Chemosphere 108 (2014) 10–16

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

A new *in silico* classification model for ready biodegradability, based on molecular fragments

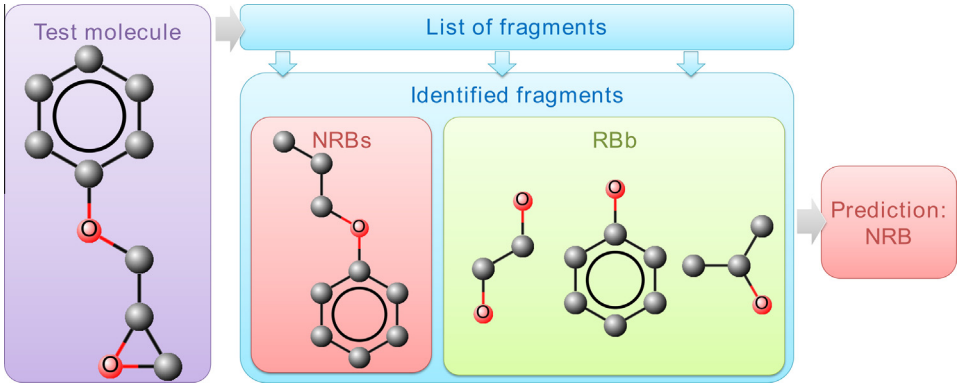
Anna Lombardo [a], Fabiola Pizzo [a], Emilio Benfenati [a,*], Alberto Manganaro [a], Thomas Ferrari [b], Giuseppina Gini [b]

[a] *IRCCS – Istituto di Ricerche Farmacologiche Mario Negri, Via G. La Masa 19, 20156 Milano, Italy*
[b] *Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, Piazza L. da Vinci 32, 20133 Milano, Italy*

HIGHLIGHTS

- A new fragment-based model to predict ready biodegradability was developed.
- A new software to extract fragments was used: SARpy.
- Statistical and expert-based fragments were used to build the new model.
- The model is freely available and useful for regulatory purposes.
- The model has performance comparable to other existing models.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 27 November 2013
Accepted 22 February 2014

Handling Editor: S. Jobling

Keywords:
Ready biodegradability
REACH
QSAR
Fragment-based model
SARpy

ABSTRACT

Regulations such as the European REACH (Registration, Evaluation, Authorization and restriction of Chemicals) often require chemicals to be evaluated for ready biodegradability, to assess the potential risk for environmental and human health. Because not all chemicals can be tested, there is an increasing demand for tools for quick and inexpensive biodegradability screening, such as computer-based (*in silico*) theoretical models. We developed an *in silico* model starting from a dataset of 728 chemicals with ready biodegradability data (MITI-test Ministry of International Trade and Industry). We used the novel software SARpy to automatically extract, through a structural fragmentation process, a set of substructures statistically related to ready biodegradability. Then, we analysed these substructures in order to build some general rules. The model consists of a rule-set made up of the combination of the statistically relevant fragments and of the expert-based rules. The model gives good statistical performance with 92%, 82% and 76% accuracy on the training, test and external set respectively. These results are comparable with other *in silico* models like BIOWIN developed by the United States Environmental Protection Agency (EPA); moreover this new model includes an easily understandable explanation.

Abbreviations: AD, Applicability Domain; BOD, Biological Oxygen Demand; ECHA, European Chemicals Agency; FN, False Negative; FP, False Positive; MCC, Matthews Correlation Coefficient; (N)RB, (Non)Readily Biodegradable; OECD TG, Organisation for Economic Co-operation and Development-Test Guideline; PBT, Persistent, Bioaccumulative, Toxic; QSAR, Quantitative Structure–Activity Relationships; REACH, Registration, Evaluation, Authorization and restriction of Chemicals; SAR, Structure–Activity Relationships; SMARTS, SMiles Arbitrary Target Specification; SMILES, Simplified Molecular Input Line Entry System; TN, True Negative; TP, True Positive; vPvB, very Persistent very Bioaccumulative.

* Corresponding author. Tel.: +39 02 39014420; fax: +39 02 39014735.
E-mail address: emilio.benfenati@marionegri.it (E. Benfenati).

http://dx.doi.org/10.1016/j.chemosphere.2014.02.073
0045-6535/

1. Introduction

With their multiple possibilities of prolonged contact with sensitive targets, chemicals that are stable in the environment arouse concern because their potential harmful effects may last longer and become chronic. Generally if a chemical is labile it is easier to investigate its exposure scenarios and the chronic effects may be less important. It is therefore important to assess whether a chemical is persistent in the environment.

REACH legislation (Registration, Evaluation, Authorization and restriction of Chemicals) (REACH, 2006) aims to raise the level of protection for human health and the environment against the risk of exposure to chemicals. Persistence is addressed under REACH and ready biodegradability is a screening test for persistence. All chemicals produced or imported for more than one ton/year must be tested for ready biodegradability (REACH, 2006) (Annex VII of REACH). Persistent and Non-Readily Biodegradable (NRB) are not synonymous: the definitions and thresholds are different. A compound is defined as persistent if it resists degradation and remains in the environment for a long time (ECHA, 2008a). It is considered persistent if its degradation half-life reaches the thresholds of 60 d in marine water, 40 d in fresh or estuarine water, 180 d in marine sediment and 120 d in fresh or estuarine water sediment and in soil, as in the new Annex XIII of REACH (REACH, 2011).

Ready biodegradability is defined as a screening test in which a high concentration of the test substance is used and ultimate biodegradation is measured by non-specific parameters under aerobic conditions. A substance is considered Readily Biodegradable (RB) when it degrades by 60% within 28 d (OECD, 1992). This means that a RB compound is also considered non-persistent but a NRB one is not necessarily considered persistent without further tests.

The reference test for ready biodegradability is the OECD TG 301 (Organisation for Economic Co-operation and Development-Test Guideline; OECD, 1992). Besides the European Community, USA, Canada, and Japan have adopted the OECD TG 301C test for evaluating ready biodegradability (OPPTS, 2008; CEPA, 1999; Yoshioka, 2007).

Within REACH the use of Structure–Activity Relationships (SAR) and Quantitative Structure–Activity Relationships (QSAR) models is encouraged. These examine the compound's properties starting from its chemical structure, exploiting the principle that similar compounds should have similar biological activities (ECHA, 2008b). SAR focuses on the rule determining the relationship, as a classifier, while QSAR quantitatively assesses of the effect (regression model).

We used SARpy (Ferrari et al., 2011) to build up a classifier for ready biodegradation. This new general software automatically extracts knowledge from a dataset and detects the molecular structural fragments associated with the activity of interest. The model we developed, based on ready biodegradability data for the OECD TG 301C – modified MITI – I test, predicts whether a compound is RB or not, to screen its persistence for the PBT (Persistent, Bioaccumulative, Toxic)/vPvB (very Persistent very Bioaccumulative) assessment.

2. Materials and methods

2.1. Data

The dataset described in (Toropov et al., 2012) was used. Two compounds were eliminated, one inorganic and one tautomer. The final dataset of 728 compounds was split into a training set (582 compounds) and a test set (146 compounds), amounting to respectively 80% and 20% of the total maintaining the same proportions of classes as the original set in both subsets.

After the development of the model a new data set was available (Cheng et al., 2012), so their continuous and binary data were extracted and combined in a single dataset. The doubtful compounds (or data), compounds with a percentage of BOD > 100% and duplicates were eliminated. If multiple data were available for the same compounds, the arithmetic mean was maintained if the data were consistent, otherwise the compound was eliminated. From this extended dataset we used the compounds not present in the training or the test set of the model presented here, for a total of 874 new compounds, as the external set.

2.2. Software

SARpy takes in input a set of chemical structures paired with their experimental activity label and produces as output a set of structural fragments associated with the property under investigation. The input and the output structures of SARpy are all expressed as Simplified Molecular Input Line Entry System (SMILES); a SMILES is a string of characters that provides a compact representation of the structure of a molecule (http://www.daylight.com/dayhtml/doc/theory/).

SARpy applies to the input structures (the training set) a fragmentation process to extract all the substructures, within a customizable size range, expressed as the number of atoms (usually 2–18). Then, the software mines for correlations between the incidence of any molecular substructure and the activity of the molecules containing it. Finally, a subset of fragments is selected and proposed to the user in the form of rules "IF fragment THEN activity".

As outcome SARpy lists the SMILES fragments paired with an activity label (e.g., positive, negative), ordered by descending precision in identifying the property under investigation. The statistical measure used for the precision is a likelihood ratio that is computed for each fragment from the ratio of positive (True Positives, TP) to negative predicted as positive (False Positive, FP) elements in the subset of molecules containing the fragment, and the ratio of negative to positive elements in the whole training set.

$$\text{likelihood ratio} = (\text{TP}/\text{FP}) \times (\text{negatives}/\text{positives}) \quad (1)$$

The likelihood can be used as a quantitative attribute of the fragment. Thus, the first fragments in the list identify the molecules with the desired activity label with almost no errors, then come the fragments with a higher misclassification rate. A more detailed description of SARpy is in (Ferrari et al., 2011, 2013); its code is available from the authors.

SARpy can be customized to improve the specificity of the model, or in a more balanced way to improve the accuracy. We obtained different series of fragments (called rule-sets) considering as active the RB compounds (and inactive the non-ready biodegradable ones). Each rule-set was obtained using the settings specified in "Supporting Information A".

3. Results and discussion

3.1. The procedure for obtaining the rules

The fragments for this model derive both from a statistical part and an expert-based part. The modeling has been done in three steps (Fig. 1). Initially, four rule-sets of fragments were generated with SARpy: NRB fragments with high specificity (rule-set 1), NRB fragments with balanced performance (rule-set 2), RB fragments with high specificity (rule-set 3) and RB fragments with