respective prediction may be less accurate.

5.2. Method used to assess the applicability domain:

Applicability domain assessed using the Reliability Index. The Reliability Index (RI) is given as a product of two indices: $RI = SI \cdot DMCI$

- 1) SI (Similarity Index) evaluates how distant the query structure is from the whole training set, and is calculated by weighted averaging of all the individual Similarity Indices SI_j (i.e., calculated from the correlation of two predicted property value vectors for the test molecule and each of the five most similar compounds from the training set.)
- 2) DMCI (Data-model consistency index) accounts for the influence of consistency of experimental data with regard to the baseline model for the five most similar compounds on the calculations' reliability. DMCI is calculated by comparing the differences between experimental and global model-predicted baseline values for the individual most similar compounds and the suggested correction value for the test compound. The more individual differences are scattered around the calculated average, the more inconsistent are the data for the similar compounds with regards to the global baseline model.

5.3. Software name and version for applicability domain assessment:

ACD/Labs Percepta (2017 Release) - Impurity Profiling Module

Advanced Chemistry Development, Inc. (ACD/Labs). 8 King Street East, Suite 107, Toronto, Ontario, Canada M5C 1B5. info@acdlabs.com

<http://www.acdlabs.com/>

5.4. Limits of applicability:

RI < 0.3: unreliable prediction
0.3 < RI < 0.5: borderline reliability of prediction
0.5 < RI < 0.75: moderate reliable prediction
RI > 0.75: high reliable prediction

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

No

6.2. Available information for the training set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

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

No

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

No

6.5. Other information about the training set:

The entire dataset used in model development and validation consists of 7826 compounds, including 3875 positive compounds (i.e. 49.5%).

6.6. Pre-processing of data before modelling:

None

6.7. Statistics for goodness-of-fit:

Sensitivity = 85.6%; Specificity = 81.5%; Concordance = 83.6%

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:

N/A

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

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 part of the dataset used for model validation consists of 1577 compounds, including 794 positive compounds (i.e. 50.3%).

7.6.Experimental design of test set:

Random splitting of the initial dataset into the training and validation sets at about 80% to 20% ratio.

7.7.Predictivity - Statistics obtained by external validation:

Sensitivity = 87.1%; Specificity = 81.7%; Concordance = 84.6%

7.8.Predictivity - Assessment of the external validation set:

Only chemicals inside the Applicability Domain of the model (i.e. $RI > 0.3$) were considered for the calculation of statistical performances (1332 compounds, i.e., 84.5% of the entire test set).

7.9.Comments on the external validation of the model:

Compounds with unreliable predictions ($RI < 0.3$) were excluded from considerations, as by definition they fall outside of the model AD and hence provide no meaningful information about the models' performance.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model is based on both fragmental structural descriptors and toxicophores, i.e. substructures identified to be responsible for the toxic action of the molecules possessing them. To enhance a mechanistic understanding, predictions obtained by the probabilistic model can be combined with and supported by the Genotoxicity Hazard System, which is a knowledge-based expert system that identifies structural fragments that may be responsible for the mutagenic activity of the analyzed molecules.

8.2.A priori or a posteriori mechanistic interpretation:

A priori (see section 8.1).

8.3.Other information about the mechanistic interpretation:

The software displays up to 5 most similar structures (included in the training set of the model) to the analysed molecule with experimental results (positive/negative). The analysis of similar structures provides additional information to gain insight into the possible mechanisms of action and support the *in silico* prediction for the query compound.

9.Miscellaneous information

9.1.Comments:

ACD/Labs Package for Toxicity Screening of Impurities provides a battery of *in silico* tests to accurately assess the genotoxic and carcinogenic potential of impurities and degradants. The impurities package offers probabilistic predictive models for 21 different

endpoints that cover various mechanisms of hazardous activity (including Mutagenicity, Clastogenicity, DNA damage mechanisms, Carcinogenicity and Endocrine Disruption mechanisms). These predictors are supplemented with a knowledge-based expert system that identifies potentially hazardous structural fragments that could be responsible for genotoxic and/or carcinogenic activity of the compound of interest. The expert system was able to recognize >94% of mutagens in ACD/Ames test database, and >90% of compounds marked as potent carcinogens in the FDA's OFAS Food-Additive Knowledgebase.

9.2.Bibliography:

[1]ACD/Labs Percepta - Impurity Profiling Module

<http://www.acdlabs.com/products/percepta/impurities.php>

[2]Japertas P et al. A comprehensive approach for in silico risk assessment of impurities and degradants in drug products. Toxicol Lett. 2011, 205, S95.

[3]Matthews EJ et al. Regul Toxicol Pharmacol. 2007, 47, 115.

[4]Lanevskij K et al., 2012. An In Silico Test Battery for Rapid Evaluation of Genotoxic and Carcinogenic Potential of Chemicals. Poster (Mar 25, 2012, ACS Spring)

http://www.acdlabs.com/download/publ/2012/acss12_insilico.pdf

9.3.Supporting information:

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

10.Summary (JRC Inventory)

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