

(Q)SAR Model Reporting Format (QMRF)

(The present QMRF is prepared according the fields and *Help* recommendations of JRC implemented in QMRF v 1.3 and QMRF Editor v.2.0.0

(https://sourceforge.net/apps/mediawiki/qmrf/index.php?title=Main_Page)

Welcome

Model version: *In vitro* Chromosomal Aberrations v.18.18

Platform version: OASIS TIMES 2.31.2

Name: *In vitro* Chromosomal Aberrations

Author: LMC, University "Prof. As. Zlatarov", Bourgas, Bulgaria

Date: 1 December 2021

E-mail: omekenya@btu.bg

www: <http://www.oasis-lmc.org/>

Section 1. QSAR identifier

1.1. QSAR identifier (title)

In vitro Chromosomal Aberrations with S9 metabolic activation

1.2. Other related models

In vitro Ames mutagenicity with S9 metabolic activation

1.3. Software coding of the model

Model version: *In vitro* Chromosomal Aberrations v.18.18

Platform version: OASIS TIMES 2.31.2

Name: *In vitro* Chromosomal Aberrations

Developer: LMC, University "Prof. As. Zlatarov", Bourgas, Bulgaria

Coding language: Delphi 10.2

Section 2. Date of QMRF

2.1. Date of QMRF

1 December 2021

2.2. QMRF author(s) and contact details

Name: Laboratory of Mathematical Chemistry

Affiliation: Laboratory of Mathematical Chemistry, University "Prof. As. Zlatarov",
"Yakimov" St. #1, 8010 Bourgas, Bulgaria

URL: <http://www.oasis-lmc.org>

E-mail: omekenya@btu.bg

2.3. Date of QMRF update(s)

21 November 2014; 12 June 2015; 12 May 2016; 12 July 2016; 31 August 2016; 30 May 2017; 23 July 2018; 22 Jan 2020; 1 December 2021

2.4. QMRF update(s)

Information which has been modified:

Sections 2.8. Availability of information about the model; **Section 3.7.** Endpoint data quality and variability; **Section 4.2.** Explicit algorithm; **Section 4.3** Descriptors in the model; **Section 4.4.** Descriptor section; **Section 4.5.** Algorithm and descriptor generation; **Section 4.6.** Software name and version for descriptor generation; **Section 5.3** Software name and version for the applicability domain assessment; **Section 5.4.** Limits of applicability; **Section 6.** Internal validation – OECD Principle 4; **Section 6.1** Availability of the training set; **Section 6.4** Data for the dependent variable; **Section 6.7** Statistic for goodness-of-fit; **Section 6.9** Robustness – Statistics obtained by leave-many-out cross-validation; **Section 6.11.** Robustness - Statistics obtained by bootstrap; **Section 6.13.**

Comment on the internal validation of the model; **Section 7.** External validation – OECD Principle 4; **Section 8.1.** Mechanistic basis of the model.

2.5. Model developer(s) and contact details

Name: P. Petkov, A. Chapkanov, C. Kuseva, H. Ivanova, E. Kaloyanova, G. Dimitrova, D. Yordanova, R. Serafimova, M. Todorov, T. Pavlov, S. Kotov, E. Jacob, A. Aptula, O. Mekenyan

Affiliation: Laboratory of Mathematical Chemistry, University "Prof. As. Zlatarov",
"Yakimov" St. #1, 8010 Bourgas, BULGARIA

URL: <http://www.oasis-lmc.org>

E-mail: omekenya@btu.bg

2.6. Date of model development and/or publication

2007/2012

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

1. O. Mekenyan; M. Todorov; R. Serafimova; S. Stoeva; A. Aynur; R. Finking; E. Jacob. Identifying the structural requirements for chromosomal aberration by incorporating molecular flexibility and metabolic activation of chemicals. *Chem Res Toxicol*, 1927-1941, (2007).
2. O. Mekenyan, S. Dimitrov, T. Pavlov, G. Dimitrova, M. Todorov, P. Petkov & S. Kotov. 2012. Simulation of chemical metabolism for fate and hazard assessment. V. Mammalian hazard assessment, *SAR and QSAR in Environmental Research*, Vol. 23, 553-606.

2.8. Availability of information about the model

TIMES_CA mutagenicity model (+S9) is derived for identification of chemicals capable to interact with DNA and proteins. Training set of the model includes 930 chemicals (part of which are proprietary data) from different literature sources. The model is based on an alerting group approach addressing mutagenicity of parents and their generated metabolites *in vitro* liver S9 metabolic system. Details of the model is provided in the sections bellow as well as in the following link:

<http://oasis-lmc.org/products/models/human-health-endpoints/chromosomal-aberrations.aspx>

2.9. Availability of another QMRF for exactly the same model

Section 3. Defining the endpoint – OECD Principle 1

3.1. Species

Chemicals included in the training set of the TIMES_CA model are collected according to the recommendation in the OECD technical guideline 473 addressing: Chinese Hamster Ovary (CHO), Chinese Hamster lung V79 and Chinese Hamster Lung (CHL)/IU, TK6:

http://www.oecd-ilibrary.org/environment/test-no-473-in-vitro-mammalian-chromosomal-aberration-test_9789264224223-en

3.2. Endpoint

In vitro Mammalian Chromosome Aberration Test

According to JRC pre-classification list of endpoints:
No. 207 QMRF Human Health Effects, QMRF 4.10 Mutagenicity.

3.3. Comment on endpoint

The purpose of the *in vitro* chromosomal aberration (CA) test is to identify substances that cause structural chromosomal aberrations in cultured mammalian cells. Structural aberrations may be of two types, chromosome or chromatid which are detected by the *in vitro* CA test. Polyploidy 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. This test is not designed to measure aneuploidy.

3.4. Endpoint units

Qualitative – positive/ negative

3.5. Dependent variable

Obs. chromosomal aberrations with S9

3.6. Experimental protocol

OECD technical guideline 473: *in vitro* mammalian chromosomal aberration test

3.7. Endpoint data quality and variability

References associated with each documented mutagenicity data (except for proprietary data) included in the training set of the model are provided in [Appendix 1](#).

Section 4. Defining the algorithm – OECD Principle 2

4.1. Type of model

Structural alert based model

4.2. Explicit algorithm

Prediction of Chromosomal aberrations is based on modelling of the two events deemed to be crucial for the effect – interaction of the chemicals with DNA/proteins and their activation as a result of liver S9 metabolism.

CA mutagenicity predictions are obtained using an alerting group approach. Only alerts having clear interpretation of mechanisms leading to DNA mutagenicity are included in the model. To obtain predictions, a set of alerts (160) is applied on parents and each of the generated *in vitro* rat liver S9 metabolites. The *in vitro* S9 metabolic simulator is trained to reproduce documented maps for mammalian liver metabolism for 438 chemicals. Match of alerts either on parents or metabolites is sufficient for obtaining positive prediction. Chemicals are predicted to be mutagenic as parents only, parents and metabolites, or as metabolites only.

Details about the alerts included in the model are provided in the next sections.

4.3. Descriptors in the model

Descriptors in the model are structural boundaries associated with alerting groups related to interactions with DNA/proteins. Alerts in the TIMES Chromosomal aberrations (+S9) constitute expertly-derived sets of structural fragments incorporating knowledge for the interactions of chemicals (parents and metabolites) with DNA/proteins. Application of the alerts on the training set of the model forms fractions of representative chemicals for the alerts, i.e. so-called ‘local’ training sets. All chemicals captured by the alerts are considered as validation sets of the introduced expert knowledge addressing reactivity of chemicals with DNA/proteins. The procedure for obtaining local training sets includes applying the structural boundaries of the alert searching among all chemicals from the training set of the model after application of S9 metabolic simulator. According to this, local training sets contain parent chemicals in which general fragments are:

- o found in their structures;
- o not found in the parent structures but found in their metabolite(s).

Description of these alerts is provided in the next sections.

4.4. Descriptor section

Table 1 summarizes the main characteristics of each DNA and protein alert in TIMES_CA (+S9):

- Alert name (corresponding to the name of the chemical class which is addressed);
- Performance of alert (correct/incorrect predictions) which is estimated based on proportion of observed positive chemicals from all chemicals captured by the alert. Performance of each alert is provided with its confidence range. As smaller is the size of local training sets as wider are the confidence ranges and vice versa.
- P-values addressing the reliability of alert performance estimation and taking into account possible bias of positive/negative chemicals in the training set of the model. Low p-values could be obtained only if both are satisfied:

- The number of chemicals in local training set is high enough;
- The alert performance is significantly higher than the proportion of positive/negative chemicals in the model training set, i.e. so-called naïve alert.

Analogically, high p-values could be obtained in case of:

- Small number of local training set chemicals (1-2 chemicals); or
- Performance comparable to the performance of the naïve alert.

High performance associated with low *p-values* indicate for High Reliability of alerts.

The above statistical measures along with the underlying mathematical formalisms are discussed in details in **Section 6** (Internal validation).

Table 1. Main characteristics of the DNA alerts in the TIMES_CA model (+S9).

No.	Alert	Correct	Incorrect	Performance est. ¹⁾	<i>p-value</i> ¹⁾
1	N-Nitroso Compounds	40	1	0.953 (0.891 ÷ 0.999)	1.9 x 10 ⁻⁸
2	Alkylated nitrosoureas and nitrosoguanidines	18	0	0.950 (0.854 ÷ 1.000)	0.0001
3	Heterocyclic Aromatic Amines	15	0	0.941 (0.829 ÷ 1.000)	0.0004
4	N-Alkyl-N-nitrosocarbamates	15	0	0.941 (0.829 ÷ 1.000)	0.0004
5	Quinoid compounds-	46	2	0.940 (0.875 ÷ 0.993)	1.1 x 10 ⁻⁸
6	Fused-Ring Primary Aromatic Amines	14	0	0.938 (0.819 ÷ 1.000)	0.0007
7	alpha, beta - Unsaturated Carboxylic Acids and Esters	26	1	0.931	1.5 x 10 ⁻⁵

				(0.841 ÷ 0.998)	
8	Vicinal Dihaloalkanes	10	0	0.917 (0.762 ÷ 1.000)	0.0052
9	Arenesulphonamides	10	0	0.917 (0.762 ÷ 1.000)	0.0052
10	Hydroxylated phenols	43	3	0.917 (0.839 ÷ 0.984)	2.6 x 10 ⁻⁷
11	N-Substituted Aromatic Amines	20	1	0.913 (0.801 ÷ 0.998)	0.0003
12	Alkylphosphates, Alkylthiophosphates and Alkylphosphonates	8	0	0.900 (0.717 ÷ 1.000)	0.015
13	Haloalkane Derivatives Containing Chain Heteroatom	17	1	0.900 (0.772 ÷ 0.997)	0.0011
14	Polarized Haloalkene Derivatives	8	0	0.900 (0.717 ÷ 1.000)	0.015
15	Arenecarbonyl Compounds	8	0	0.900 (0.717 ÷ 1.000)	0.015
16	Polycyclic Aromatic Hydrocarbon, Naphthaleneimide and Carbazole Derivatives	7	0	0.899 (0.688 ÷ 1.000)	0.015
17	Nitrogen Mustard	7	0	0.889 (0.688 ÷ 1.000)	0.025
18	N-Nitrosoamine Derivatives	7	0	0.889 (0.688 ÷ 1.000)	0.025
19	Quinoneimines protein binding alert	15	1	0.940 (0.748 ÷ 0.997)	0.0027
20	Substituted Anilines	52	6	0.883 (0.802 ÷ 0.957)	5.1 x 10 ⁻⁷
21	Dicarbonyl Compounds	11	2	0.880 (0.605 ÷ 0.971)	0.051
22	N-Hydroxylamines	56	7	0.877 (0.797 ÷ 0.951)	3.8 x 10 ⁻⁷
23	Sulfonates and Sulfates	6	0	0.818 (0.652 ÷ 1.000)	0.042
24	Arenecarboxylic Acid Esters	6	0	0.818 (0.652 ÷ 1.000)	0.042
25	Halogenated Vicinal Hydrocarbons	13	1	0.875 (0.719 ÷ 0.996)	0.0069
26	Alpha-Activated Haloalkanes	19	2	0.870 (0.735 ÷ 0.983)	0.0019

27	Pyrimidines and Purines	12	1	0.867 (0.701 ÷ 0.996)	0.011
28	alpha, beta - Unsaturated Carbonyls and Related Compounds	24	3	0.862 (0.738 ÷ 0.971)	0.0009
29	Conjugated Nitroalkenes and Five-Membered Nitro- and Amino Heterocycles	5	0	0.857 (0.607 ÷ 1.000)	0.072
30	Haloalkane Derivatives with Labile Halogen	29	7	0.857 (0.742 ÷ 0.960)	0.0004
31	Dialkyl Alkylphosphonates	5	0	0.857 (0.607 ÷ 1.000)	0.072
32	Sulfonates and Sulfates protein binding	5	0	0.857 (0.607 ÷ 1.000)	0.072
33	Epoxides, Aziridines, Thiiranes and Oxetanes	45	7	0.852 (0.757 ÷ 0.939)	2.4 x 10 ⁻⁵
34	C-Nitroso compounds protein binding	44	7	0.849 (0.752 ÷ 0.938)	3.4 x 10 ⁻⁵
35	(Thio)Phosphates	10	1	0.846 (0.659 ÷ 0.994)	0.026
36	Substituted Phenols	54	9	0.846 (0.758 ÷ 0.928)	7.3 x 10 ⁻⁶
37	Epoxides, Aziridines and Sulfuranes	47	8	0.842 (0.747 ÷ 0.930)	3.2 x 10 ⁻⁵
38	C-Nitroso Compounds	39	7	0.833 (0.728 ÷ 0.931)	0.0002
39	Fused-Ring Nitroaromatics	4	0	0.833 (0.549 ÷ 1.000)	0.121
40	Nitrogen and Sulfur Mustards	4	0	0.833 (0.549 ÷ 1.000)	0.121
41	Quinoline Derivatives	4	0	0.833 (0.549 ÷ 1.000)	0.121
42	Quinones and Trihydroxybenzenes	14	2	0.833 (0.666 ÷ 0.977)	0.016
43	Benzoquinolines and Acridines	9	1	0.833 (0.632 ÷ 0.994)	0.041
44	Carbamates	14	2	0.833 (0.666 ÷ 0.977)	0.016
45	Isothiocyanates	4	0	0.833 (0.549 ÷ 1.000)	0.121
46	Pyrazolone and Pyrazolidine Derivatives	4	0	0.833	0.121

				(0.549 ÷ 1.000)	
47	Quinoneimine, Thionine and Phenoxazinium Derivatives	8	1	0.818 (0.602 ÷ 0.993)	0.063
48	Single-ring Substituted Primary Aromatic Amines	48	10	0.817 (0.718 ÷ 0.909)	0.0001
49	Formaldehyde Releasers	3	0	0.800 (0.473 ÷ 1.000)	0.205
50	Haloalcohols	19	4	0.800 (0.645 ÷ 0.941)	0.016
51	Heterocyclic N-Hydroxylamines	3	0	0.800 (0.473 ÷ 1.000)	0.205
52	Heterocyclic nitro compounds	3	0	0.800 (0.473 ÷ 1.000)	0.205
53	Quinone Methides	3	0	0.800 (0.473 ÷ 1.000)	0.205
54	alpha-Activated benzyls	3	0	0.800 (0.473 ÷ 1.000)	0.205
55	Carboxylic acid Anhydrides	3	0	0.800 (0.473 ÷ 1.000)	0.205
56	Gallic Acid Esters	3	0	0.800 (0.473 ÷ 1.000)	0.205
57	Carboxylic Acid Amides	10	2	0.786 (0.579 ÷ 0.968)	0.074
58	Hydrazine Derivatives	6	1	0.778 (0.524 ÷ 0.989)	0.146
59	Arenediazonium and Diazonium Salts	9	2	0.769 (0.550 ÷ 0.964)	0.107
60	Aminoacridine DNA Intercalators	2	0	0.750 (0.368 ÷ 1.000)	0.348
61	Propyne Derivatives	2	0	0.750 (0.368 ÷ 1.000)	0.348
62	Specific Imine and Thione Derivatives	8	2	0.750 (0.516 ÷ 0.959)	0.152
63	alpha, omega-Dihaloalkanes	2	0	0.750 (0.368 ÷ 1.000)	0.348
64	Bipyridilium herbicides	2	0	0.750 (0.368 ÷ 1.000)	0.348
65	Ethenyl Pyridines	2	0	0.750 (0.368 ÷ 1.000)	0.348
66	Hexahydrotriazine Derivatives	2	0	0.750	0.348

				(0.368 ÷ 1.000)	
67	N-Haloacylamides	2	0	0.750 (0.368 ÷ 1.000)	0.348
68	Propargyl Alcohol Derivatives	2	0	0.750 (0.368 ÷ 1.000)	0.348
69	Nitroaniline Derivatives	13	4	0.737 (0.543 ÷ 0.917)	0.111
70	Acyl Halides	4	1	0.714 (0.409 ÷ 0.982)	0.319
71	Sterically Hindered Piperidine Derivatives	4	1	0.714 (0.409 ÷ 0.982)	0.319
72	Nitrophenols, Nitrophenyl Ethers and Nitrobenzoic Acids	11	4	0.706 (0.495 ÷ 0.903)	0.195
73	Isocyanates and Diisocyanates	6	2	0.700 (0.432 ÷ 0.946)	0.294
74	(Thio)Acyl and (thio)carbamoyl halides, cyanides, azides, etc.	10	4	0.688 (0.467 ÷ 0.895)	0.254
75	4,4'-Bipyridinium Salts and N-Oxides	1	0	0.667 (0.224 ÷ 1.000)	0.589
76	Acridone, Thioxanthone, Xanthone, Phenazine and Other Fused-Ring Heterocyclic DNA Intercalators	3	1	0.667 (0.330 ÷ 0.974)	0.457
77	Alpha,Beta-Unsaturated Aldehydes	5	2	0.667 (0.379 ÷ 0.935)	0.397
78	Bleomycin and Structurally Related Compounds	1	0	0.667 (0.224 ÷ 1.000)	0.589
79	Chlorinated Diphenylmethane and Benzophenone Derivatives	1	0	0.667 (0.224 ÷ 1.000)	0.589
80	Diazenes	1	0	0.667 (0.224 ÷ 1.000)	0.589
81	DNA Intercalators with Carboxamide and Aminoalkylamine Side Chain	1	0	0.667 (0.224 ÷ 1.000)	0.589
82	Flavonoids	1	0	0.667 (0.224 ÷ 1.000)	0.589
83	Four- and Five-Membered Lactones	1	0	0.667 (0.224 ÷ 1.000)	0.589
84	Geminal Polyhaloalkane Derivatives	11	5	0.667 (0.454 ÷ 0.869)	0.299
85	Halofuranones	1	0	0.667	0.589

				(0.224 ÷ 1.000)	
86	Hypoxanthine Derivatives	1	0	0.667 (0.224 ÷ 1.000)	0.589
87	N-Acetoxyamines	1	0	0.667 (0.224 ÷ 1.000)	0.589
88	N-Aryl-N-Acetoxy(Benzoyloxy) Acetamides	1	0	0.667 (0.224 ÷ 1.000)	0.589
89	Nitroalkanes	1	0	0.667 (0.224 ÷ 1.000)	0.589
90	Non-Cyclic Alkyl Phosphoramides and Thionophosphoramides	1	0	0.667 (0.224 ÷ 1.000)	0.589
91	Organic Azides	1	0	0.667 (0.224 ÷ 1.000)	0.589
92	Pyrrolizidine derivatives	1	0	0.667 (0.224 ÷ 1.000)	0.589
93	Specific 5-Substituted Uracil Derivatives	1	0	0.667 (0.224 ÷ 1.000)	0.589
94	Specific Acetate Esters	1	0	0.667 (0.224 ÷ 1.000)	0.589
95	Sultones	1	0	0.667 (0.224 ÷ 1.000)	0.589
96	Thiols	1	0	0.667 (0.224 ÷ 1.000)	0.589
97	Cyanohydrines	1	0	0.667 (0.224 ÷ 1.000)	0.589
98	N-Oxycarbonyl amides, N-Acyloxy-N-alkoxyamides	1	0	0.667 (0.224 ÷ 1.000)	0.589
99	Sultones protein binding	1	0	0.667 (0.224 ÷ 1.000)	0.589
100	Amino Anthraquinones	2	1	0.600 (0.229 ÷ 0.956)	0.632
101	Hydroxamic acid	2	1	0.600 (0.228 ÷ 0.956)	0.632
102	Nitroazoarenes and p-Monosubstituted Azobenzene Derivatives	2	1	0.600 (0.228 ÷ 0.956)	0.632
103	p-Substituted Mononitrobenzenes	2	1	0.600 (0.228 ÷ 0.956)	0.632
104	Nitroarenes with Other Active Groups	3	2	0.571 (0.239 ÷ 0.895)	0.664

105	Polynitroarenes	2	2	0.500 (0.147 ÷ 0.853)	0.808
106	Azoxyalkanes	1	1	0.500 (0.094 ÷ 0.906)	0.831
107	Haloalkene Derivatives with Electron-Withdrawing Groups	3	3	0.500 (0.184 ÷ 0.816)	0.805
108	N,N-Dialkyldithiocarbamate Derivatives and Azaarene Dithiocarbamates	1	1	0.500 (0.094 ÷ 0.906)	0.831
109	Monohaloalkanes	1	3	0.333 (0.026 ÷ 0.670)	0.971
110	Organic Peroxy Compounds	0	1	0.333 (0.000 ÷ 0.776)	1
111	Quinolone Derivatives	0	1	0.333 (0.000 ÷ 0.776)	1
112	p-Aminobiphenyl Analogs	0	2	0.250 (0.000 ÷ 0.632)	1

Alerts which are not supported by chemicals from the training set (theoretical alerts) are not included in Table 1. Detailed information for each alert such as structural boundaries, mechanisms, local training sets and references associated with each observed data is provided in [Appendix 2](#).

4.5. Algorithm and descriptor generation

The structural boundaries of the alerts are derived from the chemicals included in the local training sets (see Section 4.3). For derivation of each alert mechanistically justifiable structural fragments for interaction with DNA and/or proteins are identified from the chemicals having positive data in the local training set. Additional structural fragments from the other parts of the molecules which could affect (enhance or reduce) the mutagenicity effect are also introduced to complete definition of most alerts.

4.6. Software name and version for descriptor generation

TIMES Chromosomal aberrations model version 17.17

4.7. Chemicals/Descriptors ratio

Provided in Section 4.4.

Section 5. Defining the applicability domain of the model – OECD Principle 3

5.1. Description of the applicability domain of the model

The domain consists of the following sub-domain layers:

1. General parametric requirements.

The variations of molecular parameters that may affect the quality of the measured endpoint significantly are included here (such as molecular weight, etc.). The domain of general parametric includes the range of variation of hydrophobicity ($\log K_{ow}$) and Molecular weight (MW) of chemicals in training set.

2. Structural domain.

The structural component of the model is based on the structural similarity between chemicals in the training set which were correctly predicted by the model. The structural neighborhood of atom-centered fragments (accounting for the first neighbours) extracted from correctly and incorrectly predicted parent structures from the training set is used to determine this similarity.

The target chemical could contain the following types of ACF:

- Fragments present in correctly predicted training chemicals only (i.e. correct fragments),
- Fragments found both in correctly and non-correctly predicted training chemicals (i.e. fuzzy fragments). These fragments are treated as correct fragments,
- Fragments present in non-correctly predicted training chemicals only (i.e. incorrect fragments),
- Fragments not present in the training chemicals (i.e. unknown fragments).

A chemical belongs to the structural domain of the model if it could be partitioned only on correct fragments. The user is able to analyse how important are unknown and incorrect fragments (if present in the target) and to make a decision about their effect on the quality of prediction. The distribution of structural characteristics of the target chemical and accepted thresholds is used as a criterion to determine how well the target is represented in the structural space of correctly predicted chemicals. The accepted domain thresholds for Mutagenicity are as follows:

- Correct = 100%
- Incorrect = 0%

A chemical is considered In Domain if it is classified to belong to all sub-domain levels. The information implemented in the applicability domain is extracted from the correctly

predicted training chemicals used to build the model and in this respect the applicability domain determines practically the interpolation space of the model.

5.2. Method used to assess the applicability domain

The approach used to determine and assess the domain is described in:

Dimitrov S, Dimitrova G., Pavlov T., Dimitrova N., Patlewicz G., Niemela J., Mekenyan O., A stepwise approach for defining the applicability domain of SAR and QSAR models, *J. Chem. Inf. Model.*, 45, 839-849 (2005).

5.3. Software name and version for the applicability domain assessment

The LMC software OASIS Domain Manager v.1.13 (which is embedded in OASIS platform) is used to determine the applicability domain.

<http://oasis-lmc.org/products/software/domain-manager.aspx>

5.4. Limits of applicability

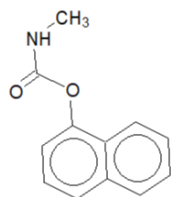
Applicability domain of the CA model (+S9) include three sub-domain layers: general parametric requirements, structural features and alerts reliability.

- General properties requirements:

As described in the Section 5.1.1, parametric domain of the model is derived based on Log *K_{ow}* and *MW*. Example demonstrating belonging of a training set chemical to the parametric layer of the model domain is provided below:

Example chemical:

- CAS: 63-25-2
- Name: Carbaryl
- 2D Depiction:



Property	Domain	Example chemical
----------	--------	------------------

<i>log K_{ow}</i>	[-13.164; 29.679]	2.348
<i>MW</i> , Da	[17.996; 1514.086]	201.211

* *K_{ow}* is calculated by EPI Suite

The values of *logK_{ow}* and *MW* of the example chemical are within the ranges of these parameters extracted from the whole training set of the model. Hence, with respect to the general parametric requirements, the example chemical is estimated to be *In Domian*.

- Structural features

Structural domain extracted from 855 training chemicals contains:

- 2588 correct fragments,
- 316 fuzzy fragments (treated as correct fragments),
- 411 incorrect fragments.

- Alerts reliability

Reliability of alerts is estimated based on:

- Alert performance of the local training set chemicals (AP);
- Number of the local training sets (N);
- Mechanistic justification (M).

According to these criteria, there are four reliability estimates for the alerts in the models:

- High reliability alerts (AP>0.6, N>10, M);
- Low reliability alerts (AP<0.6, N>10, M);
- Undetermined alerts (N<10, M);
- Undetermined theoretical alerts (M).

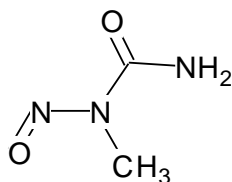
Example chemical belonging to alert with “High reliability”.

Chemical ID:

CAS: 684-93-5

Name: 1-methyl-1-nitrosourea

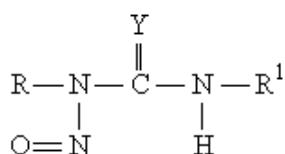
2D depiction:



Belonging to alert:

Name: Alkylated nitrosoureas and nitrosoguanidines

Structural boundaries:



where:

R = (Csp³)_n – acy at n ≥ 1, preferably linear or branched C1–C5 alkyl groups; R may also include an acyl group C(=O)Csp³ (acy) and a nitro group;

R¹ = H atom; (Csp³)_n – acy at n ≥ 1; C(=O)Csp³ (acy), Csp² (aryl), nitro group, etc.;

Y = O and NH.

Reliability:

“High reliability” based on AP=1; N=18 and M.

Currently, information for alerts reliability is provided in the model reports.

Section 6. Internal validation – OECD Principle 4

6.1. Availability of the training set

Training set of the TIMES_CA model (+S9) includes 930 organic compounds from different chemical classes.

6.2. Available information for the training set

CAS numbers, Chemical names, SMILES, documented data, literature sources and strain information are available for each compound in the model training set.

6.3. Data for each descriptor variable for the training set

Descriptors in the models are structural alerts. The main characteristics of each alert are provided in Table 1 (Section 4.4).

6.4. Data for the dependent variable for the training set

The training set of 930 chemicals include:

- 547 chemicals having positive observed CA data
- 383 chemicals having negative observed CA data.

Distribution of positive/negative chemicals in the training set of model is used for estimating performance and confidence range of the so-called *naïve alert* which is 0.588 (0.556 ÷ 0.620)¹.

1) Confidence range is calculated at 95% confidence level

6.5. Other information about the training set

The training set is compiled according to the recommendations described in the OECD TG473.

6.6. Pre-processing of data before modelling

Not available.

6.7. Statistics for goodness-of-fit

During the internal validation the original training set is separated many times randomly into two parts – one becomes a training set and the other becomes a test set. The model is re-derived many times using each new training set. Then, performance is estimated for the training sets and test sets. The averaged value of all training set performances is compared to the averaged value of all test set performances in order to assess the amount of optimism in the goodness-of-fit (GOF optimism) in the original model. GOF optimism is calculated as average performance over training sets minus average performance over test sets. Results are provided in Table 2.

Table 2. Performance of the original model over its training set (goodness-of-fit, GOF) vs. expected performance over set different from the training set (GOF – GOF optimism).

	Performance <i>est.</i> ¹⁾ model	<i>p-value</i> ¹⁾	Performance ^{2) 3)} different set
All predictions (accuracy)	0.842 (0.819 ÷ 0.865)	< 10 ⁻¹⁰	0.830
Positive chemicals (sensitivity)	0.865 (0.836 ÷ 0.893)	< 10 ⁻¹⁰	0.842
Negative chemicals (specificity)	0.808 (0.768 ÷ 0.847)	< 10 ⁻¹⁰	0.8050

¹⁾ Confidence ranges and *p-value* are calculated at 95% confidence level

²⁾ Estimated performance for training set minus GOF optimism calculated from internal validation

³⁾ Estimation of expected performance over external sets (different from training sets)

Addition information including mathematical formalism underlying the above statistical measures are provided in [Appendix 3](#).

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

Not performed

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

Method 1. *k*-fold cross-validation

In *k*-fold cross-validation the original training set is partitioned into *k* equally sized subsets. Each time a single subset is used as a test set and the remaining *k*-1 subsets are used as training set. In this manner the process is repeated *k* times and each data from the original training set is used once as a test data and *k*-1 times as a training data. The advantage of this method is that any data is used for both training and validation and each data is used exactly once as a test data. Commonly the 10-fold cross-validation is used (90% training data, 10% test data). In addition, 4-fold cross validation (75% training data, 25% test data) is also performed and the results from both procedures are provided in Table 3.

Table 3. Results from *k*-fold (10-fold and 4-fold) cross-validation.

	10-fold	4-fold

	Training sets	Test sets	Training sets	Test sets
Unique chemicals, %	90.0 (89.9 ÷ 90.1)	10.0 (9.9 ÷ 10.1)	75.0 (74.9 ÷ 75.1)	25.0 (24.9 ÷ 25.1)
Performance _{est.} , all predictions (accuracy)	0.842 (0.836 ÷ 0.849)	0.828 (0.773 ÷ 0.884)	0.842 (0.828 ÷ 0.856)	0.832 (0.790 ÷ 0.873)
<i>p-value</i> , accuracy	< 10 ⁻¹⁰	2.9 x 10 ⁻⁶	< 10 ⁻¹⁰	< 10 ⁻¹⁰
Performance _{est.} , positive chemicals (sensitivity)	0.865 (0.855 ÷ 0.875)	0.839 (0.767 ÷ 0.911)	0.865 (0.849 ÷ 0.881)	0.845 (0.800 ÷ 0.889)
<i>p-value</i> , sensitivity	< 10 ⁻¹⁰	0.0053	< 10 ⁻¹⁰	1.2 x 10 ⁻⁷
Performance _{est.} , negative chemicals (specificity)	0.808 (0.798 ÷ 0.818)	0.797 (0.712 ÷ 0.882)	0.807 (0.793 ÷ 0.821)	0.806 (0.766 ÷ 0.846)
<i>p-value</i> , specificity	< 10 ⁻¹⁰	0.0001	< 10 ⁻¹⁰	1.2 x 10 ⁻¹⁰

¹⁾ Confidence ranges and *p-value* are calculated at 95% confidence level

Method 2. Monte Carlo cross-validation

In *Monte Carlo cross-validation* the original training set is split randomly into training and test set. The advantage of this method (compared to *k-fold* cross validation) is that the proportion between training and test sets does not depend on the number of repetitions in the internal validation procedure. The *Monte Carlo cross-validation* (similarly to the *bootstrapping*) suppose creating a large number of new training/test sets (1000 – 10000). Results from application of this statistical method are provided in Table 5.

Table 4. Results from Monte Carlo cross-validation (1000 repetitions).

	75% training set		63% training set	
	Training sets	Test sets	Training sets	Test sets
Unique chemicals, %	75.0 (75.0 ÷ 75.0)	25.0 (25.0 ÷ 25.0)	63.0 (63.0 ÷ 63.0)	37.0 (37.0 ÷ 37.0)
Performance _{est.} , all predictions (accuracy)	0.842 (0.829 ÷ 0.856)	0.830 (0.789 ÷ 0.872)	0.842 (0.824 ÷ 0.861)	0.830 (0.797 ÷ 0.862)
<i>p-value</i> , accuracy	< 10 ⁻¹⁰	< 10 ⁻¹⁰	< 10 ⁻¹⁰	< 10 ⁻¹⁰
Performance _{est.} , positive chemicals (sensitivity)	0.865 (0.848 ÷ 0.881)	0.842 (0.791 ÷ 0.894)	0.864 (0.843 ÷ 0.886)	0.842 (0.802 ÷ 0.882)

<i>p-value</i> , sensitivity	< 10 ⁻¹⁰	1.8 x 10 ⁻⁵	< 10 ⁻¹⁰	9.7 x 10 ⁻⁸
Performance _{est.} , negative chemicals (specificity)	0.807 (0.785 ÷ 0.830)	0.806 (0.738 ÷ 0.873)	0.808 (0.776 ÷ 0.839)	0.808 (0.755 ÷ 0.860)
<i>p-value</i> , specificity	< 10 ⁻¹⁰	1.5 x 10 ⁻⁸	< 10 ⁻¹⁰	< 10 ⁻¹⁰

¹⁾ Confidence ranges and *p-value* are calculated at 95% confidence level

6.10. Robustness - Statistics obtained by Y-scrambling

Not performed

6.11. Robustness - Statistics obtained by bootstrap

In bootstrapping a newly derived training sets is populated from the original training set of the model by random sampling with replacement until the size of the new training set reaches the size of the original training set. The data not selected for the new training set becomes the new test set. On average, about 63% of original training set data goes into the new training set (some data appear more than once) and 37% remains in the new test set. One of the advantages of this method is that the new training sets and the original training set are equally sized. The process is repeated many times and the average results are provided in Table 5.

Table 5. Results from bootstrapping (1000 repetitions).

	Training sets	Test sets
Unique chemicals, %	63.2 (61.1 ÷ 65.3)	36.8 (34.6 ÷ 38.9)
Performance _{est.} , all predictions (accuracy)	0.842 (0.819 ÷ 0.866)	0.829 (0.796 ÷ 0.862)
<i>p-value</i> , accuracy	< 10 ⁻¹⁰	< 10 ⁻¹⁰
Performance _{est.} , positive chemicals (sensitivity)	0.865 (0.837 ÷ 0.894)	0.841 (0.798 ÷ 0.884)
<i>p-value</i> , sensitivity	< 10 ⁻¹⁰	5.5 x 10 ⁻⁷
Performance _{est.} , negative chemicals (specificity)	0.808 (0.768 ÷ 0.847)	0.808 (0.757 ÷ 0.859)

<i>p</i> -value, specificity	< 10 ⁻¹⁰	< 10 ⁻¹⁰
------------------------------	---------------------	---------------------

¹⁾ Confidence ranges and *p*-value are calculated at 95% confidence level

6.12. Robustness - Statistics obtained by other methods

Not performed

6.13. Comment on the internal validation of the model

The model shows good predictive performance for positive chemicals (Sensitivity) – 87% for training set, around 84% expected Sensitivity for chemicals different from the training set. On the contrary, prediction performance for negative chemicals (Specificity) is not that high (~81%) indicating for possible over-prediction. Lower Specificity could be due to the fact that the model training set is relatively small and not balanced – 60% of the training chemicals are observed positive which favours training of positive chemicals rather than negative ones. One should also address that *in vitro* CA tests is known to provide false positive predictions due to cytotoxicity.

Section 7. External validation – OECD Principle 4

7.1. Availability of the external validation set

93 external chemicals are available to examine performance of the model.

7.2. Available information for the external validation set

The external validation set includes pesticide chemicals from the EFSA database. Details of the EFSA database are available in the website of the agency:

<https://data.europa.eu/euodp/en/data/dataset/database-pesticide-genotoxicity-endpoints>

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

Not available.

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

Not available.

7.5. Other information about the external validation set

The EFSA genotoxicity database is biased towards negative chemicals. This justifies the fact that Sensitivity of the model is examined based on a small number of chemicals as compared with the substances used for estimating Specificity of the model. Details about the external validation set are available in:

P.I. Petkov, T. W. Schultz, M. Honma, T. Yamada, E. Kaloyanova, O. Mekenyan. 2019. Validation of the performance of TIMES genotoxicity models with EFSA pesticide data. *Mutagenesis*, Vol. 34, pp. 83-90.

7.6. Experimental design of test set

The external validation set of 93 pesticides contains:

- Positive data (10 chemicals)
- Negative data (83 chemicals)

7.7. Predictivity – Statistics obtained by external validation

Performance of the chemicals which belong and does not belong to the model domain (*In domain* and *Out of domain*) is:

- Sensitivity = 70% (7 of 10 chemicals)
- Specificity = 43% (36 of 83 chemicals)

Lower Specificity indicated that the *in vitro* liver S9 metabolic simulator associated with the model has to be modified with respect to the pesticide chemicals. These and other modifications of the metabolic simulator are performed in the latest version of the model.

7.8. Predictivity – Assessment of the external validation set

The EFSA genotoxicity database for pesticide chemicals is assumed to be a high quality database. According to the description of the database, all endpoints are evaluated using standard tests conditions as described in the corresponding tests guidelines.

7.9. Comment on the external validation of the model

Performance of the CA model (+S9) in terms of Sensitivity is reasonable accounting for the fact that all 10 chemicals with positive CA data do not belong to the model domain. Lower Specificity indicates that modifications of some alerts and/or *in vitro* S9 metabolic

simulator associated with the model are needed. Such modifications are performed in the latest versions of the model.

Section 8. Providing a mechanistic interpretation – OECD Principle 5

8.1. Mechanistic basis of the model

Only alerts extracted from the local training sets having clear interpretation of the molecular mechanism causing the mutagenicity effect are included in the model. Mechanistic rationale of each alert is provided by experts based on significant reference support from the literature.

8.2. *A priori* or *a posteriori* mechanistic interpretation

The model building followed the traditional approach:

- a. Building a hypothesis for the modelled event,
- b. Defining the alerting groups based on parent structures,
- c. Fitting of model variable to the observed data,
- d. Verification of model quality,
- e. Depending on the results found in step *d* model building could continue with step *a*, *b* or *f*,
- f. Determination of the applicability domain and practical application of the model.

8.3. Other information about the mechanistic interpretation

Additional information about the mechanistic interpretation could be found in Section 2 (2.7).

Section 9. Miscellaneous information

9.1. Comments

Model predictions are fully transparent. The user is able to analyse the whole prediction process and to verify whether it concise with his/her knowledge or purposes.

For other related models, see Section 1 (1.2).

9.2. Bibliography

Additional references are not provided.

9.3. Supporting information

Additional supporting information is not provided.