	QMRF identifier (JRC Inventory):*To be entered by JRC*
	QMRF Title: *Insubria QSAR PaDEL-Descriptor model for prediction of Endocrine Disruptors Chemicals (EDC) Estrogen Receptor (ER)-binding affinity.*
	Printing Date:*Feb 7, 2014*

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for prediction of Endocrine Disruptors Chemicals (EDC) Estrogen Receptor (ER)-binding affinity.

1.2.Other related models:

J.Li and P.Gramatica. The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders, Mol. Divers. 14, 2010, pp 687-696. [8]

1.3.Software coding the model:

[1]PaDEL-Descriptor 2.18 A software to calculate molecular descriptors and fingerprints http://padel.nus.edu.sg/software/padeldescriptor/index.html
[2]QSARINS 1.2 Software for the development, analysis and validation of QSAR MLR models paola.gramatica@uninsubria.it www.qsar.it

2.General information

2.1.Date of QMRF:

6/12/2013

2.2.QMRF author(s) and contact details:

Stefano Cassani DiSTA, University of Insubria (Varese - Italy) +390332421439 stefano.cassani@uninsubria.it www.qsar.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Stefano Cassani DiSTA, University of Insubria (Varese - Italy) +390332421439 stefano.cassani@uninsubria.it www.qsar.it
[2]Paola Gramatica DiSTA, University of Insubria (Varese - Italy) paola.gramatica@uninsubria.it www.qsar.it

2.6.Date of model development and/or publication:

July 2013

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

[1]Gramatica P., et al. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J. Comput. Chem. (Software News and Updates), 2013, 34 (24), 2121-2132 [1]
[2]QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS, submitted to J. Comput. Chem. (Software News and Updates)

2.8.Availability of information about the model:

The model is non-proprietary and published in a scientific peerreviewed journal. All information in full details are available

(e.g.training and prediction set, algorithm, ecc...).

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

No

3.Defining the endpoint - OECD Principle 1

3.1.Species:

No information available

3.2.Endpoint:

4.Human health effects 18.Endocrine Activity 4.18.a.Receptor-binding (specify receptor)

3.3.Comment on endpoint:

The training set (129 compounds) used in this study was extracted from the NCTR, EDKB [2]. The ER, binding affinities, expressed as log unit of relative binding affinity (logRBA), were tested in the rat uterine cytosol ER competitive binding assay. The external evaluation set (23 compounds) is chosen from Kuiper' article [3]. The activities (logRBA) of the 23 active compounds used in this evaluation set were normalized by Shi et al. [4].

3.4.Endpoint units:

log unit of relative binding affinity

3.5.Dependent variable:

logRBA

3.6.Experimental protocol:

Rat uterine cytosol ER competitive binding assay

3.7.Endpoint data quality and variability:

No information available

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR - Multiple linear regression model (OLS - Ordinary Least Square)

4.2.Explicit algorithm:

LogRBA Full model

The full model was then applied to 23 external chemicals

Full model equation: logRBA= -17.16 + 2.93 VadjMat + 0.15 MDEC-23 – 1.19 n6Ring + 0.55 nFG12Ring -0.65 nHBAcc + 2.98 maxHsOH + 0.15 MDEC-13 – 1.21 hmax

4.3.Descriptors in the model:

[1]VadjMat Vertex adjacency information (magnitude)

[2]MDEC-23 Molecular distance edge between all secondary and tertiary carbons

[3]n6Ring Number of 6-membered rings

[4]nHBAcc Number of hydrogen bond acceptors (using CDK HBondAcceptorCountDescriptor algorithm)

[5]nFG12Ring Number of >12-membered fused rings

[6]maxHsOH Maximum atom-type H E-State: -OH

[7]MDEC-13 Molecular distance edge between all primary and tertiary carbons
[8]hmax Maximum H E-State

4.4.Descriptor selection:

A total of 1598 molecular descriptors of differing types (0D, 1D, 2D, Fingerprints) were calculated in PaDEL-Descriptor 2.18. Constant and semi-constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (one of any two descriptors with a correlation greater than 0.98 was removed to reduce redundant information), and a final set of 280 molecular descriptors were used as input variables for variable subset selection. The models were initially developed by the all-subset-procedure, and then GA was applied to obtain the final population of models (eight variables). The optimized parameter used was Q2LOO (leave-one-out).

4.5.Algorithm and descriptor generation:

Multiple linear regression (Ordinary Least Square method) was applied to generate the model.

Molecular descriptors were generated by PaDEL-Descriptor software. The input files for descriptor calculation contain information on atom and bond types, connectivity, partial charges and atomic spatial coordinates, relative to the minimum energy conformation of the molecule, and were firstly obtained by the semi empirical AM1 method using the package HYPERCHEM. Then, these files were converted by OpenBabel into MDL-MOL format and used as input for the calculation of descriptors in PaDEL-Descriptor.

4.6.Software name and version for descriptor generation:

PaDEL-Descriptor 2.18
A software to calculate molecular descriptors and fingerprints
Yap Chun Wei, Department of Pharmacy, National University of Singapore.
http://padel.nus.edu.sg/software/padeldescriptor/index.html

HYPERCHEM - ver. 7.03
Software for molecular drawing and conformational energy optimization

OpenBabel ver.2.3.2
Open Babel: The Open Source Chemistry Toolbox. Used for conversion between HYPERCHEM files (hin) and MDL-MOL files.
http://openbabel.org

4.7.Chemicals/Descriptors ratio:

Full model: 129 chemicals / 8 descriptors = 16.13

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of the model was verified by the leverage approach and fixed thresholds has been used to define both structural and response outliers (see section 5.4). The plot of leverages (hat diagonals) versus standardised residuals, i.e. the Williams plot, verified the presence of response outliers (i.e.compounds with cross-validated standardized residuals greater than 2.5 standard deviation units) and chemicals very structurally influential in determining model parameters parameters (i.e. compounds with a leverage value (h) greater than 3p'/n (h*), where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model). For new compounds without experimental data, leverage can be used as a quantitative measure for evaluating the degree of extrapolation: for compounds with a high leverage value (h > h*), that are structural outliers, predictions should be considered less reliable.

Response and descriptor space:

Range of experimental logRBA values (for training set of 129 chemicals): -4.5 / 2.6

Range of descriptor values (for training set of 129 chemicals): VadjMat: 4 / 6.39; MDEC-23: 0 / 36.1; n6Ring: 0 / 4; nHBAcc: 0 /3; nFG12Ring: =0 / 3; maxHsOH: 0 / 0.69; MDEC-13: 0 / 8.46; hmax: 0.18 / 1.53.

5.2.Method used to assess the applicability domain:

As it has been stated in section 5.1, the structural applicability domain of the model was assessed by the leverage approach, providing a cut-off hat value (h*=0.209). HAT values are calculated as the diagonal elements of the HAT matrix:

H = X(XTX)-1XT

The response applicability domain can be verified by the standardized residuals, calculated as: r'i = ri / s√(1-hii), where ri = Yi-Ŷi.

5.3.Software name and version for applicability domain assessment:

QSARINS 1.2

Software for the development, analysis and validation of QSAR MLR models

paola.gramatica@uninsubria.it

www.qsar.it

5.4.Limits of applicability:

FULL model applied to 23 external chemicals domain: outliers for structure, hat>0.209 (h*): fulvestrant (129453-61-8), Chlordecone (143-50-0), Phenolsulfonphthalein (143-74-8), 1,3-Diphenyltetramethyldisiloxane (56-33-7), Hexestrol (5635-50-7). Outliers for response, standardised residuals > 2.5 standard deviation units: Phenolphthalin (81-90-3), 3,6,4'-Trihydroxyflavone (no CAS)

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:Yes
INChI:No
MOL file:Yes
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:
The predictive capability of the proposed equation was verified applying the full model (n training= 129) to an external evaluation set, composed of 23 chemicals. No splitting methodology was applied a priori.
6.6.Pre-processing of data before modelling:
No information available
6.7.Statistics for goodness-of-fit:
R^2= 0.76; CCCtr [5]=0.86; RMSE= 0.87
6.8.Robustness - Statistics obtained by leave-one-out cross-validation:
Q^2LOO= 0.72; CCCcv=0.84; RMSEcv= 0.94
6.9.Robustness - Statistics obtained by leave-many-out cross-validation:
Q^2LMO= 0.74
6.10.Robustness - Statistics obtained by Y-scrambling:
R^2y-sc= 0.06
6.11.Robustness - Statistics obtained by bootstrap:
No information available (since we have calculated Q^2LMO)
6.12.Robustness - Statistics obtained by other methods:
No information 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:Yes
INChI:No
MOL file:Yes
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:

The predictive capability of the proposed equation was verified applying the full model (n training= 129) to an external evaluation set, composed of 23 chemicals. The range of logRBA for these compounds is: -2.51 / 1.41.

7.6.Experimental design of test set:

No other information available

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [4]= 0.74; Q^2extF2 [6]= 0.69; Q^2extF3 [7]= 0.89; CCCex=0.82; RMSE= 0.59

7.8.Predictivity - Assessment of the external validation set:

Training and evaluation sets are balanced according to both structure and response. In particular, for response the range of logRBA values are [-4.5 / 2.6][-2.51 / 1.41] respectively for training and evaluation set.

As much as concern structural representativity, the range of descriptors values is:

VadjMat: training set (4 / 6.39), evaluation set (5 / 6.09);
MDEC-23: training set (0 / 36.1), evaluation set (7.76 / 33.69);
n6Ring: training set (0/ 4), evaluation set (1 / 4);
nHBAcc: training set (0/ 3), evaluation set (0 / 3);
nFG12Ring: training set (0/ 3), evaluation set (0 / 3);
maxHsOH: training set (0/ 0.69), evaluation set (0.34 / 0.60);
MDEC-13: training set (0/ 8.46), evaluation set (0 / 3.50);
hmax: training set (0.18 / 1.53), evaluation set (0.38 / 0.71).

7.9.Comments on the external validation of the model:

no other information available

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic basis for this physico-chemical property was set a priori, but a mechanistic interpretation of molecular descriptors was provided a posteriori (see 8.2).

8.2.A priori or a posteriori mechanistic interpretation:

The DRAGON model published in Li and Gramatica [8] was: LogRBA= −3.76 − 0.03 TIE + 0.04 TIC1 + 2.35 ATS4m − 2.08 EEig02d + 61.69 JGI10 + 3.08 E1s −12.60 Dv − 1.25 nArOR

where:

TIE: E-state topological parameter
TIC1: total information content index (neighborhood symmetry of 1-order)
ATS4m: Broto-Moreau autocorrelation of a topological structure - lag 4 / weighted by atomic masses

EEig02d: Eingenvalue 02 from edge adj. matrix weighted by dipole moments

JGI10: mean topological charge index of oreder 10

E1s: 1st component acessibility directional WHIM index / wheighted by atomic electrotopological state

Dv: D total accesibility index / weighted by atomic vand der Waals volumes

nArOR: number of ethers (aromatic)

The equation of the new PaDEL-descriptor model included in QSARINS is := -17.16 + 2.93 VadjMat + 0.15 MDEC-23 – 1.19 n6Ring + 0.55 nFG12Ring -0.65 nHBAcc + 2.98 maxHsOH + 0.15 MDEC-13 – 1.21 hmax where

VadjMat= Vertex adjacency information (magnitude) MDEC-23= Molecular distance edge between all secondary and tertiary carbons

n6Ring= Number of 6-membered rings

nHBAcc= Number of hydrogen bond acceptors (using CDK HBondAcceptorCountDescriptor algorithm) nFG12Ring= Number of >12-membered fused rings

maxHsOH= Maximum atom-type H E-State: -OH MDEC-13= Molecular distance edge between all primary and tertiary carbons hmax= Maximum H E-State

The DRAGON published model for ER Binding contains two 3D descriptors (E1s and Dv), while the new PaDEL-Descriptor model is simpler being based only on 2D variables.

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

No other information available

9.2.Bibliography:

[1]Gramatica P., et al. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J. Comput. Chem. (Software News and Updates), 2013, 34 (24), 2121-2132.

[2]National Center for Toxicological Research (NCTR) endocrine disruptor knowledge base (EDKB) http://edkb.fda.gov/databasedoor.html

[3]Kuiper GGJM, Lemmen JG, Carlsson B, Corton JC, Safe SH, van der Saag PT, van der Burg B, Gustafsson JA (1998) Interaction of estrogenic chemicals and phytoestrogens with estrogen receptor beta. Endocrinology 139:4252–4263

[4]Shi L.M. et al. QSAR Models Using a Large Diverse Set of Estrogens, J. Chem. Inf. Comput. Sci. 41 (2001) 186–195.

[5]Chirico N. and Gramatica P., Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and

the Need for Scatter Plot Inspection, J. Chem. Inf. Model. 2012, 52, pp 2044– 2058
[6]Schuurman G. et al. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean, J. Chem. Inf. Model. 48 (2008) 2140-2145.
[7]Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669-1678
[8]J.Li and P.Gramatica. The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders, Mol. Divers. 14, 2010, pp 687-696.

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