TyPol - a new methodology for organic compounds clustering based on their molecular characteristics and environmental behavior

Rémi Servien, Laure Mamy, Ziang Li, Virginie Rossard, Eric Latrille, Fabienne Bessac, Dominique Patureau, Pierre Benoit

TyPol - a new methodology for organic compounds clustering based on their molecular characteristics and environmental behavior

Rémi Servien[a,b,*], Laure Mamy[c], Ziang Li[d], Virginie Rossard[b], Eric Latrille[b], Fabienne Bessac[e,f,g], Dominique Patureau[b], Pierre Benoit[d]

[a] *INRA, Université de Toulouse, UMR 1331 Toxalim, Research Centre in Food Toxicology, F-31027 Toulouse, France (present address)*

[b] *INRA, UR 050, Laboratoire de Biotechnologie de l'Environnement, Avenue des Etangs, F-11100 Narbonne, France*

[c] *INRA, UR 251 PESSAC, Route de St Cyr, F-78026 Versailles, France*

[d] *UMR 1091 INRA-AgroParisTech, Environnement et Grandes Cultures, F-78850 Thiverval-Grignon, France*

[e] *Université de Toulouse, INPT, Ecole d'Ingénieurs de Purpan, Equipe DINA, 75 voie du TOEC, BP 57611, F-31076 Toulouse Cedex 03, France*

[f] *Université de Toulouse, UPS, IRSAMC, Laboratoire de Chimie et Physique Quantiques, 118 route de Narbonne, F-31062 Toulouse, France*

[g] *CNRS (UMR 5626), F-31062 Toulouse, France*

[*] Corresponding author. Address: INRA, Université de Toulouse, UMR 1331 Toxalim, Research Centre in Food Toxicology, F-31027 Toulouse, France. Tel.: +33 (0)5 61 19 32 82; fax: +33 (0)5 61 19 39 17.

E-mail address: remi.servien@toulouse.inra.fr (R. Servien)

HIGHLIGHTS

- An innovative methodology, TyPol, was developed to classify organic compounds
- The classification is based on environmental behavior and molecular descriptors
- It relies on partial least squares analysis and hierarchical clustering
- The degradation products of organic compounds are considered
- The environmental fate of a "new" compound can be assessed from its affiliation to one cluster

ABSTRACT

Following legislation, the assessment of the environmental risks of 30 000 to 100 000 chemical substances is required for their registration dossiers. However, a very small proportion of these chemicals was already studied through actual measurements because it is time-consuming and/or cost prohibitive. Therefore, the objective of this work was to develop a methodology to classify organic compounds, and their degradation products, according to both their behavior in the environment and their molecular properties. The strategy relies on partial least squares analysis and hierarchical clustering. The calculation of molecular descriptors is based on an in silico approach, and the environmental endpoints (i.e. environmental parameters) are extracted from several available databases and literature. The classification of 215 organic compounds in 6 different clusters showed that the combination of some specific molecular descriptors was directly related to a particular behavior in the environment. TyPol also provided an analysis of similarities between organic compounds and their metabolites. Among the 24 metabolites that were inputted, 58% were found in the same cluster as their parents. The robustness of the method was tested and shown to be good. TyPol can help to predict the environmental behavior of a

"new" compound (parent compound or metabolite) from its affiliation to one cluster but also to select representative substances from a large data set in order to answer some specific questions regarding their behavior in the environment.

Keywords:
Pesticides
Metabolites
Clustering
Molecular modeling
Environmental fate
Partial least squares

1. Introduction

New legislations such as the REACH (Registration, Evaluation, Authorization and restriction of CHemicals) regulation in the EU will require that manufacturers of substances and formulators register and provide prescribed eco/toxicological data for substances with volume higher than one metric ton per year. It is estimated that about 30 000 existing substances have to be registered by 2018 by member states (Ahlers et al., 2008). The needed information has to be equivalent to the standard information requirement and adequate to draw overall conclusions with respect to the regulatory endpoints classification and labeling. Beyond specific regulatory needs, the same questions concern chemical substances that are potentially present in the environment and that originate from various sources. According to authors, from 30 000 to 100 000 chemical substances may be concerned by environmental risks assessment (Muir and

Howard, 2006). However, a very small proportion of these chemicals are studied through actual measurements either in laboratory tests or in environmental monitoring because it is time-consuming and/or cost prohibitive. Consequently, in silico approaches based on intrinsic properties of the substances represent an important challenge and have been a matter of large research efforts for the last 15 years (Mackay et al., 2001; Walker and Carlsen, 2002). Systematic approaches able to classify compounds according to their environmental behavior or eco/toxicological effects will help both regulators and scientists facing the problem of the constant increase in the diversity and in the number of the chemical substances which will be concerned by environmental risk assessment.

One of the most used in silico methods is the QSAR (Quantitative Structure Activity Relationships) which allows the estimation of the properties of a chemical from several descriptors (OECD, 1993a). The two main types of descriptors are experimental descriptors (such as octanol-water distribution coefficient or water solubility), and molecular descriptors that include constitutional, geometrical, topological, electrostatic, and quantum properties of the molecules (Sabljic, 2001; Doucette, 2003). These molecular descriptors represent the way chemical information is transformed and coded to deal with (bio)chemical, pharmacological and toxicological problems (Todeschini and Gramatica, 1997). Contrary to approaches based on molecular descriptors, approaches based on experimental descriptors are prone to experimental errors in the input variables. However, molecular descriptors accuracy also depends on the approximations chosen to make the calculations. The calibration of the theoretical calculations is driven by the compromise between accuracy and efficiency (Lohninger, 1994; Karelson et al., 1996). Another advantage of the exclusive use of molecular descriptors is that they are calculable for not yet synthesized compounds. Therefore, the objective of this work was to develop a

methodology, TyPol, to classify organic compounds and their degradation products according to both their behavior in the environment and their molecular properties.

Contrary to QSAR, this methodology (i) does not estimate the values of environmental properties of organic compounds but classifies these compounds according to both their behavior in the environment and their molecular characteristics, (ii) considers simultaneously several environmental processes (described by appropriate environmental parameters), (iii) considers a high diversity of organic compounds, not only those belonging to similar chemical families, (iv) allows rigorous classification and comparison of the compounds, (v) considers the degradation products.

TyPol, is based on statistical analyses combining environmental endpoints (i.e. environmental parameters like degradation half-life or bioconcentration factor) and molecular descriptors (molecular surface, dipole moment…). The calculation of molecular descriptors is based on in silico approach and the environmental parameters are extracted from available databases and from literature. Based on the values of relevant molecular descriptors, TyPol will allow the classification of one organic compound (parent or metabolite) in a group of compounds having the same environmental behavior.

The choice of the statistical method involved in TyPol is crucial for the reliability of the clustering. Principal components analysis (PCA) is often used in multivariate chemical characterizations to determine linearly uncorrelated variables that summarize the information contained in variables (Jackson, 1991; Snarey et al., 1997; Harju et al., 2002; Eriksson et al., 2006). These uncorrelated variables can also be used as an excellent basis to select a representative set of chemicals using clustering methods. Various clustering techniques have been employed in chemical mapping such as strategies based on PCA and hierarchical clustering

for selecting dissimilar organic substances (Rännar and Anderson, 2010), bayesian classifiers for chemical toxicity predictions (Mishra et al., 2011), network clustering (Saito et al., 2010), PCA-based method (Rännar and Anderson, 2011) or other statistical tools (Vogt and Bajorath, 2012). However, the problematic of TyPol is different than these ones because there are two sets of variables (molecular descriptors and environmental parameters) which are different by nature. Partial least squares regression (PLS) (Wold, 1996; Eriksson et al., 2006) can be used to find the fundamental relation between two sets of variables using a latent variable approach to model the covariance structures in these two spaces. PLS model tries to find the multidimensional directions in the observable variables (i.e. molecular descriptors) space that explain the maximum multidimensional variance direction in the predicted variable (i.e. environmental parameters) space. So PLS, as PCA, constructs uncorrelated variables which summarizes the information, but PLS takes into account the information of both observable and predictive variables. Moreover, PLS easily deals with missing values. For these reasons, PLS was preferred to PCA methods. After the PLS analysis, a validation step is performed to remove the chemicals that are not properly represented by the PLS model. Then, a hierarchical clustering algorithm, based on Ward clustering (Ward, 1963), is used on the new scores to cluster the organic compounds.

2. Materials and methods

2.1. Organic compounds

For the development of TyPol, 215 organic compounds (191 parent compounds and 24 degradation products) were selected (Tables A1, A2). The selection of these compounds was done according to three criteria: (i) high diversity of chemical families for the parent compounds,

(ii) wide ranges of variation of the values of environmental parameters and molecular descriptors (Tables 1, 2), (iii) availability of data for the environmental parameters (see 2.2.). The 191 parent compounds include (i) 117 pesticides taken in the main groups of pesticides (carbamates, organochlorines, organophosphorous, strobilurins, triazines, urea, phenoxyacids...), (ii) 30 polychlorinated biphenyls (PCB), (iii) 12 polycyclic aromatic hydrocarbons (PAH), (iv) 10 polychlorinated dibenzofurans (PCDF), (v) 9 phthalates, (vi) 7 polychlorinated dibenzodioxins (PCDD), and (vii) 6 miscellaneous compounds (drugs, auxine, hormone…) (Table A1). The ability of TyPol to classify metabolites compared to their parent substance was tested using 24 metabolites deriving from chloride pesticides (Table A2). As some metabolites are common to several parent substances, 26 pairs of parent-metabolite were inputted in TyPol.

2.2. Environmental processes and parameters

Five of the main processes involved in the fate of organic substances in the environment were retained: (i) dissolution, to describe the expected distribution of the compound between liquid, solid and gaseous phases; (ii) volatilization, which is related to the risk of transfer to atmosphere; (iii) adsorption, which is linked to the risk of transfer to water; (iv) degradation which controls the dissipation and/or the persistence; and (v) bioaccumulation, to consider the impacts on the organisms and the food chain. Each of these environmental processes can be described by several “endpoints” or “environmental parameters”. In this work, water solubility (*Sw*) and octanol-water partition coefficient (*Kow*) were selected as environmental parameters to describe dissolution, vapor pressure (P_{vap}) and Henry constant (K_H) for volatilization, adsorption coefficient normalized to soil carbon organic content (*Koc*) for adsorption, half-life (*DT50*) for degradation, and bioconcentration factor (*BCF*) for ecotoxicity (Table 1). These parameters were

chosen because they are the most common to represent the five environmental processes and because of the availability of the corresponding data in numerous databases.

The values of environmental parameters were mainly taken from the PPDB (2013) but also from literature. When values were not available in PPDB (mainly for metabolites), the missing values were collected from Mackay et al. (2006) and ChemSpider (2013). However, considering that a large amount of data of ChemSpider is estimated instead of measured, the use of this database was limited. When several values were available for one environmental parameter, the mean value was retained. For the 215 compounds, 1460 environmental parameters were inputted in TyPol, and there were only 3.9% of missing values. The ranges of values of the parameters are indicated in Table 1 for the 215 compounds.

2.3. Molecular descriptors: selection and calculation

The selection of the molecular descriptors was based on a literature review focused on the QSAR that were developed to estimate *Sw*, *Kow*, P_{vap}, K_H, *Koc*, *DT50*, and *BCF*. This review allowed the determination of the molecular descriptors that were best correlated to the selected environmental parameters.

More than one hundred publications were used for this selection. As indicated in the introduction, we focused on molecular descriptors (molecular surface, dipole moment…) rather than on experimental descriptors like *Kow* or *Sw*. In addition, five criteria were defined to choose the descriptors: (i) their relevance to estimate the environmental parameters (see the cited references below), (ii) their common use for the estimation of the seven parameters, (iii) the absence of redundancy between descriptors, (iv) the possibility to calculate the descriptors with molecular modeling, and (v) their ranges of variation. Finally, 40 constitutional, geometrical, topological, electrostatic and quantum descriptors were retained (Table 2) (see for example

OECD, 1993b; Katritzky et al., 2000; Sabljic, 2001; Dearden and Schüürmann, 2003; Doucette, 2003; Yang et al., 2003; Pavan et al., 2008).

CHEM-3D of ChemOffice Ultra 12.0 (2009) molecular modeling software was used to build three-dimensional chemical structures (3D-structures) in order to calculate the electrostatic and the quantum molecular descriptors (Table 2). As the values of these molecular descriptors are highly dependent on the 3D-structures, a conformational search was done as follows: structures were energy-minimized in MOPAC (Molecular Orbital PACkage) using the semi-empirical method AM1 (Austin Model parameterization) and ground electronic states were obtained as closed-shell molecular orbital wave functions in the restricted Hartree-Fock framework. Analytical frequency calculations have been performed at AM1 level to ensure the obtained structures are minima on the potential energy surface (PES). For each compound, we proceeded by successive steps calculating a large number of conformations deriving from each other by rotations around the different chemical bonds in order to find the global minimum. As first estimate, the descriptors of acido-basic molecules were calculated for their neutral form. The Excel function of ChemOffice was then used to calculate the molar masses and the Connolly surfaces. Finally, the constitutional (except the molar mass) and the topological descriptors were calculated with Dragon 5.5 (2007). For the 215 compounds, 8600 values of molecular descriptors were inputted in TyPol. Their ranges of values are indicated in Table 2.

2.4. Partial least squares regression

PLS regression is a latent variable method that compresses all the information in some new uncorrelated variables to summarize linear relations between two sets of variables. Traditionally, individuals are presented as plots with two components however two axes are not

always the optimal choice. Therefore, in this work, the optimal number of axes to perform clustering will be selected using the PRESS (Prediction Sum of Squares) criterion. In addition, PLS can deal with missing values by using the NIPALS (Non-linear Iterative PArtial Least Squares) algorithm. This algorithm allows performing PLS without removing the individuals with missing values and without estimating these missing values (Tenenhaus, 1998). However, the less there are missing values the more accurate the final results are.

2.5. Domain of validity

The knowledge of the domain of validity of the final clustering is important to avoid erroneous conclusions. However, as explained previously, the objective of this work was not to develop QSAR, so, the determination of a domain of validity cannot be done as in standard QSAR procedures (Boethling and Costanza, 2010). A priori, TyPol does not have a domain of validity and can be applied to all compounds. However, the use of the PLS algorithm can lead to compounds that are declared atypical by the algorithm. These compounds can be identified using the T^2 of Hotelling (Tenenhaus, 1998). If the T^2 value of a compound is above a calculated threshold, the compound is atypical on the PLS axes. Nevertheless, as one of the objective of the method is to assess the properties of novel compounds, the clustering of these atypical compounds is done by TyPol.

2.6. Hierarchical clustering

Clustering algorithms are used to assign similar objects into groups (called clusters) based on a similarity criterion chosen by the user. The algorithm used in this study is based on the Ward clustering, which keeps the growth of errors as small as possible by merging individuals or clusters. The final number of clusters is chosen after comparison of the heights of

the dendrogram, a statistical map which resumes Ward clustering. For the convenience of analyzing clustering of the compounds and their relevant metabolites, arrows linking the parent compounds to their metabolites were represented on the main axes of the PLS. The multivariate analysis was done in R 2.10.0.1 with the "mixOmics" (version 2.8-1) and "cluster" (version 1.13.1) packages.

2.7. Robustness of the method

To assess the robustness of the clustering method, a classical cross-validation algorithm was used. A fixed percentage of the whole sample is removed from the sample and the PLS is performed. Then, all compounds (including those which were removed to compute the PLS) are projected on the PLS axes and clustered by the hierarchical clustering algorithm. As external compounds (i.e. which were not included during the PLS algorithm) are added in this step, this method can assess the robustness of our methodology and, by consequence, its relevance. Finally, the obtained clustering is compared to the targeted clustering obtained with the PLS calculated on the whole sample. The closer the clustering is to the targeted one, the more robust the method is. This cross-validation study was performed a hundred times for different percentages of removed compounds and the clusterings were compared using the Adjusted Rand Index (Hubert and Arabie, 1985; Nguyen et al., 2009). This index is a measure of the similarity between two different clusterings. The closer it is to 1 (respectively to 0), the more (respectively less) the two clusterings are similar.

2.8. Computing tools

The information system is based on a management system for relational database MySQL DBMS-R (version 5.1), an Apache web server (version 2.2), and the statistical R software (also

used for graphs). The system is installed in a distribution Debian 6.0. The environmental parameters and molecular descriptors are inserted into the management system relational database server which interfaces with Tcl/Tk (Tool Command Language/ToolKit) made from the R software and "RODBC" library (version 1.3-2). Annotations on the data or results are also stored in the same database. Since the web interfaces are easily editable, statistical analyses of data are treated and helped by the R software Tcl/Tk interfaces. All data that are stored in the DBMS MySQL-R can be viewed via the web interface phpMyAdmin (version 3.3). Data can be imported from phpMyAdmin and new data can easily be inserted. Finally, TyPol was designed in order to easily adapt to other research questions giving the user the choice of the variables and the compounds of the study.

3. Results and discussion

The first step in the use of the TyPol methodology is the chemical mapping to select the number of components for the subsequent classification, then a hierarchical clustering is performed to identify the optimal number of clusters to classify the organic compounds. As this article is focused on the presentation of the development of TyPol, the results that are given as an illustration of the outputs of TyPol are not analyzed in details.

3.1. Chemical mapping by PLS

The choice of the number of PLS components is critical for the subsequent analysis and classification. The number of components which gave the lowest PRESS was therefore selected, it corresponded to the fourth first axes of the PLS.

The domain of validity of the analysis was studied by calculating the T^2 of Hotelling for the 215 compounds of the study. It appeared that 7 compounds were found as atypical by the four components of the PLS: chlordecone, mirex, kelevan, fosetyl, di-isodecyl, di-isononyl, and benzo(g,h,i)perylene. Indeed, it is well known that these compounds have an extreme behavior in the environment: for example, chlordecone, mirex and kelevan are very persistent (Marchand, 1989; ATSDR, 1995; Cabidoche et al., 2009; Dolfing et al., 2012) contrary to fosetyl which has a very low *DT50* (PPDB, 2013); and di-isodecyl, di-isononyl, and benzo(g,h,i)perylene have very high *Kow* values (PPDB, 2013). Chlordecone, benzo(g,h,i)perylene, mirex and kelevan also have very high connectivity indexes. Nevertheless, these compounds were taken into consideration for the subsequent analysis because they could be representative of other compounds.

The four-component PLS model has good statistical results: $R^2_X=0.77$, $R^2_Y=0.90$ and $Q^2_Y=0.44$. The first two components were the most important ones. The closer the compounds are in this score-plot, the more similar they are (Fig. 1). The main characteristic of the first component, which explains 40% of the variance, is the strong positive loadings for all the geometric and topological descriptors, and constitutional descriptors like the number of chlorine or halogen atoms. A contrario, the dipole moment and the total energy have strong negative loadings therefore have an opposite effect. The second axis explains 16% of the variance. On this axis, variables such as the number of chlorine or halogen atoms have a positive loading whereas the number of rotatable, double or simple bonds or the number of hydrogen, oxygen or total atoms have a negative loading (Fig. 2). Figure 2 also shows that many variables seem to be correlated, mainly the different connectivity and valence connectivity indexes. Similar analysis can be done for the third and the fourth axes selected by the PLS.

3.2. Clustering

Using a hierarchical clustering algorithm, several clustering, from 1 (all compounds in the same cluster) to 215 (all compounds in a different cluster), were obtained. The selection of the number of clusters is an important and difficult task, which is usually performed by plotting the heights of the dendrogram's node and looking for a break. The results showed that the best choice was to classify the compounds in 6 clusters. The size of the six clusters varied from 3 for to 55 compounds (Fig. 1), each cluster being characterized by specific features.

The cluster 1 contains 48 compounds and groups together all the thiocarbamates (4 compounds) and nearly 50% of the triazines, carbamates and ureas inputted in TyPol. This cluster is characterized by high values of total energy and polarizability and low values of different connectivity indexes, related to low K_H therefore low volatility and low *DT50* that is low persistence in the environment. Twenty-one of the 30 compounds of cluster 2 are PCB (over 31 inputted in TyPol). There are also 5 organochlorines and 3 PAH. Compounds of cluster 2 have low electric dipole moment and high total energy, linked to low *Sw* and high *DT50*, therefore high persistence in the environment. Cluster 3 shares some common traits with cluster 2 in the first two axes. Nevertheless, these two clusters are well separated in the two other axes of the PLS which are not plotted here in a sake of compactness. So, cluster 3 is composed of 55 compounds, including all PCDF, 10 organochlorines, 9 PCB, 90% of the PCDD, and 8 PAH (11 in the study). Similarly to cluster 2, the combination of high molecular masses and low number of hydrogen atoms is related to low values of *Sw* and very high *DT50* (high persistence). The cluster 4 contains 37 compounds including 7 strobilurin compounds, 6 of the 10 phthalates and 4 of the 5 triazoles. The main characteristics of this cluster are very high connectivity indexes,

polarizability, and number of hydrogen and carbon atoms connected to low *DT50* and low values of the P_{vap} and *Sw*. Among the 45 compounds of the cluster 5, there are 5 organophosphates, 4 triazines, all dinitroanilines and all chloroacetamides. This cluster is characterized by an important electric dipole moment and number of rotatable bonds for the descriptors and a low P_{vap} connected with a high *Sw*. The differences between this cluster and cluster 1, which are not obvious on the first two axes for all compounds, are more easily noticeable in the fourth axes of the PLS. Finally, as showed on Figure 1, cluster 6 is an extreme one. It contains mirex, kelevan and chlordecone. As discussed above, these 3 organochlorine insecticides have very particular chemical structures and an exceptional persistence in the environment (bishomocubane family). They have extraordinary high values of connectivity or valence connectivity indexes, polarizability, molecular mass, number of chlorine and other halogen atoms and K_H; and extremely low values of number of multiple bonds, total energy, HOMO energy and *Koc*. In addition, chlordecone could also be formed from kelevan (Dolfing et al., 2012; PPDB, 2013). Even on the third and the fourth axes of the PLS, these three compounds have extreme locations and cannot be aggregated with any other cluster. The other compounds that were detected as atypical by the T^2 of Hotelling are clustered in nearly all the clusters: cluster 1 for fosetyl, cluster 3 for benzo(g,h,i)perylene or cluster 4 for di-isodecyl and di-isononyl.

The robustness of the method was assessed, using the cross-validation method described above, and found to be high and not depending on a low number of values. The Adjusted Rand Index values were 0.92, 0.87, 0.84 and 0.80 if 1%, 10%, 20% and 50% of the compounds were removed, respectively. TyPol allows the classification of organic compounds according to a particular behavior in the environment which is related to the combination of the values of some specific molecular descriptors.

3.3. Parents-metabolites relationships

To test the ability of TyPol to classify degradation products compared to their parent compounds, 26 pairs of parents and metabolites were inputted (Table A2). The clustering made above using all compounds was retained for the analysis. Figure 3 shows the classification of the metabolites compared to their parents. Among all metabolites, 58% (i.e. 15 metabolites) were in the same cluster as their parents. Conversely, 42% (i.e. 11 metabolites) were not in the same cluster as their parents: 6 metabolites originating from parents in clusters 4 and 5 were in cluster 1; 2 metabolites of parent in cluster 3 were in cluster 2; and 3 metabolites of parents in clusters 1 and 4 were in cluster 5. These results are due to similarities (or dissimilarities) in terms of structure and behavior between parent compounds and their metabolites, but further tests need to be performed with other chemical families. The classification of metabolites compared to the parent compounds will allow the prediction of the behavior in the environment of potential metabolites and/or of metabolites for which no data are available. In addition, the different routes of degradation, i.e. biotic, abiotic (oxidation, dehalogenation…) will be added in the future to investigate if the change in cluster between a compound and its metabolite(s) is related to the type of degradation mechanism.

4. Conclusion

A novel approach for clustering organic compounds according to both their behavior in the environment and their molecular descriptors is presented. The approach is based on PLS regression and hierarchical clustering.

The classification of 215 organic compounds in 6 different clusters showed that the combination of the values of some molecular descriptors is directly related to a particular

behavior in the environment. The robustness of the method was studied and demonstrated to be high. Therefore, TyPol can help to predict the environmental behavior of a "new" compound from its affiliation to one cluster or to select representative substances from a large data set in order to answer some specific questions regarding their behavior in the environment. In addition, TyPol takes into account the metabolites of organic compounds. The analysis is based on the same methodology as above and highlights the similarities (or dissimilarities) between a parent substance and its degradation product. One of the next steps of this work will investigate if the change in cluster between a compound and its metabolite(s) is related to the type of degradation mechanism (oxidation, epoxidation, hydroxylation…). Additional environmental and ecotoxicological parameters will also be included in TyPol to refine the classification of compounds.

Acknowledgements

The authors acknowledge the Projet Innovant of the "Environnement et Agronomie" Department of INRA and the AIP DEMICHLORD (Etudes exploratoires de la dégradation microbienne de la chlordécone) of INRA for financial supports, and the FIRE (Fédération Ile-de-France de Recherche sur l'Environnement) for Ziang Li's grant. They are also grateful to Anaïs Labrunie and Sophie Vitrant for their contribution to this work.

Appendix A. Supplementary material

Supplementary Tables A1, A2

References

Ahlers, J., Stock, F., Werschkun, B., 2008. Integrated testing and intelligent assessment-new challenges under REACH. Environ. Sci. Pollut. Res. 15, 565-572.

ATSDR (Agency for Toxic Substances and Disease Registry), 1995. Toxicological profile for mirex and chlordecone. US Department of Health and Human Services, Public Health Services, http://www.atsdr.cdc.gov/toxprofiles/tp66.pdf

Boethling, R. S., Costanza, J., 2010. Domain of EPI suite biotransformation models. SAR QSAR Environ. Res. 5, 415- 443.

Cabidoche, Y.-M., Achard, R., Cattan, P., Clermont-Dauphin, C., Massat, F., Sansoulet, J., 2009. Long-term pollution by chlordecone of tropical volcanic soils in the French West Indies: A simple leaching model accounts for current residue. Environ. Pollut. 157, 1697-1705.

ChemOffice, 2009. ChemOffice Ultra 12.0 molecular modelling software, Cambridge Soft, Perkin Elmer.

ChemSpider, 2013. The free chemical database. http://www.chemspider.com/

Dearden, J. C., Schüürmann, G., 2003. Quantitative structure-property relationships for predicting Henry's law constant from molecular structure. Environ. Toxicol. Chem. 22, 1755-1770.

Dolfing, J., Novak, I., Archelas, A., Macarie, H., 2012. Gibbs free energy of formation of chlordecone and potential degradation products: implications for remediation strategies and environmental fate. Environ. Sci. Technol. 46, 8131-8139.

Doucette, W. J., 2003. Quantitative structure-activity relationships for predicting soil-sediment sorption coefficients for organic chemicals. Environ. Toxicol. Chem. 22, 1771-1788.

Dragon 5.5, 2007. Software for the calculation of molecular descriptors, Talete s.r.l. http://www.talete.mi.it/

Eriksson, L., Andersson, P., Johansson, E., Tysklind, M., 2006. Megavariate analysis of environmental QSAR data. Part I - A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). Mol. Diversity 10, 169-186.

Harju, M., Andersson, P. L., Haglund, P., Tysklind, M., 2002. Multivariate physicochemical characterisation and quantitative structure-property relationship modeling of polybrominated diphenyl ethers. Chemosphere 47, 375-384.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193-218.

Jackson, J. E. A., 1991. User's Guide to Principal Components, eds. John Wiley and Sons, New York.

Karelson, M., Lobanov, V. S., 1996. Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. Chem. Rev. 96, 1027-1043.

Katritzky, A.R., Maran, U., Lobanov, V.S., Karelson, M., 2000. Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties. J. Chem. Inf. Comput. Sci. 40, 1-18

Lohninger, H., 1994. Estimation of soil partition coefficients of pesticides from their chemical structure. Chemosphere 29, 1611-1626.

Mackay, D., McCarty, L. S., McLeod, M., 2001. On the validity of classifying chemicals for persistence, bioaccumulation, toxicity, and potential for long-range transport. Environ. Tox. Chem. 20, 1491-1498.

Mackay, D., Shiu, W. Y., Ma, K.-C., Lee, S. C., 2006. Handbook of physical-chemical properties and environmental fate for organic chemicals, second ed. CRC Press, Taylor and Francis Group, Boca Raton.

Marchand, A. P., 1989. Synthesis and chemistry of homocubanes, bishomocubanes and trishomocubanes. Chem. Rev. 89, 1011-1033.

Mishra, M., Potetz, B., Huan, J., 2011. Bayesian classifiers for chemical toxicity prediction. BIBM'11, Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine, 595-599.

Muir, D. C. G., Howard, P. H., 2006. Are there other persistent organic pollutants ? A challenge for environmental chemists. Environ. Sci. Technol. 40, 7157-7166.

Nguyen, X., Epps, J., Bailey, J., 2009. Information theoretic measures for clustering comparison: Is a correction for chance necessary? ICML'09: Proceedings of the 26th Annual International Conference on Machine Learning, San Francisco, 1073-1080.

OECD (Organisation for economic co-operation and development), 1993a. Application of structure-activity relationships to the estimation of properties important in exposure assessment. Environment monographs No 67, OECD, Paris.

OECD (Organisation for economic co-operation and development), 1993b. Structure-activity relationships for biodegradation. Environment monograph No 68, OECD, Paris.

Pavan, M., Netzeva, T. I., Worth, A. P., 2008. Review of literature-based quantitative structure-activity relationship models for bioconcentration. QSAR Comb. Sci. 27, 21-31.

PPDB (Pesticide properties database), 2013. http://sitem.herts.ac.uk/aeru/footprint/index2.htm

Rännar, S., Andersson, P. L., 2010. A novel approach using hierarchical clustering to select industrial chemicals for environmental impact assessment. J. Chem. Inf. Mod. 50, 30-36.

Rännar, S., Andersson, P. L., 2011. A multivariate chemical similarity approach to search for drugs of potential environmental concern. J. Chem. Inf. Mod. 51, 1788-1794.

Sabljic, A., 2001. QSAR models for estimating properties of persistent organic pollutants required in evaluation of their environmental fate and risk. Chemosphere 43, 363-375.

Saito, S., Ohno, K., Sese, J., Sugarawa, K., Sakuraba, H., 2010. Prediction of the clinical phenotype of Fabry disease based on protein sequential and structural information. J. Hum. Genet. 55, 175-178.

Snarey, M., Terrett, N. K., Willet, P., Wilton, D. J., 1997. Comparison of algorithms for dissimilarity-based compound selection. J. Mol. Graphics Modell. 15, 372-385.

Tenenhaus, M., 1998. La regression PLS, théorie et pratique, ed. Technip, Paris.

Todeschini, R., Gramatica, P., 1997. 3D-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors. Quant. Struct.-Act. Relat. 16, 113-119.

Vogt, M., Bajorath J., 2012. Chemoinformatics: a view of the field and current trends in method development. Bio. Med. Chem. 20, 5317-5323.

Walker, J. D., Carlsen, L., 2002. QSARs for identifying and prioritizing substances with persistence and bioconcentration potential. SAR QSAR Environ. Res. 13, 713-725.

Ward, J. H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236-244.

Wold, H., 1996. Estimation of principal component and related models by iterative least squares, in: Krishnaiaah, P.R. (Ed.), Multivariate Analysis, Academic Press, New York, pp. 391-420.

Yang, P., Chen. J., Chen, S., Yuan, X., Schramm, K.-W., Kettrup, A., 2003. QSPR models for physicochemical properties of polychlorinated diphenyl ethers. Sci. Tot. Environ. 305, 65-76.