Introduction of jumping fragments in combination with QSARs for the assessment of classification in ecotoxicology

Sylvain Lozano, Guillaume Poezevara, Marie-Pierre Halm, Elodie Lescot-Fontaine, Alban Lepailleur, Ryan Bissell-Siders, Bruno Crémilleux, Sylvain Rault, Bertrand Cuissart, Ronan Bureau

▶ **To cite this version:**

Sylvain Lozano, Guillaume Poezevara, Marie-Pierre Halm, Elodie Lescot-Fontaine, Alban Lepailleur, et al.. Introduction of jumping fragments in combination with QSARs for the assessment of classification in ecotoxicology. Journal of Chemical Information and Modeling, 2010, 50 (8), pp.1330-1339. hal-01011322

HAL Id: hal-01011322
https://hal.science/hal-01011322

Submitted on 23 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Introduction of Jumping Fragments in Combination with QSARs for the Assessment of Classification in Ecotoxicology

Sylvain Lozano,[†] Guillaume Poezevara,[‡] Marie-Pierre Halm-Lemeille,[†] Elodie Lescot-Fontaine,[†] Alban Lepailleur,[†] Ryan Bissell-Siders,[‡] Bruno Crémilleux,[‡] Sylvain Rault,[†] Bertrand Cuissart,[‡] and Ronan Bureau*[,†]

Centre d'Etudes et de Recherche sur le Médicament de Normandie, UPRES EA-4258, FR CNRS INC3M, Université de Caen Basse-Normandie, UFR des Sciences Pharmaceutiques, Boulevard Becquerel, 14032 Caen Cedex, France and Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen, CNRS UMR 6072, Campus Côte de Nacre, 14032 Caen Cedex, France

Starting from a random set of structures taken from the European Chemical Bureau (ECB) Web site, an estimation of the classification by acute category in ecotoxicology was carried out. This estimation was based on two approaches. One approach consists in starting with global quantitative structure−activity relationship (QSAR) equations, analyzing the results and defining an interpretation in terms of overall results and mode of action. The other starts with the notion of emerging fragments and more specifically with the introduction of a particular concept: the jumping fragments. This publication studies the scopes and limitations of each approach for the classification of the derivatives. A promising combination of the two methods is proposed for the classification and also for bringing new information about the importance, for the ecotoxicity, of specific chemical fragments considered alone or in association with others.

INTRODUCTION

The first and most essential step for a safe use of chemicals is to know their identities and hazards, information leading to the implementation of the appropriate protective measures. Classification and labeling (CL) of chemicals involves an evaluation of the intrinsic hazard of substances and mixtures.[1] This evaluation must be made for any chemical manufactured within or imported into the European Union (EU). CL is based on physical, health, and environmental hazards. To assess the hazards for the aquatic environment, acute or chronic aquatic toxicity data are required. The core part of the harmonized classification system for environmental hazards consists of three acute and three chronic classification categories (GHS).[1,2] Acute aquatic toxicity means the intrinsic property of a material to cause injury to an aquatic organism in a short-term exposure. Three endpoints are determined using a fish 96 h LC50, a crustacean species 48 h EC50, and/or an algal species 72 or 96 h EC50 (OECD test guideline 201-202-203). In function of the data associated to these endpoints, chemicals are classified[2] with a hazard statement codes (HSC) in acute categories 1−3: H400 (very toxic, L(E)C50 $\leq$ 1 mg/L), H401 (toxic, 1 mg/L < L(E)C50 $\leq$ 10 mg/L), and H402 (harmful, 10 mg/L < L(E)C50 $\leq$ 100 mg/L). H400 may be subdivided to include a lower band at L(E)C50 $\leq$ 0.1 mg/L.

Traditionally, such information has arisen from the use of in vivo animal testing but under REACH[3,4] legislation (Registration, Evaluation, Authorization of Chemicals), QSAR models are expected to play a significant role. In fact, for reasons of resources and animal welfare, it is important to reduce the number of tests where it is significantly justifiable. Category approaches and QSARs are alternatives that can be used to save resources and accelerate hazard and risk assessments. A considerable number of QSARs for individual classes of chemicals are available and give correct results for industrial chemicals.[5] For ecotoxicological endpoints, studies on QSAR are numerous[6,7] but only a few researches have concerned the prediction of classification categories. Two studies[8,9] described the impact of ecotoxicity endpoints in European notification procedure on CL for the aquatic environment. Concerning HSC data in the databases, no HSC for a chemical, in relation with the hazardous for aquatic environment, results either from no biological effect (experimentations described in this case) or from missing ecotoxicological tests. This is the main point concerning the difficulties associated to the consideration of a chemical set formed by nontoxic compounds in our study (no HSC assigned). This article is focused on H400 and H402 HSC (vide infra for an explanation). Starting from a heterogeneous set of derivatives, several strategies are possible to estimate their toxicities by QSAR equations. One of the major approaches is divided into two steps. First, the substance is affiliated to a previously defined class, and second, a QSAR equation that is specific for this class is applied. Among the classification schemes, we can mention the ECOSAR classification[10] which is used to predict the aquatic toxicity of chemicals based on their similarity of structure to chemicals for which the aquatic toxicity has been previously measured. To date, the ECOSAR package includes more than 150 QSARs for more than 50 chemical classes. However, it has been realized in the last decade that an appreciation of the

* Corresponding author. E-mail: ronan.bureau@unicaen.fr. Telephone: (33)2-31-56-68-20.
[†] CERMN, Université de Caen Basse-Normandie.
[‡] Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen.

mode of action (MOA), which describes the effect of a toxicant at the organism level, is crucial for the development of reliable QSARs. As a consequence, a classification scheme which retrieves information on the possible MOA of a target chemical is more relevant for subsequently identifying its potential analogues. Yet, by considering the three ecotoxicological aquatic endpoints, the definition of a possible MOA by compounds is not trivial. Several softwares, like Toxtree[11] or QSAR toolbox,[12] give some information about a possible MOA. A previous study was carried out in our group[13] starting from the EPAFHM[14] set (toxicity data for fish) for which the MOA was indicated with a probability associated to each MOA. Unfortunately, the same type of data set is not available for daphnia and algae. A second approach starts from global QSAR equations. These equations are mainly based on the integration of parameters in relation with hydrophobic, steric, or electronic characteristics of chemicals. They are designed to estimate the toxicity in relation with a nonspecific MOA, like baseline, polar, and ester narcoses and some classes of reactive chemicals (the toxicity of high-reactive chemicals is underestimated). In these global equations, polar and nonpolar compounds are handled in the same models, and they are able to give at least the baseline toxicity for chemicals. However, derivatives with a specific MOA are not integrated in these equations. Starting from a random set of derivatives, this point will be discussed in this study.

The second point of this study concerns the potential application of jumping fragments for CL. Emerging fragments have been introduced in chemoinformatics in 2006[15] in the context of molecular and compound classification. Their applications concern the extraction of key molecular features from very few known active compounds and classify molecules according to different potency levels. In our case, we have started from data sets representing an important diversity in terms of structures and biological profiles (potential MOA), and we introduce the notion of jumping fragments in this field. A recent data mining algorithm from our laboratory[16] was experimented enabling an automatic extraction of substructures which appear frequently in one class (H400) and never appear in another class (H402). In this publication, such a substructure is named a jumping fragment. The overall objective of this study is to analyze the potential of these approaches starting from a random set of derivatives for the estimation of their potential classifications.

MATERIALS AND METHODS

QSAR Models. The 96 h acute toxicity (mol/L) to fish was estimated from a global QSAR model[13,17] (eq 1). The equation was refined starting from a referential data set concerning fish acute toxicity (EPAFHM).[14]

$$\log LC_{50} = -0.509 \log P_{OW} - 0.005MW + 0.067E_{LUMO} - 1.977 \quad (1)$$

$$n = 566,\ r^2 = 0.65,\ s = 0.81$$

The 48 h acute toxicity (mol/L) to daphnia (eq 2) was estimated from a global QSAR model.[18]

$$\log EC_{50} = -0.57 \log P_{OW} + 0.45E_{LUMO} - 2.44 \quad (2)$$

$$n = 61,\ r^2 = 0.54,\ s = 0.71$$

The 72 h acute toxicity (mol/L) to algae (eq 3) was estimated from a baseline narcosis model.[19] Until now, to our knowledge, no global equation was defined for algae.[20] With a MOA as a narcotic estimated to be associated to more than 50% of organic chemicals, we have chosen to consider this equation:

$$\log EC_{50r} = -1.00 \log P_{OW} - 1.23 \quad (3)$$

$$n = 10,\ r^2 = 0.93,\ s = 0.17$$

Quality of the Prediction and Classification. From s values (see eqs 1−3) associated to each QSAR equation, the residuals can lead to a misclassification from one acute classification category to the nearest acute classification category. We have chosen to consider, as a predictive error, the derivatives with an estimated toxicity higher than two intervals from the real class (H400 instead of H402). As a consequence, in the initial data set, all the derivatives with H401 HSC were discarded.

Calculation of Descriptors. The log P_{OW} values were calculated by KOWWIN[21] in agreement with data in relation with eq 1 (EPAFHM data set).[14] Three-dimensional (3D) atomic coordinates were generated, and energy minimizations were carried out (clean force field).[22] E_{LUMO} (lowest unoccupied molecular orbital energy) values were calculated for each chemical using VAMP[23] and AM1 for Hamiltonian.

Data Set. An initial data set of 72 563 chemicals (IUCLID,[24] version 4) has been retrieved from the ECB Web site[25] and recorded in a structure data file (SDF) format. For each chemical, the data are associated with CAS and EINECS numbers and an IUPAC chemical name. A PERL script has extracted, for each chemical (EINECS number), the attributed R-phrases leading to 933 chemicals with R50/R51/R52 HSC (annex I of directive 67/548/EEC). In the new harmonized classification, R50 corresponds to H400, R51 to H401, and R52 to H402. On this set, all nonorganic chemicals were removed (metals, organometallics, etc.). The structures with an H401 HSC were also discarded, salts were cleaned up, and a filter was applied on molecular weight ($30 < MW < 490$ g/mol) and on log P_{OW} values ($1 < \log P_{OW} < 8$) in order to keep only the chemicals respecting the validity domain common to the three QSAR equations. These selections finally led to 436 out of 933 derivatives. Hereafter, this set was referred as the ECB_H400H402 data set. This data set was separated in two sets, one representing 372 derivatives with a H400 HSC and a second representing 64 derivatives with a H402 HSC. Classifications using jumping fragments were conducted on the whole ECB_H400H402 data set. For QSAR models, LUMO values could not be calculated for three chemicals leading to 433 chemicals selected.

Jumping Fragments. We assume here that the level of toxicity for a chemical may be influenced by the presence of a specific fragment. Such a fragment may have a strong foothold in the toxic chemicals and may be missing from the nontoxic chemicals. We recently have designed an algorithm that automatically extracts such fragments.

A *fragment* denominates a connected part of a chemical structure containing at least one chemical bond. Given a set

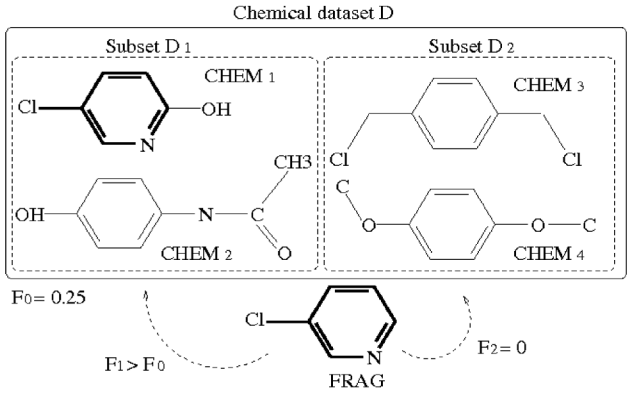

Figure 1. The notion of jumping fragment. FRAG is a jumping fragment.

of molecules D, a fragment is *frequent* in D if its frequency of occurrence exceeds a given frequency threshold. Let D be partitioned into two subsets D_1 and D_2. A *jumping fragment* from D_2 to D_1 is a frequent fragment in D_1 that never appears in D_2 (see Figure 1). A molecular structure is depicted by a labeled graph, its molecular graph. The extraction of the jumping fragments from a set of molecules is a demanding problem, computationally speaking. The number of fragments to take into account is huge, and the search deals with two nonpolynomial graph problems[26] (graph and subgraph isomorphism). We specifically consider here the jumping fragments from the H402 chemicals to the H400 chemicals; they correspond to the fragments that are frequent within the H400 structures and that never appear within the H402 ones. In this context, a jumping fragment should be an indicator of the level of toxicity for a molecule.

The notion of jumping fragments relies on the notion of (jumping) emerging patterns[27] introduced recently in chemoinformatics;[15] it has been named Emerging Chemical Patterns (ECPs). Auer and Bajorath have extracted ECPs and have used them to conduct several informative experimental studies.[28,29] For these studies, molecules were described using discretized descriptors from the Molecular Operating Environment[30] (MOE). Consequently, an ECP corresponds to a set of MOE descriptors. This differs from a jumping fragment which is directly extracted from the 2D structures of the molecules. We recently have designed the first method to extract the frequent emerging graph patterns.[16] Extraction of the jumping fragments consists in a specialization of this method. Previously, two methods have been proposed to extract other patterns but which can be specialized to extract jumping fragments. Borgelt and Berthold have introduced the notion of *discriminative fragment*.[31] Given two frequency thresholds, f_{D1} the frequency of the fragment in D_1 and f_{D2} the frequency of the fragment in D_2, a fragment is discriminative from D_2 to D_1 if its frequency in D_1 exceeds f_{D1} and its frequency in D_2 is below f_{D2}. A discriminative fragment with f_{D2} set at zero is a jumping fragment. Ting and Bailey have introduced the notion of *contrast subgraph*.[32] A graph is a contrast subgraph from D_2 to D_1 if it occurs as a subgraph in D_1 but never in D_2. A contrast subgraph which is connected and so frequent as to be statistically significant is a jumping fragment. Although the notion of contrast subgraph is very interesting, it requires substantial computation time. To the best of our knowledge, the calculation of the contrast subgraphs from D_2 to D_1 is not performed more efficiently than by considering each graph in D_1 in turn and mining the

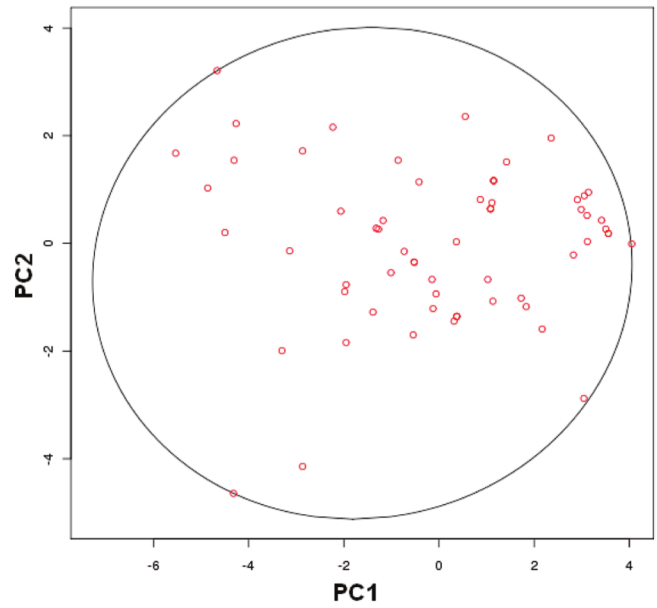

Figure 2. PCA analysis on the 64 H402 chemicals.

subgraphs of a graph with more than 20 vertices is a significant challenge. The jumping fragments can be extracted with either of the former methods. The following experiments assess the jumping fragments as descriptors for ecotoxicity studies.

Ecotoxicity Experiment using Jumping Fragments. The input of this study is the ECB_H400H402 data set partitioned into two subsets: 372 molecules in H400 class and 64 molecules in H402 class. In order to enable a five-fold cross-validation scheme, each of these subsets is randomly split into five samples of equal size. The five-fold cross-validation scheme allows us to use 80% of the data set as training data. The resulting classification model leaves a large testing sample on which we can measure the model's average performance, and it is almost as powerful as the full model which is trained on 100% of the data set.[33] The five-fold cross-validation is performed by reserving each of the five parts of the data set, in turn, to play the role of the testing sample, and by extracting jumping fragments, using the remaining four parts of the data as the learning sample. We perform a five-fold cross-validation scheme and not a leave-one-out scheme. This choice enables the study of the variability between the sets of jumping fragments extracted from the different folds; this variability is measured in Table 5.

Constitution of the Data Set. With only 64 H402 structures, the data set is not well-balanced. Each fold of the cross-validation study contains less than 55 H402 molecules in its learning set. This small number of H402 molecules may lead to the extraction of unjustified jumping fragments which are then treated as toxicophores; when they are tested on test data, they are not predictive. Consequently, we have studied the diversity of this set of 64 H402 molecules with a principle component analysis (PCA) (see Figure 2). As attributes, we have chosen eleven descriptors: Alog *P*; molecular weight (MW); numbers of: H donors, H acceptors, rotatable bonds, atoms, rings, and aromatic rings; molecular solubility; molecular surface area; and molecular polar surface area. The first two principal components explain 75% of the variance in the values of these descriptors on

Table 1. Results of Classification by Acute Categories for the ECB_H400H402 Data Set

real classification	predicted classification				failure rate
	H400	H401	H402	no classification	
H400	**176**	122	70	1	19%
H402	3	18	**43**	0	5%

the H402 molecules. The data set ECB_H400H402 clearly lacks a sufficiently diverse set of less toxic molecules in which to differentiate the toxic molecules. So, we have added to this initial training set 126 new structures with H402 classification drawn from our internal database.[17] The resulting data set has the same 372 H400 structures, 192 instead of 64 H402 structures.

Within each fold of the cross-validation experiment, the data set is arbitrarily divided into learning (80%) and test (20%) data. A confidence level for the χ^2 test is decided. This level determines a minimum frequency threshold of statistical significance within training data. The jumping fragments which are present with at least this frequency in the H400 molecules are extracted. Each molecule of the test sample is then described only by the jumping fragments which it contains. An illustrative decision rule is finally applied on the testing set. As an illustration of how to balance the confidence with which jumping fragments individually imply H400 and their coverage of H400, we study a *decision rule*: "a molecule is toxic just in case it contains a jumping fragment".

For each fold of the cross-validation process, the following measures are determined: (i) the number of jumping fragments and their distribution by size, (ii) the percentage of jumping fragments which generalize from the learning set to the test set, (iii) the redundancy of the jumping fragments extracted in this and other folds, (iv) the degree to which the jumping fragments cover the H400 molecules; this is the proportion of H400 molecules that contain at least one jumping fragment, and (v) the success rate of the illustrative decision rule applied to the H402 molecules. The set of discovered jumping fragments is memorized in order to study the variation in the set of jumping fragments from one fold to another and to determine a set of jumping fragments, which is always extracted.

RESULTS AND DISCUSSION

Classifications using QSAR Models. The results for the predictions of the classifications are presented in Table 1. For the H400 compounds, a failure rate of 19% was obtained (H402 or no HSC instead of H400). 176 derivatives are correctly classified (48%), 122 derivatives have intermediate classification (33%, H401) and 71 derivatives have a wrong prediction (H402) and particularly for one derivative not classified. The failure rate is really lower for compounds classified as H402 (5%) with only three derivatives having a bad prediction (H400 instead of H402).

Table 2 represents the data corresponding to the 70 structures predicted H402 instead of H400. In the same class (H400), the structure predicted without classification (overall predicted acute ecotoxicity values >100 mg/L) corresponds to methyl bromide (compound **1**, see Figure 3). Used as a pesticide, **1** has a potential reactivity to macromolecules, but the MOA is still not understood. For the descriptors, the calculated log P_{ow} of **1** is correct (equal to its experimental value, log $P = 1.19$). The LUMO value is slightly positive (AM1 for Hamiltonian). By considering PM3 instead of AM1, the LUMO value decreases leading to a H402 classification. To launch an analysis on the remaining 70 structures, a clustering was first carried out with FCFP4[23] as fingerprints (functional fingerprints), Tanimoto metric for the distance between records and for the clustering method, and a relocation method based on maximal dissimilarity partitioning.[33] From the clustering of the 70 derivatives (see clusters in Table 2), only the centroids (cluster centers, **2**−**16**) are represented in Figure 3. The analysis of the clusters has shown rapidly that some chemical groups are related to specific biological activities. For instance, cluster 1 (**2** for the centroid) corresponds to triazines with herbicide activity (3 out of 4 compounds). Cluster 2 (**3** for the centroid) corresponds to derivatives with aniline functions. Cluster 11 (19 chemicals, **12** for the centroid) corresponds to carbamate and phenyl urea derivatives (insecticides, fungicides, and herbicides). In fact on the overall set, 46 compounds out of 70 correspond to pesticides (64%, clusters 1, 4, 7−9, 11, and 13). The general QSAR models are unable to integrate toxicity data resulting from specific interactions with some receptors. So, the results for pesticides are logical. For instance, 3-(3-chloro-4-methylphenyl)-1,1-dimethylurea (phenyl urea herbicide) has specific interactions with a protein (D1 protein) involved in photosynthesis. This interaction leads to a high toxicity for algae. The predictions were good for fish (15 mg/L compared to 20 mg/L) and daphnia (43 mg/L compared to 67 mg/L)[34] and incorrect for algae (0.024 mg/L instead 48 mg/L). The other data (24 nonpesticides out of 70 derivatives) concern mainly derivatives with aniline functions (13 structures). These derivatives are classified as polar narcotics[35] and must be correctly estimated by the models. Indeed, it is globally the case for fish toxicities, like 4-methylaniline with a predicted value of 81 mg/L (real value of 115 to 171 mg/L in function of the species),[36] *m*-toluidine with a predicted value of 79 mg/L (real value of 34 mg/L),[36] *o*-toluidine with a predicted value of 87 mg/L (real value between 68 and 100 mg/L),[36] and 3,4-dichloroaniline with a predicted value of 13 mg/L (real value between 2.4 to 13 mg/L).[37] The H400 HSC for aniline derivatives comes from an important toxicity for daphnia, not understood by only a polar narcosis as MOA. One study from Ramos et al.[38] has discussed this point, but the mechanisms behind the high sensitivity of daphnia to aromatic amines remains unclear.

For the compounds classified H402, three compounds have a wrong classification (see Figure 4). For carbon tetrachloride **17**, the recent data for *Chlamydomonas reinhardtii*[39] has shown, in fact, a greater sensitivity for algae with an EC_{50r} of 0.246 mg/L and an EC_{10} of 0.072 mg/L. These data, not present in the IUCLID4[24] file, are considered valid, and so the acute classification category is really H400. For the compound **18**, the overall result is curious. Data from the ECB[25] site showed no information concerning ecotoxicity data on fish, daphnia, and algae. This compound is sensible to moisture, and all isocyanates react with water to form insoluble urea derivatives. So, this compound is really an outlier, and we do not understand the H402 classification

Table 2. Description of the 70 Derivatives with H402 Classification Instead of H400[a]

CAS number	IUPAC name	cluster	
122-34-9	6-chloro-N2-N4-diethyl-1,3,5-triazine-2,4-diamine	1	P
1014-69-3	N2-isopropyl-N4-methyl-6-(methylthio)-1,3,5-triazine-2,4-diamine	1	P
1912-24-9	6-chloro-N2-ethyl-N4-isopropyl-1,3,5-triazine-2,4-diamine	1	P
2095-02-5	2,4-diethyl-6-methylbenzene-1,3-diamine	1	C
21725-46-2	2-(4-chloro-6-(ethylamino)-1,3,5-triazin-2-ylamino)-2-methylpropanenitrile	1	P
95-53-4	*o*-toluidine	2	C
95-76-1	3,4-dichloroaniline	2	C
100-61-8	*N*-methylaniline	2	C
100-63-0	phenylhydrazine	2	C
106-47-8	4-chloroaniline	2	C
106-49-0	4-methylaniline	2	C
108-44-1	*m*-toluidine	2	C
532-82-1	4-[(*Z*)-phenyldiazenyl]benzene-1,3-diamine hydrochloride	2	
2051-79-8	N1,N1-diethyl-3-methylbenzene-1,4-diamine	2	C
36341-27-2	biphenyl-4,4′-diamine acetate	2	C
68479-98-1	3-(pentan-3-yl)benzene-1,2-diamine	2	C
107-05-1	3-chloroprop-1-ene	3	
764-41-0	(*E*)-1,4-dichlorobut-2-ene	3	C
10061-01-5	(1*Z*)-1,3-dichloroprop-1-ene	3	C
2497-07-6	*O*,*O*-diethyl S-2-(ethylsulfinyl)ethyl phosphorodithioate	4	P
23560-59-0	7-chlorobicyclo[3,2,0]hepta-2,6-dien-6-yl dimethyl phosphate	4	P
30864-28-9	methyl (2*E*)-3-[(dimethoxyphosphorothioyl)oxy]-2-methylprop-2-enoate	4	P
700-13-0	2,3,5-trimethylbenzene-1,4-diol	5	P
1570-64-5	4-chloro-2-methylphenol	5	
2095-01-4	4,6-diethyl-2-methylbenzene-1,3-diamine	5	C
533-74-4	3,5-dimethyl-1,3,5-thiadiazinane-2-thione	6	P
150-68-5	3-(4-chlorophenyl)-1,1-dimethylurea	7	P
1746-81-2	3-(4-chlorophenyl)-1-methoxy-1-methylurea	7	P
2782-57-2	1,3-dichloro-1,3,5-triazinane-2,4,6-trione	7	P
62-73-7	2,2-dichlorovinyl dimethyl phosphate	8	P
300-76-5	1,2-dibromo-2,2-dichloroethyl dimethyl phosphate	8	P
501-53-1	benzyl carbonochloridate	8	C
771-29-9	1-hydroperoxy-1,2,3,4-tetrahydronaphthalene	8	C
1918-16-7	2-chloro-*N*-isopropyl-*N*-phenylacetamide	8	P
50563-36-5	2-chloro-*N*-(2,6-dimethylphenyl)-*N*-(2-methoxyethyl)acetamide	8	P
148-79-8	4-(1H-benzo[d]imidazol-2-yl)thiazole	9	P
3878-19-1	2-(furan-2-yl)-1H-benzo[d]imidazole	9	P
10605-21-7	methyl 1H-benzo[d]imidazol-2-ylcarbamate	9	P
96-05-9	allyl methacrylate	10	
97-86-9	isobutyl methacrylate	10	
21087-64-9	4-amino-6-tert-butyl-3-(methylthio)-1,2,4-triazin-5(4H)-one	10	P
63-25-2	naphthalen-1-yl methylcarbamate	11	P
114-26-1	2-isopropoxyphenyl methylcarbamate	11	P
116-06-3	(*E*)-2-methyl-2-(methylthio)propanal O-methylcarbamoyl oxime	11	P
149-30-4	benzo[d]thiazole-2-thiol	11	P
315-18-4	4-(dimethylamino)-3,5-dimethylphenyl methylcarbamate	11	P
1563-66-2	2,2-dimethyl-2,3-dihydrobenzofuran-7-yl methylcarbamate	11	P
2032-59-9	4-(dimethylamino)-3-methylphenyl methylcarbamate	11	P
2425-10-7	3,4-dimethylphenyl methylcarbamate	11	P
2631-40-5	2-isopropylphenyl methylcarbamate	11	P
3766-81-2	2-*s*-butylphenyl methylcarbamate	11	P
15545-48-9	3-(3-chloro-4-methylphenyl)-1,1-dimethylurea	11	P
17804-35-2	methyl 1-(butylcarbamoyl)-1H-benzo[d]imidazol-2-ylcarbamate	11	P
18691-97-9	1-(benzo[d]thiazol-2-yl)-1,3-dimethylurea	11	P
19937-59-8	3-(3-chloro-4-methoxyphenyl)-1,1-dimethylurea	11	P
23564-05-8	methyl *N*-[2-([(methoxycarbonyl)amino] methanethioylamino)phenyl]carbamothioylcarbamate	11	P
29973-13-5	2-(ethylthiomethyl)phenyl methylcarbamate	11	P
34014-18-1	1-(5-tert-butyl-1,3,4-thiadiazol-2-yl)-1,3-dimethylurea	11	P
34681-10-2	(*E*)-3-(methylthio)butan-2-one O-methylcarbamoyl oxime	11	P
75-08-1	ethanethiol	12	C
650-51-1	2,2,2-trichloroacetate	12	C
23103-98-2	2-(dimethylamino)-5,6-dimethylpyrimidin-4-yl dimethylcarbamate	13	P
25366-23-8	1,3-dimethyl-1-(5-(trifluoromethyl)-1,3,4-thiadiazol-2-yl)urea	13	P
51235-04-2	3-cyclohexyl-6-(dimethylamino)-1-methyl-1,3,5-triazine-2,4(1H,3H)-dione	13	P
69581-33-5	*N*-(3-chlorophenyl)-*N*-(2-oxotetrahydrofuran-3-yl)cyclopropanecarboxamide	13	P
95-69-2	4-chloro-2-methylaniline	14	C
2312-76-7	sodium 2-methyl-4,6-dinitrophenolate	14	P
7580-31-6	2-ethylhexanoic acid	15	
26530-20-1	2-octylisothiazol-3(2H)-one	15	P

[a] P is pesticides, and C is CMR.

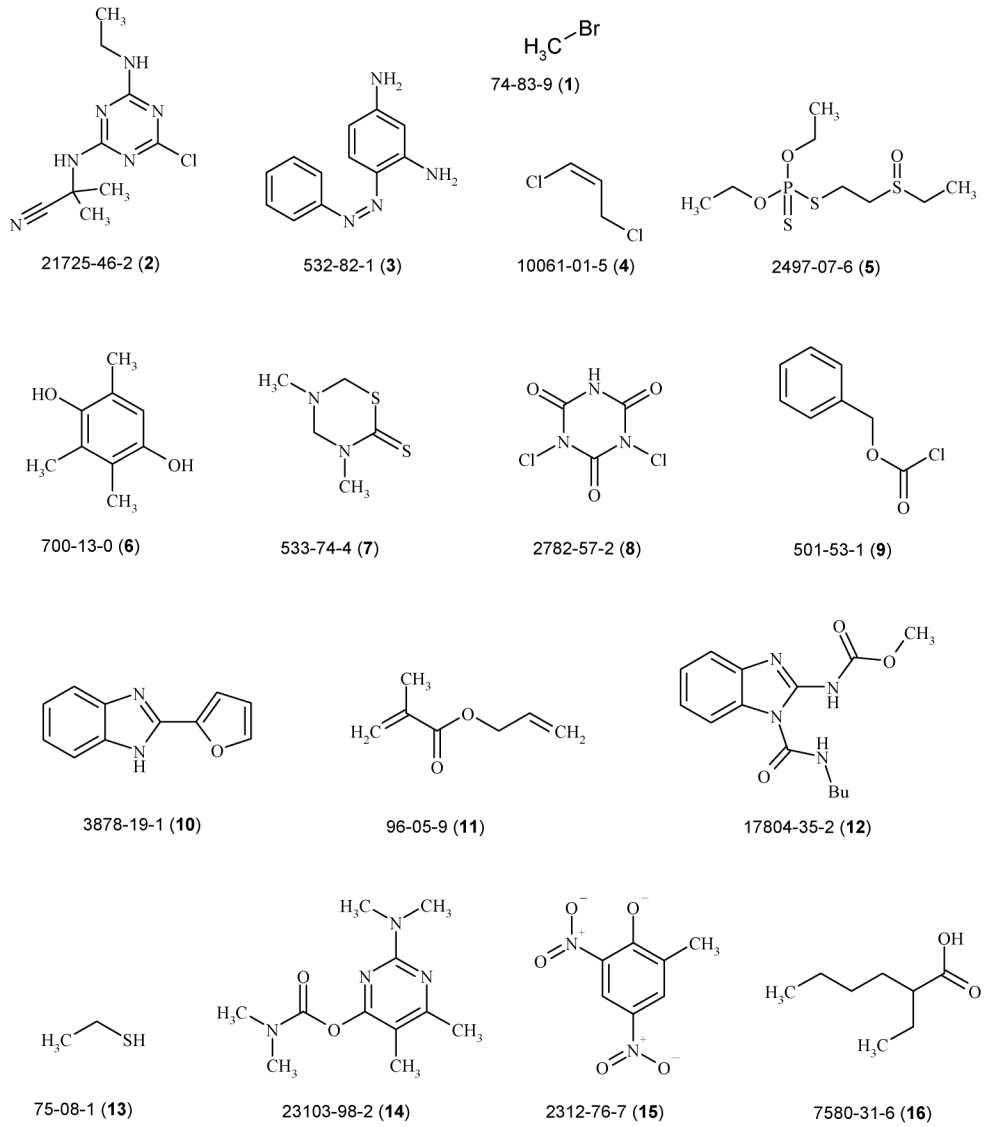

Figure 3. Representation of the centroids of each cluster for the 70 derivatives classified H402 instead of H400. The derivative 1 corresponds to the only compound without classification for the prediction.

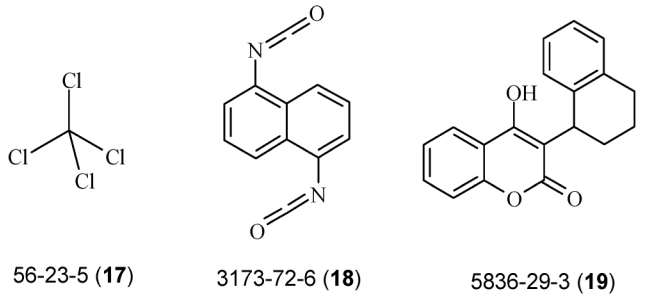

Figure 4. Structures associated to the derivatives estimated H400 instead of H402 for the classification.

starting straight from its structure (degradation/biodegradation?). The compound **19** is a pesticide with the respective values for fish (53 mg/L), *Daphnia magna* (>14 mg/L) and algae (>18 mg/L). This is well described in a recent report[40] of EU for the biocides.

Generalization of the Jumping Fragments. We have extracted those jumping fragments which occur sufficiently in D_1 to refute the hypothesis that the molecules containing the fragment are drawn independently from both classes. Consequently, a χ^2 test of this independent hypothesis has been used to select statistically significant jumping fragments. The only free parameter in this method is the level of confidence of the χ^2test, which determines a minimum frequency threshold in D_1 for the jumping fragments. Under this scheme, a jumping fragment satisfies the following three properties:

(i) It occurs in D_1.
(ii) It does not occur in D_2.
(iii) It is significantly more present in D_1 than in D_2.

In our experiment in which D_1 is the class of H400 molecules and D_2 is the class of H402 molecules and the size of one fold in the cross-validation scheme is a fifth of the data set, the minimum frequency thresholds fixed by the χ^2 test are 2.6% (8 molecules) and 4.3% (13 molecules) for the confidence thresholds of 95 and 99%, respectively.

In Table 3, the number of jumping fragments and their average distribution by size (over the five folds) are reported for both confidence thresholds of 95 and 99%. We put no constraint on the size of a fragment other than conditions i−iii. We do not eliminate small or large fragments, except to the extent that, in practice, large fragments fail condition iii and small fragments fail condition ii. As the confidence

Table 3. Number of Jumping Fragments and Their Average Distribution by Size over the Five Folds

	average number	
fragment size	95%	99%
2	2.2	0.8
3	6.8	2.2
4	21.4	9.8
5	53.8	18.6
6	96.2	30.6
7	163.4	44.8
8	275.8	63
9	383.8	59.4
10	397.6	30.2
11	322.2	10
12	205.2	1.4
13	93.2	0
14	22.8	0
15	3.4	0

threshold is relaxed, the frequency threshold decreases, and larger fragments with more atoms are extracted. We now restrict our attention to a confidence level of 99% to perform a study of the generalization of properties ii and iii from training to test data. We choose 99% as the highest value, which allows a sufficiently large sample of jumping fragments to be extracted.

We have assessed the extent to which the size of the jumping fragments might correlate with the rate to which the fragment's jumping property ii and property of statistical significance iii generalize to the test sample. The results are summarized in Table 4. The jumping property ii and the statistical property iii are statistically equivalent for the smallest, small, medium, and largest quartiles of the jumping fragments, with some preference for being satisfied by small fragments.

We have assessed whether in passing from the (four folds of the) training to (one fold of the) test data the three properties i−iii of a jumping fragment generalize. We have considered properties i and ii together as the jumping properties and iii as the property of statistical significance. Results are displayed in Table 4. Property ii generalizes between 60.31% of extracted fragments for fold 1 to 98.71% for fold 4, with an average of 82.42%. Property iii remains true in fold 1 for 80.62% of the fragments extracted from the other four folds and remains true in each of folds 3 and 5 for 100% of the fragments extracted from the other four folds, with an average of 93.21%. Both the jumping, ii, and statistical, iii, properties of jumping fragments generalize well from the learning to testing set, for each fold of the cross-validation process. See also Table 8 for a view of some jumping fragments, such as Clc(c(ccCl)Cl)c, which are clearly associated to ecotoxicological properties. Thus, to find a toxicity model on jumping fragments is statistically justified.

Table 5. Percentage of Common Jumping Fragments between Each Fold of the Cross-Validation Process

fold	1	2	3	4	5
1	−	39.13	45.05	44.94	48.34
2	39.13	−	63.94	68.94	61.15
3	45.05	63.94	−	68.17	67.18
4	44.94	68.64	68.17	−	68.28
5	48.34	61.15	67.18	68.28	−

Use of the Jumping Fragments. We have considered each pair of folds of the cross-validation experiment, in turn, to examine whether they extract the same set of jumping fragments. The redundancy in the extracted jumping fragments between each pair of folds of the cross-validation process is shown in Table 5. For example, folds 1 and 2 have extracted generally different jumping fragments with only 39% redundancy, the lowest observed. The highest redundancy, 69%, was observed between folds 2 and 4. For each pair of folds, this 39−69% represents a large set of

Table 4. Generalization of Jumping Fragments from the Learning to Testing Set with a Confidence Threshold of 99%[a]

fold	measure	smallest	small	medium	largest	all
1	size (min, max, average)	3, 7, 5.7	7, 8, 7.7	8, 9, 8.8	9, 12, 10.2	3, 12, 8.1
	jumping property ii (%)	80.25	65.43	64.2	31.71	60.31
	statistical significance (%)	91.36	82.72	66.67	81.71	80.6
	failure on both (%)	8.64	12.35	19.75	10.98	12.92
2	size (min, max, average)	2, 6, 5.1	6, 8, 7.2	8, 9, 8.4	9, 12, 10.3	2, 12, 7.7
	jumping property ii (%)	85.48	90.48	88.89	79.37	86.06
	statistical significance (%)	95.16	93.65	90.48	79.37	89.64
	failure on both (%)	4.84	6.35	9.52	20.63	10.36
3	size (min, max, average)	2, 6, 5.1	6, 8, 7.1	8, 9, 8.2	9, 12, 9.8	2, 12, 7.5
	jumping property ii (%)	76.39	76.71	90.28	98.63	85.52
	statistical significance (%)	100	100	100	100	100
	failure on both (%)	0	0	0	0	0
4	size (min, max, average)	2, 7, 5.2	7, 8, 7.3	8, 9, 8.4	9, 12, 9.9	2, 12, 7.7
	jumping property ii (%)	100	100	98.28	96.61	98.71
	statistical significance (%)	100	100	98.28	96.61	98.71
	failure on both (%)	0	0	1.72	1.69	0.86
5	size (min, max, average)	2, 7, 5.4	7, 8, 7.3	8, 9, 8.3	9, 11, 9.8	2, 11, 7.6
	jumping property ii (%)	92.06	90.63	79.69	92.19	88.63
	statistical significance (%)	100	100	100	100	100
	failure on both (%)	0	0	0	0	0
average	size (min, max, average)	2.2, 6.6, 5.3	6.6, 8, 7.3	8, 9, 8.4	9, 11.8, 10	2.2, 11.8, 7.7
	jumping property ii (%)	86.01	83.19	83.14	77.42	82.42
	statistical significance (%)	97.02	94.69	89.94	91.2	93.21
	failure on both (%)	2.98	4.13	6.8	6.74	5.17

[a] Partitioned according to the size of the fragment.

Table 6. Success Rate of the Prediction Rule on the H402 Molecules from the Testing Set

	frequency threshold (%)			
fold	1	2.6	4.3	10
1	53	77	93	100
2	64	80	95	98
3	49	77	92	95
4	64	90	97	100
5	49	82	95	100
average	55.8	81	94.3	98.6

fragments which is extracted by both folds. Some fragments are statistically significant in one fold but not in the other. Finally, the fragments of one fold which did not generalize property ii are not extracted by the other fold. There is a balance between the influence of properties ii and iii on the difference between two folds. The jumping fragments extracted in every single fold of the experiment represent 18.51% of the fragments extracted in any fold. These fragments are candidate *toxicophores*—substructures which cause toxicity—to be presented to experts or used in a decision process.

Classification Model using Jumping Fragments. Tables 6 and 7 report the success of the jumping fragments in the prediction of risk. This is the raw performance of the jumping fragments, without the benefit of a classification model. For a minimum frequency threshold of 10%, almost no jumping fragment are extracted, so that no chemical is predicted to carry a risk of H400. With a frequency threshold of 4.3%, the predictive value of the extracted fragments is very good. Hundreds of jumping fragments are extracted (Table 8 lists those extracted with a frequency at least of 5%), and each jumping fragment is almost completely absent in the H402 chemicals of the testing set. With a frequency threshold of 2.6%, thousands of jumping fragments are extracted from the learning set, and they are by and large absent in the H402 chemicals. The poor result with the threshold of 1% shows clearly that jumping fragments with weak frequency in the H400 molecule are not, one by one, sufficient evidence to rule out the hypothesis that a molecule containing the fragment is H402. Taken as a whole, these results show that the minimum frequency threshold we set for the extraction of jumping fragments in the learning set has an important influence of the value of each jumping fragment as a prediction rule. Setting a low threshold leads to the extraction of many fragments, which cover well the H400 chemicals but none of which is confidently absent in the H402 chemicals. Setting a high threshold leads to the extraction of fragments, each of which is strong evidence against the hypothesis that a molecule is H402. These fragments poorly cover the H400 chemicals, as is shown in Table 7, and so they do not explain the H400 risk.

Evaluation of an Illustrative Decision Rule. We have considered a wide range of frequency thresholds—varying from 5 to 0.6%—with which to parametrize the simplest possible decision rule based on jumping fragments; a molecule is H400 just in case it contains a jumping fragment. For every frequency threshold and every fold of the cross-validation experiment, the jumping fragments have been extracted, and the corresponding decision rules have been assessed (Table 7). Consequently, a dynamic view on the behavior of the decision rule has been obtained. The *coverage rate* denominates the fraction of the H400 training molecules which contains at least one of the extracted jumping fragments. This rate varies from 34.3% for a frequency threshold of 5 to 84.3% for a frequency threshold of 0.6%. The standard deviation of this coverage rate varies from 6.43 for a frequency threshold of 5% to 0.74 for a frequency threshold of 0.6%. These results show that the correlation of the coverage rate between any fold of cross-validation increases as the frequency threshold decreases. Analogous to the coverage rate, the H400 *success rate* is the fraction of the H400 test molecules which contains at least one of the extracted jumping fragments. This rate measures the performance of the decision rule for the classification of H400 structures. The H400 success rate varies from 38.3% for a frequency threshold of 5 to 81.9% for a frequency threshold of 0.6%. The fragments extracted from a set of H400 molecules also occur within H400 molecules outside of this set. The H402 success rate varies from 95.8% for a frequency threshold of 5 to 47.1% for a frequency threshold of 0.6%. These results show a correlation between the number of structures associated to a jumping fragment (H400 subset) and the occurrence of the same jumping fragment in the H402 set. The H402 success rate drops when the frequency threshold changes from 2.6 to 1%. The corresponding support in the 300 structures of the learning set is 8 (2.6%) and 3 (1%) structures. Consequently, the occurrences of the jumping fragments in the H402 molecules are very low when their supports in the H400 molecules are, in this study, around 10 structures. For full information, Table 8 lists all the jumping fragments associated to H400 (frequency of 5%) for which the H402 success rate is 95.8%. The success rate of the fully classifying version (see Table 7) of our illustrative decision rule, "a molecule is toxic just in case it contains a jumping fragment", varies from 58.1% for a frequency threshold of 5 to 69.9% for a frequency threshold of 0.6%. One may note that with a

Table 7. Coverage Rate of the Jumping Fragments on the H400 Molecules of the Learning Set and the Rate of Generalization of Property ii to the Testing Set

		frequency threshold (%)					
		5	4.3	3	2.6	1	0.6
learning set	support in H400 molecules	15	13	9	8	3	2
	coverage rate on H400 (%)	34.3	41.5	60.4	62.9	81	84.3
	coverage rate on H400 (SD)[a]	6.43	4.9	3.83	2.7	1.27	0.74
testing set	H400 success rate (%)	38.3	42.9	62.6	66.9	79	81.9
	H402 success rate (%)	95.8	94.3	85	81.0	55.8	47.1
	overall success rate (%)	58.1	60.6	69.9	70.7	71	69.9

[a] SD is standard deviation.

Table 8. Smiles Representation For the JF with a Threshold of 5%[a]

JEP (34 to 25)	JEP (25 to 22)	JEP (22 to 19)	JEP (19 to 18)	JEP (18 to 17)	JEP (17 to 16)	JEP (16 to 15)
cOP	Clc(c(cc1)Cl)cc1	cccOP(OC)(O)=S	c(ccc1OP=S)cc1	c(ccc(NO)cc)O	cccOC(N)=O	cccc(OC(N)=O)c
cOPO	Clcc(cccc)Cl	cccOP(OC)OC	cccc(OP(=S)O)c	c(ccc1NO)(O)cc1	C(C(CC1)C)CC1	cccc(OC(N)=O)cc
cOPOC	Clccccccl	cccOP(OC)(OC)=S	cccc(OP(=S)O)cc	ccc(cccN=O)O	cOP(OCC)=S	c(ccc1OC(N)=O)cc1
ccOP	ccc(c(cc)Cl)Cl	ccc(OC=O)cc	c(ccc1OP(=S)O)cc1	ccc(cccN(=O)O)O	ccOPOCC	cccccOCN
ccOPO	ccOP(O)O	ccccOC=O	ccccc(OP=S)c	ccc(cccNO)O	n(cNC)c	ccccc(OCN)c
cccOP	c(ccO)C	cccc(OC=O)c	ccccc(OP(=S)O)c	cc(OP(OC)=S)c	c(ccc1ccc)cc1	ccccOC(N)=O
cccOPO	cOC=O	cccc(OC=O)cc	cccccOP=S	ccc(OP(OC)=S)c	ccccOP(=S)(O)O	ccccc(OC(N)=O)c
cOP=S	ccOC=O	c(ccc1OC=O)cc1	cccccOP(=S)O	cccc(OP(OC)=S)cc	ccccOP(OC)O	cccccOCN
cOP(=S)O	cccOC=O	ccccOC=O	c(ccO)(C)c	cccc(OP(OC)=S)c	ccccOP(OC)(O)=S	cccccOC(N)=O
ccOPOC	cc(OP)c	ccccc(OC=O)c	c(cc(O)c)C	ccccc(OP(OC)=S)cc	ccccOP(OC)OC	cccOPOCC
OP(=S)(O)O	ccOP(=S)(O)O	cccccOC=O	c(cc(O)c)(C)c	c(ccc1OP(OC)=S)cc1	ccccOP(OC)(OC)=S	n(c(NC)n)c
O(P(OC)(O)=S)C	cc(OPO)c	cc(OPOC)c	ccc(ccO)C	ccccc(OP(OC)=S)c	ccccOP(=S)(O)O	n(c(nc)NC)c
O(C)P(=S)(O)O	ccOP(OC)O	ccc(OPOC)c	ccc(cc(O)c)C	cccccOP(OC)=S	ccccOP(OC)O	nc(NC)n
cOP(OC)=S	ccOP(OC)(O)=S	ccc(OPOC)cc	c(cc(c1)C)cc1O	cCOC	ccccOP(OC)(O)=S	cc(OP(O)O)c
cccOPOC	ccOP(OC)OC	ccccOP=S	ccc(ccC)O	ccCOC	ccccOP(OC)OC	ccc(OP(O)O)c
ccOP=S	ccOP(OC)(OC)=S	ccccOP(=S)O	ccc(cc(C)c)O	cc(COC)c	ccccOP(OC)(OC)=S	ccc(OP(O)O)cc
ccOP(=S)O	cccc(OP)c	cccc(OPOC)c	cccc(ccO)C	O(P(OCC)(O)=S)CC	cCOCC	cccc(OP(O)O)c
ccOP(OC)=S	cccc(OP)cc	cccc(OPOC)cc	cccc(ccC)O	O(P(OCC)(O)=S)C	ccCOCC	cccc(OP(O)O)cc
cccOP=S	ccc(OPO)c	c(ccc1OPOC)cc1	Clc(cc(cCl)Cl)c	O(CC)P(=S)(O)O	cc(COCC)c	c(ccc1OP(O)O)cc1
cccOP(=S)O	ccc(OPO)cc	ccccOP=S	Clcc(ccCl)Cl	n(cO)c	ccc(COC)cc	ccccc(OP(O)O)c
ClccCl	ccOP(O)O	ccccOP(=S)O	cOPOCC	ccccOP(O)O	ccccCOC	ccccccOP(O)O
Clc(cCl)c	cccc(OP)c	ccccc(OPOC)c	ncNC	ccccOP(O)O	cccc(COC)c	cccCOCC
Clcc(cc)Cl	cccc(OP)cc	cccccOPOC	c(cccN=O)O	Clc(c(ccCl)Cl)c	cccc(COC)cc	ccc(COCC)c
Clcc(ccc)Cl	c(ccc1OP)cc1	ncO	c(cccN=O)(O)c	Clc(c(cc(Cl)c)Cl)c	c(ccc1COC)cc1	c(cC=O)C
cOP(O)O	cccc(OPO)c	ccccOP(OC)=S	c(ccc(N=O)c)O	Clc(c(cc1Cl)Cl)cc1	ccccCOC	c(c(C=O)c)C
cccOP(OC)=S	cccc(OPO)cc	ccccOP(OC)=S	c(ccc(N=O)c)(O)c	Clc(ccc(cCl)Cl)c	ccccc(COC)c	cOCNC
cOP(=S)(O)O	c(ccc1OPO)cc1	c(ccccO)C	c(ccc(N=O)cc)O	Clc(c(cccCl)Cl)c	cccccCOC	cOC(NC)=O
cOP(OC)O	ccccOPOC	c(ccccO)(C)c	c(ccc1N=O)(O)cc1	Clcc(cccCl)Cl	cc(OCN)c	ccOCNC
cOP(OC)(O)=S	ccccc(OP)c	c(cccc(O)c)C	c(cccN(=O)O)O	Clcc(cc(cc)Cl)Cl	cc(OC(N)=O)c	ccOC(NC)=O
cOP(OC)OC	ccccc(OPO)c	cS	c(cccN(=O)O)(O)c	Clccc(c(cc)Cl)Cl	ccc(OCN)c	cccOCNC
cOP(OC)(OC)=S	ccccOPOC	cc(OP=S)c	c(ccc(N(=O)O)c)O	Clccc(cccCl)Cl	ccc(OCN)cc	cccOC(NC)=O
ccccOP	cccccOP	cc(OP(=S)O)c	c(ccc(N(=O)O)c)(O)c	cccCOC	ccc(OC(N)=O)c	ccOP(OCC)=S
ccccOPO	ccccccOPO	ccc(OP=S)c	c(ccc(N(=O)O)cc)O	ccc(COC)c	ccc(OC(N)=O)cc	
cccccOP	cc(OC=O)c	ccc(OP=S)cc	c(ccc1N(=O)O)(O)cc1	cOCN	ccccOCN	
cccccOPO	ccc(OC=O)c	ccc(OP(=S)O)c	c(cccNO)O	cOC(N)=O	cccc(OCN)c	
Clc(c(Cl)c)c	c(ccc1OCC)cc1	ccc(OP(=S)O)cc	c(cccNO)(O)c	ccOCN	cccc(OCN)cc	
Clc(c(cc)Cl)c	cccOP(=S)(O)O	cccc(OP=S)c	c(ccc(NO)c)O	ccOC(N)=O	c(ccc1OCN)cc1	
Clc(c(ccc)Cl)c	cccOP(OC)O	cccc(OP=S)cc	c(ccc(NO)c)(O)c	cccOCN	ccccOC(N)=O	

[a] In brackets the number of occurrences is given.

frequency threshold of 0.6% a fragment is frequent as soon as it occurs in two H400 molecules. This constraint seems very weak, and it may explain the decrease of the performances related to the frequency threshold of 0.6%. The frequency threshold which produces the optimal decision rule is 2.6% for which the decision rule is accurate on 71% of the data set.

Classifications using QSAR and Jumping Fragments. If we consider a methodology for an overall estimation of the classification associated to a compound, an approach combining the two methods should be interesting. For this publication and by considering the previous results (see Table 6), we just check if a new classification could be done for the 70 structures predicted as H402 classification (instead of H400, QSAR analysis) by using the jumping fragments associated to several frequency thresholds. From the frequency threshold of 5 to 2%, only five structures were removed from this set. By considering the correlation between the presence of these jumping fragments and a classification as H402, the probability is high that these five structures have a H400 classification. Starting from a frequency threshold of 1%, only 2 derivatives belonging to the cluster 4 will still remain in the H402 subset. By analyzing these last results, the structures, for instance, with the carbamate functions (cluster 11) are classified as H400 only with a frequency threshold of 1% (6 structures with carbamate functions are in the H402 set). This result shows that the carbamate function (often associated to acetylcholinesterase inhibitors) is not the main criteria to define a high toxicity to the organisms, but there exists a jumping fragment for which an association with the carbamate functions leads to this H400 classification. This situation with the same potential "toxic" chemical fragments in two sets (very toxic and harmful), but overall toxicity controlled by the nature of the other chemical fragments, will be particularly studied in the future.

CONCLUSION

This study has shown for the first time the potential of a new approach combining quantitative structure−activity relationship (QSAR) analysis with physicochemical descriptors and jumping fragments. Two objectives could be reached. The first one concerns an estimation of the potential classification of a derivative in ecotoxicology. The second objective is the possibility to get more information about the potential mode of action (MOA) associated to a derivative. Indeed, general QSAR equations are fitted to structures with a nonspecific MOA leading to first information concerning at least the baseline toxicity for each compound. The second method will give more information about the possibility to have a particular MOA associated to specific jumping fragments. Actually, the relationship between jumping fragments and a particular MOA is not really pointed out, except through the definition of the chemical subsets (depending of the frequency threshold) for which common jumping fragments are observed. To get a better view about the importance of some chemical features (in terms of MOA),

the future steps will be to refine these jumping fragments (through the notion of common substructure) and to analyze more clearly their weights toward the toxicity of the chemicals (straight importance or in association with another fragments). This will be carried out in the next research.

ACKNOWLEDGMENT

We thank Agence Nationale de la Recherche (ANR, ANR-07-CP2D-09-02) for financial support. We thank the regional council of Basse-Normandie for financial support ("programme emergence").

REFERENCES AND NOTES

(1) Guidance to Regulation (EC) No 1272/2008 on Classification, Labelling and Packaging of substances and mixtures. *European Commission*, **2008**; http://ecb.jrc.ec.europa.eu/documents/Classification-Labelling/CLP_Guidance_to_Regulation.pdf.
(2) Globally Harmonized System for the classification and labelling of chemicals. *United Nations Economic Commission for Europe*, **2007**; http://www.unece.org/trans/danger/publi/ghs/ghs_rev03/English/04e_part4.pdf.
(3) REACH. Registration Evaluation Authorization and restriction of Chemicals. *European Commission Environment*, **2007**; http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm.
(4) Rogers, M. D. The European Commission's White Paper "Strategy for a Future Chemicals Policy": A Review. *Risk Anal.* **2003**, *23* (2), 381–388.
(5) Benfenati, E.; Clook, M.; Fryday, S.; Hart, A. QSARs for regulatory purposes: the case for pesticide authorization. In *QSARs for regulatory purposes: the case for pesticide authorization*; Benfenati, E., Ed.; Elsevier: Milano, Italy, 2007; pp 1−54.
(6) Benfenati, E.; Gini, G. Computational predictive programs (expert systems) in toxicology. *Toxicology* **1997**, *119* (3), 213–225.
(7) Cronin, M. T. D.; Livingstone, D. J. *Predicting chemical toxicity and fate*; CRC Press: New York, 2004; pp 263−373.
(8) Faucon, J. C.; Bureau, R.; Faisant, J.; Briens, F.; Rault, S. Ecotoxicological endpoints in European notification procedure impact on the classification for the aquatic environment. *Chemosphere* **1999**, *38* (12), 2849–2863.
(9) Weyers, A.; Vollmer, G. Algal growth inhibition: effect of the choice of growth rate or biomass as endpoint on the classification and labelling of new substances notified in the EU. *Chemosphere* **2000**, *41* (7), 1007–1010.
(10) *ECOSAR, Estimation Programs Interface Suite (EPI Suite) for Microsoft Windows, v 4.00*; United States Environmental Protection Agency: Washington, DC, 2009.
(11) *TOXTREE*; European Central Bank: Frankfurt am Main, Germany; http://ecb.jrc.ec.europa.eu/qsar/home.php?CONTENU=/qsar/qsar-tools/qsar_tools_toxtree.php.
(12) (Q)SAR Application Toolbox. *OASIS*; Laboratory of Mathematical Chemistry: Bourgas, Bulgaria, 2008; http://toolbox.oasis-lmc.org/.
(13) Lozano, S.; Lescot, E.; Halm, M.-P.; Lepailleur, A.; Bureau, R.; Rault, S. Prediction of acute toxicity in fish by using QSAR methods and chemical modes of action. *J. Enzyme Inhib. Med. Chem.* **2009**, *25* (2), 195–203.
(14) EPAFHM: EPA Fathead Minnow Acute Toxicity; U. S. Environmental Protection Agency: Washington, DC; http://www.epa.gov/ncct/dsstox/sdf_epafhm.html.
(15) Auer, J.; Bajorath, J. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection. *J. Chem. Inf. Mod.* **2006**, *46* (6), 2502–2514.
(16) Discovering Emerging Graph Patterns from Chemicals. In *Lecture Notes in Computer Science*; Poezevara, G., Cuissart, B., Crémilleux, B., Eds.; Springer-Berlin: Heidelberg, Germany 2009; Vol. 5722, pp 45−55.
(17) Faucon, J. C.; Bureau, R.; Faisant, J.; Briens, F.; Rault, S. Prediction of the fish acute toxicity from heterogeneous data coming from notification files. *Chemosphere* **1999**, *38* (14), 3261–3276.
(18) Faucon, J. C.; Bureau, R.; Faisant, J.; Briens, F.; Rault, S. Prediction of the Daphnia acute toxicity from heterogeneous data. *Chemosphere* **2001**, *44* (3), 407–422.
(19) Van Leeuwen, C. J.; Van Der Zandt, P. T. J.; Aldenberg, T.; Verhaar, H. J. M.; Hermens, J. L. M. The application of QSARs, extrapolation and equilibrium partitioning in aquatic effects assessment for narcotic pollutants. *Sci. Total Environ.* **1991**, *109−110*, 681–690.
(20) Netzeva, T.; Pavan, M.; Worth, A. Review of data sources, QSARs and integrated testing strategies for aquatic toxicity. EUR 22943 EN. http://ecb.jrc.ec.europa.eu/documents/QSAR/EUR_22943_EN.pdf.
(21) Meylan, W. M.; Howard, P. H. Atom/fragment contribution method for estimating octanol-water partition coefficients. *J. Pharm. Sci.* **1995**, *84* (1), 83–92.
(22) Hahn, M. Receptor Surface Models. 1. Definition and Construction. *J. Med. Chem.* **1995**, *38* (12), 2080–2090.
(23) *PipelinePilot*; SciTegic, Inc: San Diego, CA; http://www.scitegic.com/.
(24) IUCLID, International Uniform Chemical Information Database. http://iuclid.eu/.
(25) ECB, European Chemicals Bureau. http://ecb.jrc.ec.europa.eu/documentation/ (accessed 2008).
(26) Garey, M. R.; Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, W. H., Ed. 1979.
(27) Dong, G.; Li, J., Efficient mining of emerging patterns: Discovering Trends and differences. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press: New York, 1999; pp 43−52.
(28) Auer, J.; Bajorath, J. Distinguishing between bioactive and modeled compound conformations through mining of emerging chemical patterns. *J. Chem. Inf. Mod.* **2008**, *48* (9), 1747–53.
(29) Auer, J.; Bajorath, J. Simulation of sequential screening experiments using emerging chemical patterns. *Med. Chem.* **2008**, *4* (1), 80–90.
(30) MOE, *Molecular Operating Environment*. Chemical Computing Group Inc: Montreal, Canada, 2007.
(31) Borgelt, C.; Berthold, M. R. Mining molecular fragments: finding relevant substructures of molecules. *Proceedings of the IEEE International Conference on Data Mining* **2002**, 51–58.
(32) Ting, R. M. H.; Bailey, J. Mining minimal contrast subgraph patterns. *Proceedings of the sixth SIAM international conference on data mining.* **2006**, 638–642.
(33) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Divers* **1996**, *2*, 1381–1991.
(34) PPDB Pesticide Properties Database; IUPAC: Research Triangle Park, NC; http://sitem.herts.ac.uk/aeru/iupac/.
(35) Schultz, T. W.; Lin, D. T.; Arnold, L. M. QSARs for monosubstituted anilines eliciting the polar narcosis mechanism of action. *Sci. Total Environ.* **1991**, *109−110*, 569–580.
(36) INCHEM; International Programme on Chemical Safety (IPCS): Geneva, Switzerland; http://www.inchem.org/.
(37) 3,4-dichloroaniline. European Chemicals Bureau (ECB); ECB: ; http://ecb.jrc.ec.europa.eu/IUCLID-DataSheets/95761.pdf.
(38) Ramos, E. U.; Vaal, M. A.; Hermens, J. L. M. Interspecies sensitivity in the aquatic toxicity of aromatic amines. *Environ. Toxicol. Pharmacol.* **2002**, *11* (3−4), 149–158.
(39) Brack, W.; Rottler, H. Toxicity testing of highly volatile chemicals with green algae. *Environ. Sci. Pollut. Res. Int.* **1994**, *1* (4), 223–228.
(40) Coumatetralyl. *European Chemicals Bureau*; ECB: ; http://ecb.jrc.ec.europa.eu/documents/Biocides/ANNEX_I/ASSESSMENT_REPORTS/AnnexI_AR_5836−29-3_PT14_en.pdf.

CI100092X