	QMRF identifier (JRC Inventory): **Q13-410-0064**
	QMRF Title: **TOPS-MODE QSAR for mammalian cell mutagenicity of alpha,beta-unsaturated carbonyl compounds**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

TOPS-MODE QSAR for mammalian cell mutagenicity of alpha,beta-unsaturated carbonyl compounds

1.2.Other related models:

1.3.Software coding the model:

2.General information

2.1.Date of QMRF:

December 2009

2.2.QMRF author(s) and contact details:

[1]Pérez-Garrido A Environmental Engineering and Toxicology Dpt., Catholic University of San Antonio San Antonio, Guadalupe, Murcia,Spain aperez@pdi.ucam.edu

[2]Morales Helguera A Department of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas Santa Clara, Villa Clara, Cuba aliuskamhelguera@yahoo.es

[3]Cordeiro MNDS REQUIMTE, Chemistry Department, Faculty of Sciences, University of Porto Porto, Portugal

[4]López GC Department of Food and Nutrition Technology, Catholic University of San Antonio Guadalupe, Murcia, Spain

[5]Garrido Escudero A Environmental Engineering and Toxicology Dpt., Catholic University of San Antonio Guadalupe, Murcia,Spain

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Pérez-Garrido A Environmental Engineering and Toxicology Dpt., Catholic University of San Antonio San Antonio, Guadalupe, Murcia,Spain aperez@pdi.ucam.edu

2.6.Date of model development and/or publication:

2009

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

Pérez-Garrido A, Helguera AM, López GC, Cordeiro MN & Escudero AG (2010). A topological substructural molecular design approach for predicting mutagenesis end-points of alpha, beta-unsaturated carbonyl compounds. Toxicology 268, 64-77. http://www.ncbi.nlm.nih.gov/pubmed/20004227

2.8.Availability of information about the model:

Training and test sets are available. Algorithm available.

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

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Various cell lines

3.2.Endpoint:

4.Human Health Effects 4.10.Mutagenicity

3.3.Comment on endpoint:

Mutagenicity measured using cell lines or strains with or without exogenous metabolic activation (S9): L5178Y mouse lymphoma cells, CHO, AS52 and V79 lines of Chinese hamster cells. A compound was categorized as a mutagen if at least one the mammalian test result was positive while a compound was categorized as nonmutagen if exclusively negative mammalian test results one or more were reported. Mammalian cell mutagenicity according to OECD guidelines.

3.4.Endpoint units:

no units

3.5.Dependent variable:

MCGM =1 positive result; MCGM=-1 negative result.

3.6.Experimental protocol:

The data were obtained according to the OECD 476 Test Guideline.

3.7.Endpoint data quality and variability:

Mammalian cell gene mutation test using cell lines L5178Y mouse lymphoma cells, CHO, AS52 and V79 lines of Chinese hamster cells. Data from a large compilation, from many different sources.

The data were extracted from the Chemical Carcinogenesis Research Information System (http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS). The test results was extracted of Chemical Carcinogenesis Research Information System (CCRIS).CCRIS is a toxicology data file of the National Library of Medicine's (NLM) Toxicology Data Network (TOXNET?). It is a scientifically evaluated and fully referenced data bank, developed and maintained by the National Cancer Institute (NCI). Data are derived from studies cited in primary journals, current awareness tools, NCI reports, and other special sources. Test results have been reviewed by experts in carcinogenesis and mutagenesis. We categorized a compound as a mutagen if at least one the mammalian test result was positive while a compound was categorized as nonmutagen if exclusively negative mammalian test results one or more were reported.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR

4.2.Explicit algorithm:

QSAR

QSAR derived by two-group Linear Discriminant Analysis

MCGM = 4.143

? 2.548?$_2{}^{\text{Dip}}$

+ 3.012?$_2{}^{\text{Dip2}}$

?1.54 x (10?4?7$^{\text{Pol}}$)

+ 5.271?$^{1\text{Gas}}$

4.3.Descriptors in the model:

[1]2Dip TOPS-MODE descriptor

[2]2Dip2 TOPS-MODE descriptor

[3]7Pol TOPS-MODE descriptor

[4]1Gas TOPS-MODE descriptor

4.4.Descriptor selection:

variable selection. This was used to select the variables (descriptors) with the highest influence on mutagenicity but in contrast to regression analysis, which minimizes the standard deviation, we minimized the Wilk's Lambda.

4.5.Algorithm and descriptor generation:

The spectral moments of the edge adjacency matrix are defined as the traces. That is the sum of the main diagonal of the different powers of such matrix. Several bond weights such as standard bond distance (Std), standard bond dipole moments (Dip, Dip2), hydrophobicity (H), polar surface area (Pols), polarizability (Pol), molar refractivity (Mol), van der Waals radii (vdW), and Gasteiger–Marsilli charges (Gas) were used for computing the spectral moments of the bond matrix.

4.6.Software name and version for descriptor generation:

Modeslab

http://www.modeslab.com/

4.7.Chemicals/Descriptors ratio:

9.75 (39 chemicals / 4 descriptors)

Whereas the ratio Chemicals / Descriptors in final equation (39/4=9.75) is given, the ratio relative to the “originally screened” parameters is not given. As a matter of fact, this is the critical parameter that –if not respected- may produce chance correlations (ref 15 sect 9.2)

The pool of original descriptors was 676.

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

In the Williams plot, i.e. the plot of standardized residuals versus leverage values (h), the applicability domain is established inside a squared area within x standard deviations and a leverage threshold h*=0.307 (h* is generally fixed at 3p/n, where n is the number of training compounds and p the number of model parameters, whereas x = 3).

See Pérez-Garrido et al. (2009).

5.2.Method used to assess the applicability domain:

Method based on leverage values (Gramatica, 2007).

5.3.Software name and version for applicability domain assessment:

Statistica v 7.0

http://www.statsoft.com

5.4.Limits of applicability:

Substances that had a leverage value igreater than the threshold (h*=0.307) are outside of the applicability domain. Leverage threshold

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

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:

39 compounds: 27 positives; 12 negatives

6.6.Pre-processing of data before modelling:

Data taken as they are from a compilation.

6.7.Statistics for goodness-of-fit:

The goodness-of-fit was evaluated by checking the:
accuracy: the percentage of all chemicals correctly identified by the model;
sensitivity: the percentage of mutagenic (positive) chemicals correctly identified (calculated out of the total number of positives);
specificity: the percentage of non-mutagenic (negative) chemicals correctly identified (calculated out of the total number of negatives);
Squared Mahalanobis Distances (D2), the Wilk’s lambda (?), Fisher function, FIT(?) and Kappa (?)
The parametrer FIT(?) is similar to Kubinyi function in regression analysis, defined by:
FIT(?)=(1-?)(n-k-1)/(n+k2)?, where n is the number of compounds in the training set, k is the number of variables in the equation that describe the model, and ? is the Wilk´s
Lambda. The FIT(?) criterion has a low sensitivity toward changes in k values, as long as they are small numbers, and a substantially increasing sensitivity for large k values.
The ? index (Cohen, 1960) excludes matching due solely to chance. However, a commonly cited scale is represented in by Landis and Koch (1977):
?<0 Less than chance agreement;
? between 0.01 and 0.20 Slight agreement;
? between 0.21 and 0.40 Fair agreement;
? between 0.41 and 0.60 Moderate agreement;
? between 0.61 and 0.80 Substancial agreement;? between 0.81 and 0.99 Almost perfect agreement;?=0.412; p<10-5; F=12.107 (Fisher function); FIT(?)=0.881 ; ?=0.727

(Kappa), D2=6.343;

Sensitivity: 92.59%; Specificity: 75.17%; Accuracy: 87.67%; False positives=7.41%; False negatives=24.83%

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

We have implemented the leave-group-out (LGO) procedure

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

The LGO procedure was applied, leaving out 20% of the training set by random extraction and then recalculating the model and the statistics with the remaining chemicals. This LGO procedure was repeated 300 times. The mean values of the accuracy, sensitivity, and specificity for both training and test sets, as well as the mean values of Wilk's ? (?Cross) and squared Mahalanobis distances (D2Cross) , are reported.

?Cross=0.399; D2 Cross=6.756; Sensitivity Training: 93.35%; Specificity Training: 75.17%; Accuracy Training: 87.67%; Sensitivity Test: 89.87%; Specificity Test: 69.67%; Accuracy Test: 84.10%

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: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

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

All

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

All

7.5.Other information about the external validation set:

9 datapoints: 7 positive values; 2 negative values

7.6.Experimental design of test set:

k-Means Cluster Analysis (k-MCA) was used to extract the test set. The training set contained 80% (39/48) of the original data whereas the test set the remaining 20%. The k-MCA analysis was separately made for each group: mutagenic and non-mutagenic. Selection of the training and test sets was then carried out by taking compounds belonging to each cluster, proportionally to the size of the cluster. The pool of descriptors was formed for the entire Dragon descriptors family. We also made an inspection of the standard deviation between and within clusters, the respective Fisher ratio and p level of significance (ought to be lower than 0.05) (McFarland and Gans, 1995, Johnson and Wichern, 1988). Table

1.

Table 1. Standard deviation between and within clusters, degrees of freedom (df), Fisher ratio (F) and level of significance (p) of the variables in the k-means cluster analysis.

	Variable	Between SS	df	Within SS	df	F	signif. p
	BEHv5	31.98	4	0.76	29.00	304.66	<10-5
	BELv5	31.66	4	1.39	29.00	165.67	<10-5
Mutagenic	BEHe5	32.43	4	1.18	29.00	198.72	<10-5
	BEHp5	31.75	4	0.77	29.00	297.44	<10-5
	H1v	26.40	4	3.91	29.00	48.98	<10-5
Non-mutagenic	Mor19u	10.34	1	4.04	12.00	30.74	1.27 10-4
	Mor19e	10.37	1	3.83	12.00	32.54	9.85 10-5

7.7.Predictivity - Statistics obtained by external validation:

Sensitivity: 85.71%; Specificity: 100%; Accuracy: 88.89%; False positives=0%; False negatives=14.29%

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:

Electro-donating substituents such as methyl groups in position ? reduce the reactivity of this moiety by Michael type mechanism and electron-withdrawing substituents in the double bound increase the mutagenicity of the substance (Aptula and Roberts, 2006; Schultz et al, 2005). These features are consistent with an Michael addition type mechanism since the stabilization of the positive charge on the terminal carbon, the preferred site of nucleophilic attack (Feron, 1991, Dearfield, 1991) are determinant in its reactivity.

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori interpretation based on variables of the equation.

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

TOPological Substructural MOlecular DEsign (TOPS-MODE) descriptors are based on the calculation of the spectral moments of the so-called bond matrix (Estrada, 1996 and 1997).

These are graph-based descriptors: the molecular structure of each compound is represented by its molecular graph and then, the bond adjacency matrix (B) is derived. B is a squared symmetric matrix whose entries are ones or zeros if the corresponding bonds are adjacent or not. The order of this matrix (m) is the number of bonds in the molecular graph, being two bonds adjacent if they are incident to a common atom. Furthermore, weights are introduced in the diagonal entries of this matrix to mirror fundamental physicochemical properties that might relate to the target endpoint being modelled. The weights included the standard bond distance (Std), standard bond dipole moments (Dip, Dip2), as well as contributions from the following atomic properties: hydrophobicity (Hyd), polar surface area (Pols), polarizability (Pol), molar refractivity (Mol), van der Waals radii (vdW), Gasteiger–Marsilli charges (Gas), atomic masses (Ato), solute excess molar refraction (Ab-R$_2$), solute dipolarity/polarizability (Ab-?$_2^H$), effective hydrogen-bond basicity (Ab-??$_2^0$, Ab-??$_2^H$) and solute gas hexadecane partition coefficient (Ab-logL^{16}). Finally, the spectral moments are defined as the traces (i.e., the sum of the main diagonal elements) of the different powers of the weighted B matrix.

9.2.Bibliography:

[1]Duchowicz PR, Castro EA, Fernndez FM (2006). Alternative algorithm for the search of an optimal set of descriptors in qsar-qspr studies. MATCH Communications in Mathematical and in Computer Chemistry 55, 179–192.

[2]Gramatica P (2007). Principles of QSAR models validation: internal and external. QSAR & Combinatorial Science 26, 694-701.

[3]ement for nominal scales. Journal of Educational and Psychological Measurement 20, 37–46.

[4]Landis JR & Koch GG (1977). The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

[5]McFarland JW & Gans DJ (1995). Chemometric methods in molecular design. pp. 295-307. VCH, Weinheim.

[6]Johnson RA & Wichern DW (1988). Applied MultiVariate Statistical Analysis. Prentice- Hall, New York.

[7]Feron VJ, Til HP, de Vrijer F, Woutersen RA, Cassee FR & van Bladeren PJ (1991). Aldehydes: Occurrence, carcinogenic potential, mechanism of action and risk assessment. Mutation Research 259, 363–385.

[8]Dearfield KL, Harrington-Brock K, Doerr CL, Rabinowitz JR & Moore MM (1991). Genotoxicity in mouse lymphoma cells of chemicals capable of michael addition. Mutagenesis 6, 519–525.

[9]Chemical Carcinogenesis Research Information System http://toxnet.nlm.nih.gov/cgibin/ sis/htmlgen?CCRIS

[10]Estrada E (1996). Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 1. Definition and Applications to the Prediction of Physical Properties of Alkanes. Journal of Chemical Information and Computer Sciences 36, 844–849.

[11]Estrada E (1997). Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 2. Molecules Containing Heteroatoms and QSAR Applications. Journal of Chemical Information and Computer Sciences 37, 320–328.

[12]Schultz TW, Yarbrough JW & Johnson EL (2005). Structure-activity relationships for reactivity of carbonyl-containing compounds with glutathione. SAR and QSAR in Environmental Research 16, 313–322.

[13]Aptula AO & Roberts DW (2006). Mechanistic applicability domains for non-animal based toxicological end points: General principles and application to reactive toxicity. Chemical Research in Toxicology 19, 1097–1105.

[14]Pérez-Garrido A, Helguera AM, López GC, Cordeiro MN & Escudero AG (2010). A topological substructural molecular design approach for predicting mutagenesis end-points of alpha, beta-

unsaturated carbonyl compounds. Toxicology 268, 64-77.

http://www.ncbi.nlm.nih.gov/pubmed/20004227

[15]Topliss & Costello (1972). Journal of Medicinal Chemistry 15, 1066.

9.3.Supporting information:

TOPS MODE Training_39.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q13-410-0064/attachment/A760
TOPS MODE Test_9.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q13-410-0064/attachment/A761

Test set(s)

Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q13-410-0064

10.2.Publication date:

2013-07-02

10.3.Keywords:

TOPS-MODE;mutagenicity;alpha;beta-unsaturated carbonyl compound;

10.4.Comments:

former Q19-30-8-242