Environmental Research 142 (2015) 161–168

Contents lists available at ScienceDirect

Environmental Research

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

Prioritization of *in silico* models and molecular descriptors for the assessment of ready biodegradability

Alberto Fernández [a], Robert Rallo [b,*], Francesc Giralt [a]

[a] Departament d'Enginyeria Quimica, Universitat Rovira i Virgili, Tarragona, Catalonia, Spain
[b] Departament d'Enginyeria Informatica i Matematiques, Universitat Rovira i Virgili, Tarragona, Catalonia, Spain

ARTICLE INFO

Article history:
Received 28 March 2015
Received in revised form
22 May 2015
Accepted 25 June 2015
Available online 7 July 2015

Keywords:
Prioritization
In silico
Qsar
Biodegradability

ABSTRACT

Ready biodegradability is a key property for evaluating the long-term effects of chemicals on the environment and human health. As such, it is used as a screening test for the assessment of persistent, bioaccumulative and toxic substances. Regulators encourage the use of non-testing methods, such as *in silico* models, to save money and time. A dataset of 757 chemicals was collected to assess the performance of four freely available *in silico* models that predict ready biodegradability. They were applied to develop a new consensus method that prioritizes the use of each individual model according to its performance on chemical subsets driven by the presence or absence of different molecular descriptors. This consensus method was capable of almost eliminating unpredictable chemicals, while the performance of combined models was substantially improved with respect to that of the individual models.

1. Introduction

Chemicals identified as Persistent, Bioaccumulative and Toxic (PBT) or very Persistent and very Bioaccumulative (vPvB) are of great concern due to their potential impact on the environment and human health. Even if the emission of a persistent substance into the environment is stopped, this may not necessarily result in a reduced environmental concentration and, subsequently, in biota. Accordingly, persistent chemicals may result in chronic exposure and long term and cumulative adverse effects that can propagate through food webs. The assessment of biodegradation constitutes a very important issue being addressed by governmental regulatory agencies. The European regulation on Registration, Evaluation, Authorization and Restriction of Chemicals (REACH, 2006) requires an evaluation of persistence for prioritization (i.e., the identification of the most dangerous chemicals). This constitutes a critical decision-making process since massive testing and analyses are needed to fill the enormous gaps of missing information that is required. The huge quantity of chemicals that have to be evaluated makes the experimental testing of every single compound an impossible task. Thus, in order to minimize economic costs and time, data prediction and evaluation for REACH require the development of new methods suitable to optimize existing data and integrate information gathered by non-testing methods such as quantitative structure–activity relationship ((Q)SAR) models. In fact, the European Chemicals Agency guidance (ECHA, 2008) indicates that before any new test is carried out, all available data should be used, including data from valid *in silico* models.

Methods for data integration have been reported to be a suitable approach for increasing the reliability in decision making when multiple sources of information are available. These methods have been extensively used for the combination of raw data obtained from several sensors (sensor fusion) in environmental applications (Ashraf et al., 2012). A different approach consists in the integration of different models through a boosting procedure to obtain a new, more robust and accurate model (Rallo et al., 2005). On the other hand, weight-of-evidence approaches are quantitative decision-making methods that integrate a set of evidences by weighting them according to experts' criteria (Benedetti et al., 2012; Morales-Caselles et al., 2008). Alternatively, decision-making approaches can be implemented to integrate data by using optimization methods such as multicriteria decision analysis (Passuello et al., 2012; Zabeo et al., 2011). Qualitative consensus models, such as Dempster–Shafer theory of evidence, have also been reported for consensus reasoning with discrete datasets (Fernández et al., 2009). A similar approach for quantitative consensus can be applied to integrate the output of different *in silico* models for environmental endpoints, in a statistically sound manner by using Bayesian rules (Fernández et al., 2012).

The aim of this study is to develop a new consensus approach suitable for integrating diverse non-testing methods for ready biodegradability assessment. The purpose of this integration is to reduce economic costs and to optimize the use of available information. Instead of developing a new model for ready

* Corresponding author.

http://dx.doi.org/10.1016/j.envres.2015.06.031
0013-9351/

biodegradability, the goal was to take advantage of the good properties of freely available *in silico* models and to improve their classification scores. To this end, all sources of information were prioritized according to their efficiency in predicting the ready biodegradability of chemicals in subsets driven by the presence or absence of different molecular descriptors. More concretely, the prioritization of existing QSAR models was based on the highest predictivity scores that these models obtained for several subsets of chemicals containing a concrete list of molecular descriptors. In case of conflict between the classifications obtained from two distinct models, a consensus is reached on the most probable class assignment. This consensus is achieved by means of selecting the classification delivered by the QSAR model with the highest predictivity score for the particular list of descriptors present in the chemical substance under study.

2. Materials and methods

2.1. Ready biodegradability data

The Japanese Ministry of International Trade and Industry (MITI) (I) test, described in the Organization for Economic Cooperation and Development guideline 301 C (OECD, 1992), is a screening test used to determine the ready biodegradability. The MITI (I) test measures the biochemical oxygen demand (BOD) in aerobic aqueous medium for 28 days. Chemicals with a BOD value greater or equal than 60% are considered as readily biodegradable (RB), whereas those with a BOD value lower than 60% are regarded as not readily biodegradable (NRB). The MITI data collected from four sources were used in the current study.

A dataset containing 884 chemical compounds was obtained from BIOWIN's help manual (US EPA, 2012; Tunkel et al., 2000). The dataset consisted of 385 RB chemicals and 499 NRB chemicals. The dataset was divided by BIOWIN developers into a training dataset, with 589 compounds, and a validation dataset, with 295 compounds. Since the training dataset was used to develop the BIOWIN5 and BIOWIN6 models and both models were included in the consensus approach, only the 295 compounds in the validation dataset (131 RB and 164 NRB) were kept, so that BIOWIN models did not have any advantage over the other QSAR models.

A second biodegradability dataset was obtained from Cheng et al. (2012). It contained 1631 compounds (595 RB and 1036 NRB), all of them tested under the Japanese MITI (I) test protocol. This dataset was obtained from several sources, including BIOWIN's dataset; therefore, repeated chemical substances had to be excluded.

Additional experimental data of the Japanese MITI (I) test were obtained from Mansouri et al. (2013), who gathered a dataset containing 1725 chemicals (547 RB and 1178 NRB). These data were collected from several sources of information, including some molecules from the dataset compiled by Cheng et al. (2012), which had to be removed to avoid duplicates.

The fourth dataset was obtained from Lombardo et al. (2014), and it was composed of 728 MITI (I) test values. This dataset was divided by VEGA developers (VEGA, 2014) into a training dataset, containing 582 compounds, and a validation dataset, containing 146 compounds. Since the training dataset was used to develop the VEGA model, and this model was included in the consensus approach, only the 146 compounds in the validation dataset (69 RB and 77 NRB) were kept, so that VEGA did not have any advantage over the other QSAR models. In addition, some of these substances had to be removed because they had been initially extracted from BIOWIN and therefore they were duplicated in our dataset.

Chemical compounds from the four datasets were compared using the Chemical Abstracts Service Registry Number (CASRN), what resulted in a unified dataset containing 1873 distinct substances. In order to obtain a highly reliable dataset, we discarded 344 chemicals appearing in just one dataset and from the remaining 1529 compounds we removed 17 cases where the reported experimental values were contradictory. Last, we discarded chemicals that had been used for training BIOWIN or VEGA models, resulting in a final dataset of 757 chemical substances. 104 (14%) of these substances were present in the BIOWIN validation set, 730 (96%) were present in the Cheng et al. (2012) dataset, 751 (99%) of these chemicals were present in the Mansouri et al. (2013) dataset, and 52 (7%) were present in the VEGA validation set. Regarding the classification of these chemical substances into one of the two ready biodegradability classes, there were 232 (31%) RB and 525 (69%) NRB.

2.2. Molecular descriptors

Dragon v6.0 (Talete srl, 2014) was applied to compute molecular descriptors for all the chemicals in our dataset. Only integer-valued descriptors were considered because they are best to split the chemicals dataset into two subsets: one containing the chemicals with presence (a value greater than zero) of a given descriptor, and another subset containing the chemicals with absence (a null value) of the same descriptor. For example, the molecular descriptor counting the number of ring systems (NRS) splits the compounds dataset into one subset of 509 chemicals containing rings, and another subset of 248 chemicals without rings. These two subsets were going to be used separately to assess the performance of each *in silico* model. Dragon calculates a total of 4885 molecular descriptors classified in twenty-nine logical molecular descriptors blocks. Non-informative descriptors, with constant values in our chemicals dataset, were removed. The 1048 integer-valued descriptors extracted from Dragon fell inside the seven molecular descriptors blocks listed in Table 1.

For efficiency reasons, molecular descriptors that were equivalent to our purposes (i.e., descriptors that split our dataset into the same two presence/absence subsets) were put together into groups of descriptors, simplifying the initial 1048 individual descriptors into 637 groups of descriptors. For example, the following ring descriptors were grouped together because they divided the chemicals dataset into the same two subsets: cyclomatic number of rings (nCIC), number of circuits (nCIR), total ring size (TRS), ring perimeter (Rperim), and number of ring systems (NRS).

2.3. QSAR models for ready biodegradability

Only freely available QSAR models that predict ready biodegradability were considered in the current work. Any commercial software was excluded from the analysis (Pavan and Worth, 2008). The performance of some of these models has been recently compared elsewhere (Pizzo et al., 2013; Boethling, 2014). The

Table 1
Number of Dragon molecular descriptors used in this work, divided in seven logical molecular descriptors blocks.

Block	Descriptors
Constitutional indices (CI)	20
Ring descriptors (RD)	17
Functional group counts (FGC)	112
Atom-centered fragments (ACF)	100
Atom-type E-state indices (ESI)	42
Chemically Advanced Template Search 2D descriptors (CATS)	129
2D Atom pairs (AP)	628
Total	1048