	QMRF identifier (JRC Inventory): **Q13-301-0040**
	QMRF Title: **Catalogic Hybrid Expert System for biodegradation**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

Catalogic Hybrid Expert System for biodegradation

1.2.Other related models:

1.3.Software coding the model:

Catalogic

Version 5.10.7

Ovanes Mekenyan

http://oasis-lmc.org/?section=software&swid=1

2.General information

2.1.Date of QMRF:

24 June 2009

2.2.QMRF author(s) and contact details:

Grace Patlewicz DuPont Haskell Global Centers for Health and Environmental Sciences, DuPont, Newark DE19701 grace.y.tier@usa.dupont.com

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Ovanes G Mekenyan Laboratory of Mathematical Chemistry, University "Prof. As. Zlatarov", Yakimov St, 8010 Bourgas, Bulgaria Ovanes G Mekenyan omekenya@btu.bg http://oasis-lmc.org

2.6.Date of model development and/or publication:

The first publication cites 2002 as being the year when the Catabol software (now termed Catalogic) was first developed. This QMRF describes Catalogic Version 5.10.7.

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

[1]Jaworska J, Dimitrov S, Nikolova N & Mekenyan O (2002). Probablistic assessment of biodegradability based on metabolic pathways. CATABOL system. SAR & QSAR in Environmental Research 13, 445-455.

[2]Information on Catalogic Version 5.10.7 and installation is available on the LMC website. http://oasis-lmc.org/?section=software&swid=1

2.8.Availability of information about the model:

The model is commercal and includes a mixture of both proprietary data and published data. The published data comprises the MITI database for degradation studies and the metabolic pathways are taken from the University of Minnesota Biocatalysis/Biodegradation database (http://umbbd.msi.umn.edu/). The proprietary data comprises biodegradation studies from Procter and Gamble.

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

None to date.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Not applicable. A solution, or suspension, of the test substance in a mineral medium is inoculated and incubated with activated sludge under aerobic conditions in the dark or in diffuse light.

3.2.Endpoint:

2.3.a.Persistence: Biodegradation. Ready/not ready biodegradability 301Ready Biodegradability

3.3.Comment on endpoint:

A BOD value of between 0-100% is the endpoint result used. Catalogic converts this to a value between 0 and 1

3.4.Endpoint units:

A percentage of the BOD (biological oxygen demand) is determined.

3.5.Dependent variable:

BOD (biological oxygen demand)

3.6.Experimental protocol:

OECD 301C

3.7.Endpoint data quality and variability:

(Tokyo, Japan) and can be downloaded from http://www.cerij.or.jp/ceri_en/otoiawase/otoiawase_menu.html. These databases have been used extensively for QSAR model development and validation. These data sets are generally available and are regarded as of a high quality.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Hybrid Expert System

4.2.Explicit algorithm:

Hybrid Expert System

The expert system comprises a mix of structure-activity, QSAR and structure-microbial metabolism rules

4.3.Descriptors in the model:

Epiwin Epiwin's Kowwin and Water solubility are embedded into Catalogic to provide an automated means of calculating LogKowin, MW and solubility values as part of the applicability domain assessment.

4.4.Descriptor selection:

4.5.Algorithm and descriptor generation:

4.6.Software name and version for descriptor generation:

4.7.Chemicals/Descriptors ratio:

Not applicable. This is a hybrid expert system that relies on a mix of structure-activity and structure-metabolism rules.

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Catalogic utilises a multi-stage applicability domain that has been described by Dimitrov et al. (2005). The first stage comprises a global requirements domain with simple cut-offs for physicochemical parameters

such as logKow, molecular weight (MW) and water solubility. These upper and lower thresholds for logKow, MW and solubility are based on the training set information. The next stage consists of defining the structural domain which is used to gauge the extent to which a chemical is structurally similar to those in the training set of chemicals. The structural domain can be characterised by the atom centred fragments of these chemicals. A series of rules was proposed to reflect the effect of different neighbours on a specified atom. Hydrogen atoms are treated as an inherent part of the atoms to which they were bound, whereas the first, second, and so on neighbours were used to characterise the specified atom. In practice this means that specific substructures are extracted from a set of compounds to characterise the structural domain for a given atom. The conservativeness of structural domain is defined by the atomic neighborhood taken into account for atom-centred fragments. This affects the chemicals which are encompassed in the domain and the correctness of the predictions. The metabolic simulator domain takes into account the reliability of the generated metabolites. The reliability of metabolites is determined as a product of reliabilities of transformations used to generate these metabolites. An overall call for the domain status is reported as the total domain. If one of the domains fails for a given chemical, the overall outcome is reported as ‘outside of domain’. Catalogic provides an applicability domain definition for each of its component models which are applied in a sequential manner. The advantage of processing query chemicals through all of the stages is the increased reliability of prediction for those chemicals that satisfy all conditions for inclusion in the AD. The cost of applying this rigorous approach is that the number of chemicals for which reliable predictions are eventually made is reduced but this increases confidence in reliability of the final prediction.

5.2.Method used to assess the applicability domain:

The approach use to determine and assess the domain is described in Dimitrov et al (2005). Default settings within the program are used to assess the domain

5.3.Software name and version for applicability domain assessment:

Catalogic

Version 5.10.7

Ovanes G Mekenyan

5.4.Limits of applicability:

These are integrated within the Catalogic program. The thresholds of the domain settings are conservative. The global parameters of LogKowin, MW and solubility are set based on the training set and are respectively: LogKow -4-24; MW 44-959 and solubility 0-1000000 mg/L. These values are derived from the embedded Epiwin software that are used to calculate these physicochemical properties.

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: 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:

Some

6.5.Other information about the training set:

The MITI training underpinning this model is available directly from the Chemicals Evaluation Research Institute (Tokyo, Japan) and can be downloaded from http://www.cerij.or.jp/ceri_en/otoiawase/otoiawase_menu.html. The Procter & Gamble dataset (109 proprietary chemicals) is not available. The MITI training set included in the model is viewable within the Catalogic software.

6.6.Pre-processing of data before modelling:

The percentage BOD values were transformed into ratios from 0-1.

6.7.Statistics for goodness-of-fit:

According to the original Catabol publication (Jaworska et al, 2002): The model allows for identifying potentially persistent catabolic intermediates and their molar amounts. The data in the training set agreed well with the calculated BODs (r2 = 0.90) in the entire range i.e. a good fit was observed for readily, intermediate and difficult to degrade chemicals. After introducing 60% ThOD as a cut off value, the model predicted correctly 98% ready biodegradable structures and 96% not ready biodegradable structures.

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

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

According to the original Catabol publication (Jaworka et al, 2002): cross-validation four times leaving 25% of data resulted in Q^2 = 0.88 between observed and predicted values.

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:

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:

All

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

All

7.5.Other information about the external validation set:

7.6.Experimental design of test set:

7.7.Predictivity - Statistics obtained by external validation:

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:

Catalogic is a mechanistic modelling approach for the quantitative assessment of biodegradability in biodegradation pathways. It can be considered as a hybrid system, containing a knowledge-based expert system for predicting biotransformation pathway combined with a probabilistic model that calculates probabilities of the individual transformation and overall BOD and/or extent CO_2 production.
The core is its biodegradability simulator including a library of hierarchically ordered individual transformations (catabolic steps) and a matching substructure engine providing their subsequent performance. Biodegradation is based on the entire pathway not the parent structure alone. It contains over 550 principle transformations; they often include more than one real biodegradation step to improve speed of predictions. Before computing the transformation of a target fragment, adjacent fragments are checked for inhibiting fragments. These inhibiting fragments can completely prevent the execution of the transformation or may assign a lower probability for the reaction to take place. There are three or four inhibiting fragments per transformation and thus, over 2000 combinations of principal transformations and inhibiting fragments in the system. The Catalogic system is trained to predict ready biodegradation within 28 days, under ready biodegradation conditions, on the basis of 743 chemicals from MITI database and another training set of 109 proprietary chemicals from Procter & Gamble (P&G) obtained with the OECD 301C and OECD 301B tests, respectively. In the first database biodegradation is expressed as the oxygen uptake relative to theoretical uptake, while in the P&G database biodegradation is measured by CO_2production. The catabolic steps are derived from a set of most plausible metabolic pathways

predicted by experts for each chemical in the training set. The MITI-I database is used to provide the widest structural diversity and the most consistent biodegradability assessments (O2 yield during OECD 301 C test) among existing data collections. For some transformations, fragments called “masks” are attached to a source fragment. These inactivating fragments prevent the performance of a specific transformation. However, the same reactions may occur for the second time with lower probability but no masks. The consequence is that if such a reaction is not executed the first time it is encountered because of the mask it will be executed later but with a lower probability. Currently the set of transformations includes 141 abiotic and biologically mediated reactions, which occur very rapidly, compared to the duration of the biodegradation tests. These rapid biotransformations were predicted to occur with the following highly reactive groups and intermediates: oxiranes, ketenes, acyl halides, thiocarboxylic acids, hydroperoxides, nitrenes and geminal diols. Various chemical equilibrium processes like carboxylic acids hydrolysis, keto-enol tautomerism, thiolthiol tautomerism and cyanuric acid isomerisation were also included in this class of transformation. Many of the other 465 metabolic transformations such as oxidation, hydrolysis, decarboxylation and dehalogenation were grouped into subsets of reactions depending on the similarity of their target fragment and transformation products. The probabilities of 324 rate-determining reactions grouped in 50 subsets were estimated on the basis of experimental biodegradation data. Due to lack of sufficient probabilities the remaining 141 reactions were determined on the basis of expert knowledge. The principle transformation steps are divided into two types of reactions: spontaneous and catabolic. Spontaneous transformations may be biotic or abiotic, including, for example, spontaneous hydrolysis. Catabolic transformations describe only biotic processes. The hierarchy of transformations is set according to descending probabilities of individual transformations that are derived from the model. Catalogic was created to predict the most probable biodegradation pathway, the distribution of stable metabolites and the extent of biological oxygen demand or CO2 production compared to theoretical limits. Catalogic matches the parent molecule with the source fragment associated with each transformation starting with the transformation having the highest probability of occurrence. When a match is identified, the molecule is metabolised and transformation products are treated as parent molecules. The procedure is repeated for the newly-formed metabolite until the product of probabilities of consecutive performed transformations reaches a user defined threshold. The sequence of transformations that is obtained represents the most plausible catabolic pathway for the biodegradation of the parent chemical.

8.2.A priori or a posteriori mechanistic interpretation:

A priori

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]MITI-I database http://www.cerij.or.jp/ceri_en/otoiawase/otoiawase_menu.html.

[2]Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemala J & Mekenyan O (2005). A stepwise approach for defining the applicability domain of SAR and QSAR models. Journal of Chemical Information and Computer Science 45, 839-849.

9.3.Supporting information:

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

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q13-301-0040

10.2.Publication date:

2013-06-27

10.3.Keywords:

Catalogic;expert system;biodegradation;Biochemical Oxygen Demand;BOD;

10.4.Comments:

former Q9-21-18-132