Environment International 88 (2016) 250–260

Contents lists available at ScienceDirect

Environment International

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

Full length article

A new integrated in silico strategy for the assessment and prioritization of persistence of chemicals under REACH

Fabiola Pizzo [a,*], Anna Lombardo [a], Marc Brandt [b], Alberto Manganaro [c], Emilio Benfenati [a]

[a] IRCCS — Istituto di Ricerche Farmacologiche "Mario Negri", Department of Environmental Health Sciences, Laboratory of Environmental Chemistry and Toxicology, Via La Masa, 19, 20159 Milan, Italy
[b] Umweltbundesamt (UBA) — German Federal Environment Agency, Wörlitzer Platz 1, 06844 Dessau-Roßlau, Germany
[c] Kode s.r.l., Via Nino Pisano, 14, 56122 Pisa, Italy

ARTICLE INFO

Article history:
Received 13 October 2015
Received in revised form 30 November 2015
Accepted 16 December 2015
Available online xxxx

Keywords:
Persistence
k-NN
Structural alerts
Chemical classes
PBT
In silico

ABSTRACT

The fact that chemicals can be recalcitrant and persist in the environment arouses concern since their effects may seriously harm human and environmental health. We compiled three datasets containing half-life (HL) data on sediment, soil and water compartments in order to build in silico models and, finally, an integrated strategy for predicting persistence to be used within the EU legislation Registration, Evaluation, Authorisation and restriction of CHemicals (REACH). After splitting the datasets into training (80%) and test sets (20%), we developed models for each compartment using the *k*-nearest neighbor algorithm (*k*-NN). Accuracy was higher than 0.79 and 0.76 respectively in the training and test sets for all three compartments. To support the *k*-NN predictions, we identified some structural alerts, using SARpy software, with a high-true positive percentage in the test set and some chemical classes related to persistence using the software IstChemFeat. All these results were combined to build an integrated model and to reach to an overall conclusion (based on assessment and reliability) on the persistence of the substance. The results on the external validation set were very encouraging and support the idea that this tool can be used successfully for regulatory purposes and to prioritize substances.

1. Introduction

According to the International Council of Chemical Associations (ICCA) (ICCA, 1996) persistence is defined as the ability of a substance to remain unchanged in the environment for a long time. Such chemicals persist in the environment since their physical, chemical and biological degradation is slow, leading to the possibility of accumulation in the environment and biota, and even, to cause chronic effects. Persistence refers in particular to chemicals with degradation half-lives (HL) higher than certain trigger values in water, sediment or soil compartments (EChA Guidance, 2014). HL is normally used as a parameter to calculate persistence and is defined as the time needed to remove half of the starting amount of a substance from the environment (ECETOX, 2003).

Persistence is also linked with the ability of a substance to be present in environments distant from the emission source, or to degrade slowly in laboratory conditions (Boethling et al., 2009). Therefore, considering the high potential for adverse effects and transport to distant environments, assessment of the persistence of these compounds requires careful investigation (ECETOX, 2003).

The need to identify persistent substances is also connected to the prioritization and screening of persistent, bioaccumulative, and toxic (PBT) or very persistent, very bioaccumulative (vPvB) chemicals, under several worldwide regulations (Gramatica and Papa, 2007).

Under the new EU chemical legislation Registration, Evaluation, Authorisation and restriction of CHemicals (REACH) which aims to protect human and environmental health, assessment of persistent, bioaccumulative and toxic (PBT) properties is mandatory for substances manufactured and/or imported at least of 1 tonne/year (REACH, 2006).

In REACH ready biodegradability is used as a screening test for the assessment of PBT/vPvB substances (Lombardo et al., 2014). A substance that biodegrades in an experimental test system is considered not persistent in the environment (Pavan and Worth, 2007).

Computer software can be used to examine the fate and degradation properties of substances. With powerful computers and sophisticated software, we can now assess the persistence of chemicals using models (ECETOX, 2003). REACH regulation clearly encourages and promotes the use of computational models, such as (quantitative) structure–

Abbreviations: REACH, Registration, Evaluation, Authorisation and restriction of CHemicals; P, persistent; nP, non-persistent; vP, very persistent; nP/P, non-persistent/persistent; P/vP, persistent/very persistent; HL, half-lives; DT50, disappearance time 50; PBT, persistent, bioaccumulative, toxic; PBT/vPvB, persistent, bioaccumulative, toxic/very persistent, very bioaccumulative; POPs, persistent organic pollutants; RB, readily biodegradable; nRB, non-readily biodegradable; *k*-NN, *k*-nearest neighbor algorithm; (Q)SAR, (quantitative) structure–activity relationship; SAs, structural alerts; TP, true positive; TN, true negative; FP, false positive; FN, false negative; MCC, Matthew's correlation coefficient; AD, applicability domain; ADI, applicability domain index; EChA, European Chemicals Agency; US EPA, United States Environmental Protection Agency.

* Corresponding author at: Via La Masa 19, 20159 Milan, Italy.
E-mail address: fabiola.pizzo@marionegri.it (F. Pizzo).

http://dx.doi.org/10.1016/j.envint.2015.12.019
0160-4120/

activity relationship ((Q)SAR), to get information on the toxicity of substances, for classification and labeling, risk assessment and the initial identification of potential PBT properties when no experimental data is available (REACH, 2006). However, one of the major limitations for fate modeling, is the scarcity of degradation data.

Here we propose an integrated approach that can help to assess the potential for persistence of organic chemicals. This tool can support decision-making for substance management (assessment and prioritization) by providing a final evaluation as the result of a combined system of computational models that predict persistence in sediment, soil and water compartments. To obtain the safest evaluation, the overall prediction is conservative, meaning that if different classes of persistence are predicted in the three compartments, the final assessment will be the worst class found.

2. Materials and methods

2.1. Data collection

2.1.1. Training and test sets

HL data were collected from several sources. Gouin et al. (2004) give information on HL, expressed in hours, for 233 organic compounds in nine classes (on a semi-decade log scale basis). It covers four environmental media: water (not specified whether marine or fresh water), sediment (not specified whether marine or fresh water), soil and air. An average HL is assigned to each class, which is the only value available for each chemical. Gramatica and Papa (2007) HL provide data for 250 organic compounds, referring to the same compartments and classified as in Gouin et al. (2004). Since no thresholds for air are defined for PBT/vPvB assessment, we took account only of sediment, soil and water environments. Since no information was available, for water and sediment, we considered all data as referring to fresh water environment. For all the compounds present in both datasets we double-checked the chemical structures and their correspondence with CAS number with ChemIDplus (http://chem.sis.nlm.nih.gov/chemidplus/) and Pubchem Compound (https://pubchem.ncbi.nlm.nih.gov/). Salts, mixtures, doubtful compounds and duplicates were eliminated, as well as duplicate compounds with conflicting experimental values. We obtained datasets of 297 organic compounds for sediment, 298 for soil and 298 for water. Another source was available from United States Geological Survey (USGS) only for soil, containing 318 HL. These compounds were checked as above and the continuous values were classified following the same criteria as in Gouin et al. (2004), and then added to the soil dataset, obtaining a dataset of 537 compounds.

In the RIVM Report (Linders et al., 1994) disappearance time 50 (DT50) for water and soil compartments were also available. DT50 differs from HL because HL refers to first or pseudo-first order reactions. However, we assumed DT50 as overlapping HL. These compounds too were checked, classified and added to the water and soil datasets as above, obtaining final datasets of 351 and 568 data respectively for water and soil.

Each data source used has disadvantages: USGS and RIVM contain only pesticides; moreover, RIVM data are quite old and are not HL but DT50. In Gouin et al. (2004) and Gramatica and Papa (2007) data were classified but it is impossible with these categories to distinguish P and vP compounds since they do not exactly fit in with the precise thresholds defined for PBT and vPvB (Table 1). The thresholds for P were 960 h (40 days) for fresh water, 2880 h (120 days) for soil and sediment in fresh water; for vP we used the thresholds of 1440 h (60 days) for fresh water and 4320 h (180 days) for soil and sediment in fresh water. Since RIVM and USGS sources contained continuous values and it is more logical to convert continuous values into categories than vice versa, we classified them as in Gouin et al. (2004) and Gramatica and Papa (2007). Since we had figures outside the time frames considered by Gouin et al., we added two more categories (10 and 11, Table 1 Supplementary material) and divided the values into four classes: nP compounds (with HL/DT50 below the P threshold), nP/P compounds (including both nP and P compounds), P/vP compounds (including both P and vP compounds) and vP compounds (i.e. compounds over the vP threshold). Soil and water datasets were unbalanced, with a prevalence of nP compounds, while sediment was balanced. The percentages of compounds in the four classes (nP, nP/P, P/vP, vP) for each compartment are given in Fig. 1.

2.1.2. Validation set

To test the real ability of the model to recognize harmful substances we also compiled an external dataset, using compounds present in the Candidate List of substances of very high concern for authorisation available on the European Chemicals Agency (EChA) website. Therefore 161 molecules were available; however, the information on PBT/vPvB activity was provided only for a few (26 molecules, 16%). Mixtures, unknown or variable composition, complex reaction products or biological materials (UVCB) and inorganic compounds were excluded from our selection. We finally obtained a dataset containing information on PBT assessment for 12 substances, that we used for our purpose.

2.2. Software used for modeling

2.2.1. IstKNN

IstKNN is a software developed by Kode, for rapidly developing *k*-nearest neighbor (*k*-NN) models using different settings (Manganaro et al., 2016). The customizable settings are: “k”, that is the number of similar compounds retained for the analysis; “S1” is the similarity index threshold between similar compounds, if only one similar compound is available the similarity index must to be equal to or higher than a given threshold “S2”. Finally “E”, called the enhance factor enhances the role of molecules with higher similarity values in the prediction. The *k*-NN algorithm estimates the outcome (i.e. a continuous or categorical value) for a target compound on the basis of read-across accounting for its most similar compounds (i.e. nearest neighbors) present in the model's training set for which the toxicological activity is known (Altman, 1992). The datasets were randomly split into training (80%) and test (20%) sets as suggested in Martin et al. (2012) and Davis (2014); however the splitting was random but considered the numbers of compounds in each class (nP, nP/P, P/vP, vP). Thus for each persistence class considered for *k*-NN analysis we used 80% for the training set and 20% for the test set. The software was run in-batch and several models were generated; we selected those with a good balance between general performance and number of missing values.

The general accuracy calculated by IstKNN as the ratio of true-positive (TP) to the total number of predictions, was measured in the training and test sets. To clearly understand whether the models

Table 1
Half-life classes on the basis of the available data.

	nP		nP/P		P/vP		vP	
	Class range (h)	Value assigned (h)	Class range (h)	Value assigned (h)	Class range (h)	Value assigned (h)	Class range (h)	Value assigned (h)
Sediment	0–1000	5–550	1000–3000	1700	3000–10,000	5500	10,000–1,000,000	17,000–550,000
Soil	0–1000	5–550	1000–3000	1700	3000–10,000	5500	10,000–1,000,000	17,000–550,000
Water	0–300	5–170	300–1000	550	1000–3000	1700	3000–1,000,000	5500–550,000