VIRTUAL SCREENING FOR ENVIRONMENTAL POLLUTANTS:
STRUCTURE–ACTIVITY RELATIONSHIPS APPLIED TO A DATABASE
OF INDUSTRIAL CHEMICALS

TOMAS ÖBERG*
University of Kalmar, Department of Biology and Environmental Science, SE-391 82 Kalmar, Sweden

(Received 17 May 2005; Accepted 26 September 2005)

Abstract—The current risk paradigm calls for individual consideration and evaluation of each separate environmental pollutant, but this does not reflect accurately the cumulative impact of anthropogenic chemicals. In the present study, previously validated structure–activity relationships were used to estimate simultaneously the baseline toxicity and atmospheric persistence of approximately 50,000 compounds. The results from this virtual screening indicate fairly stable statistical distributions among small anthropogenic compounds. The baseline toxicity was not changed much by halogen substitution, but a distinct increase seemed to occur in the environmental persistence with increased halogenation. The ratio of the atmospheric half-lives to the median lethal concentrations provides a continuous scale with which to rank and summarize the incremental environmental impacts in a mixture-exposure situation. Halogenated compounds as a group obtained a high ranking in this data set, with well-known pollutants at the very top: DDT metabolites and derivatives, polychlorinated biphenyls, diphenyl ethers and dibenzofurans, chlorinated paraffins, chlorinated benzenes and derivatives, hydrochlorofluorocarbons, and dichlorononylphenol. Environmentally friendly chemicals that obtained the lowest rank are nearly all hydroxylated and water-soluble. Virtual screening can assist with "green chemistry" in designing safe and degradable products and enable assessment of the efficiency in chemicals risk management.

Keywords—Quantitative structure–activity relationships Risk assessment Persistent organic pollutants Mixture exposure

INTRODUCTION

The European Inventory of Existing Commercial Chemical Substances, a database maintained by the European Community, lists approximately 100,000 substances, 80,000 of which are thought to be in use [1]. Many of these chemicals could inflict damage if released into environmental media (air, soil, and water). The current risk paradigm calls for individual consideration and evaluation of each separate compound, but this does not reflect accurately the cumulative impact of anthropogenic chemicals acting additively [2]. Furthermore, no distinct demarcation occurs between persistent and nonpersistent compounds [3]. These observations are difficult to reconcile with a stringent classification into hazardous and nonhazardous chemicals.

Another problem in risk analysis of chemicals is the lack of experimental data and observations to assist in the process of hazard identification. More chemicals are waiting to be evaluated than current laboratory resources permit, and animal testing is becoming increasingly controversial. In vitro assays and in silico computational methods provide an alternative to animal testing [4], but measured physical properties also are lacking for a large portion of the chemicals in current use. A proactive approach is hampered by the fact that new chemicals, not yet synthesized, are impossible to evaluate.

Virtual screening of chemical libraries—composed of known or constructed compounds—is now widely used in drug discovery [5,6]. The aim of virtual screening is to score, rank, or filter a set of compounds using computational methods [7]. Virtual screening often is based on quantitative structure–activity relationships (QSARs). These statistical models describe the correlation between a biological, chemical, or physical property and a group of molecular descriptors for the investigated compounds. The QSARs are already used for ranking and prioritizing in environmental risk assessment [8,9], and the number of published QSARs amount to at least 17,000 [10]. However, few of these QSARs meet the more rigorous acceptability criteria suggested for use in a regulatory context [11–13]. External validation is necessary for a realistic estimate of future prediction errors. Similarly, well-defined applicability boundaries also are needed, particularly when a QSAR model is applied to virtual screening of a chemical library. The QSARs applied in the present study have been developed with an aim to meet these criteria [14,15].

Baseline toxicity or narcosis is the minimum nonspecific toxicity exhibited by an organic compound. Most industrial chemicals, with different sizes, shapes, and functional groups, express only baseline toxicity [16]. Environmental samples (e.g., contaminated surface waters) often contain mixtures of a large number of these anthropogenic compounds. Additivity is a fundamental part of the baseline toxicity concept, and attention therefore must be given to the possible joint action of all mixture components [17–22]. Below the thresholds for specific toxicity, the baseline toxicity may still amount to an environmentally relevant effect [2,23]. Therefore, baseline toxicity likely is a prominent effect in the environment and of particular importance for persistent organic pollutants [17,24].

Baseline toxicity results from the accumulation of chemicals and the ensuing disruption of normal functions in biological membranes; thus, it is closely related to other forms of bioaccumulation. Baseline toxicity also is a feature well suited for modeling, because it essentially can be described as
a partitioning process even though the target sites may be quite specific [25]. A validated QSAR for baseline toxicity recently was reported by the author and applied to a diverse set of small organic compounds (32–285 amu) [14].

This QSAR model was developed using a bilinear-regression technique to project a multidimensional descriptor space down to a few latent variables. The Euclidean distance to the model and the Mahalanobis distance within the model space defined the domain of applicability. The predicted baseline toxicities indicated an acute hazard for 26% of the investigated compounds, with the 96-h median lethal concentration (LC50) for fish (Pimephales promelas) being less than 10 mg/L.

Toxicity itself does not constitute an environmental risk without exposure. Exposure for chemical hazards is determined by release into the environment and removal by physical, chemical, and biological processes. The degradation varies considerably between different environmental media, with air being the most reactive [26]. Reaction with hydroxyl (OH) radicals in the troposphere is a dominant removal pathway for many industrial chemicals [27]. The rate constants for this gas-phase reaction therefore can be used as an indicator of environmental persistence. Experimental data are available for a large number of small organic compounds (30–300 amu), and a validated QSAR recently was developed by the author [15]. Screening the same data set as used previously [14] indicated a slow atmospheric degradation (>2 d) for 15% of the investigated compounds. An atmospheric half-life of 2 d has been proposed as the screening criterion for persistent organic pollutants [28]. The actual environmental half-life is, of course, much longer, because only a minor portion usually is present at the same time in the atmosphere.

The aim of the present investigation is to apply the two QSAR models to a large chemical library to study the distribution of atmospheric persistence and baseline toxicity among small anthropogenic compounds. Such a screening is useful to identify compounds that are potentially hazardous to the environment (toxic and persistent) as well as those that could be more environmentally friendly.

**MATERIALS AND METHODS**

The experimental part of the investigation consisted of three steps, as in the previous QSAR studies [14,15]: Collection of structural data, generation of molecular descriptors, and data analysis.

**Collection of structural data**

The SMILESCAS Database (Syracuse Research, North Syracuse, NY, USA) has 109,611 entries for compounds with the two-dimensional (2D) structure specified by the Simplified Molecular Input Line Entry System (SMILES) together with their Chemical Abstract Service (CAS) registry numbers. The chemical names also are given for many of these compounds. The database has good coverage of most industrial chemicals and is integrated with the stand-alone programs that are part of the Estimation Program Interface suite of models (U.S. Environmental Protection Agency, Office of Pollution Prevention Toxics, Washington, DC). This database was used to generate molecular descriptors for further analysis and evaluation.

Calibration data for the two QSAR models were obtained from the Fathead Minnow Database (U.S. Environmental Protection Agency Mid-Continent Ecology Division, Duluth, MN) and the PhysProp Database (Syracuse Research), respectively.

**Generation of molecular descriptors**

The 2D molecular structures from the SMILESCAS database were used directly as input for the generation of 867 empirical descriptors using the Dragon software (Ver 5.0; Milano Chemometrics and QSAR Research Group, University of Milano–Bicocca, Milan, Italy). The molecular descriptors that were generated include constitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge-adjacency indices, burden eigenvalue descriptors, topological charge indices, eigenvalue-based indices, functional group counts, and atom-centered fragments. Most of these molecular descriptors have been reviewed in a recent textbook by Todeschini and Consonni [29].

**Data analysis**

The data analysis was carried out using the Unscrambler software (Ver 9.1: Camo Process AS, Oslo, Norway) and Matlab (Ver 7.0; MathWorks, Natick, MA, USA). The two QSAR models applied here were developed previously from two experimental data sets (of 311 and 743 compounds, respectively) using partial-least-squares regression (PLSR) [14,15]. Partial-least-squares regression is based on a linear transformation of the original descriptors to a limited number of orthogonal factors, attempting to maximize the covariance between the descriptors and the response variable. The term *latent variable* is used to denote the PLSR factors, because they can be interpreted as describing the inherent chemical properties [30]. Multivariate calibration has been reviewed by Martens and Næs [31], Wold et al. [32], and Næs et al. [33].

The QSAR models for baseline toxicity and the OH radical reaction rate constant used 218 descriptor variables projected onto five latent variables and 333 descriptor variables projected onto seven latent variables, respectively. Descriptor variables were selected for inclusion in the final models based on significance tests using jackknifing in preliminary runs. Organometallic and organosilicon compounds were not included in the derivation of the QSAR models and, therefore, were excluded this time as well.

The two QSAR models were characterized by the standard deviations of the prediction residuals for the calibration objects and the external test set: Standard error of calibration (SEC) and standard error of prediction (SEP), respectively. The explained variances are defined as the sum of squares due to regression divided by the sum of squares about the mean: $R^2_{cal}$ (square of the multiple-correlation coefficient for the calibration objects), and $Q^2_{ext}$ (square of the multiple-correlation coefficient for the external test set).

The PLSR-based QSARs define a valid domain for the descriptor variables, and new prediction objects were assessed by the residual standard deviation (the Euclidean distance to the PLSR model) and the leverage (the Mahalanobis distance to the calibration objects within the PLSR model space). These two distance measures were then used to decide if an object was within the domain of applicability. Here, the 5% significance level was chosen as the limit for the residual standard deviation, and the limit for the leverage was set to threefold the average leverage for the calibration objects.

**QSAR validation and performance**

Both QSAR models used in the present study were developed with the same approach. Initially, a large number of
RESULTS AND DISCUSSION

The two QSAR models developed previously with PLSR [14, 15] were used to estimate the LC50 values (log mmol/L) and the OH reaction rate constants at 298 K (log cm3/molecule/s). Subsequently, the model predictions were recalculated to express toxicity in terms of milligrams per liter and atmospheric half-life in terms of days.

**Bivariate distribution of toxicity and persistence**

The baseline toxicity QSAR (for acute fish toxicity [Pimephales promelas], expressed as the 96-h LC50) was applied successfully to 98,083 compounds in the database, and 66,884 of these compounds were within the domain of applicability. Twenty-six percent (25,316) of the 98,083 compounds had LC50 values of less than 10 mg/L, corresponding to categories 1 and 2 in the Globally Harmonized System of Classification and Labeling of Chemicals [34]. This is exactly the same proportion reported previously for a more limited data set [14]. However, because many compounds have specific toxicity, it can be regarded only as a minimum estimate of this proportion.

The OH reaction rate constant QSAR also was applied successfully to the same number of compounds, and 51,894 of these were within the domain of applicability. Fifteen percent (7,557) of the 51,894 compounds had half-lives of longer than 2 d, and this is almost the same proportion (16%) as that in the previously investigated, more limited data set.

When the two QSARs were applied sequentially, 50,074 compounds were within the domain of applicability for both models. The approximately log-normal bivariate distribution of toxicity and persistence among these compounds is shown in Figure 1.

The results from the virtual screening with these two QSARs thus seem to indicate a fairly stable bivariate distribution of baseline toxicity and persistence among small organic compounds. It is premature to claim that this statistical distribution will hold for the properties of chemicals in commercial use, but it certainly opens up an interesting discussion about the possibilities to generalize the observations. If baseline toxicity and persistence among the whole population of small-size anthropogenic compounds can be described by a statistical distribution like this one, then it also may be possible to use a probabilistic approach to assess the chemical risks from general exposure. Undoubtedly, this will require new approaches to risk management and risk policy making.

**Effect of halogenation**

In all, 1,653 compounds meet both the criteria for acute aquatic toxicity (LC50, <10 mg/L) [34] and environmental persistence (atmospheric half-life, >2 d) [28]. Most of these compounds (88%) are halogenated, but the effect of halogenation reach far outside of this selected group of potentially hazardous compounds. The logarithmic bivariate distributions of toxicity and persistence for nonhalogenated (39,428) and halogenated (10,646) compounds are shown separately, with two density plots (Figs. 2 and 3).

The baseline toxicity was not changed much by halogenation, but a distinct increase seemed to occur in the environmental persistence. The geometric mean and standard deviation of the estimated atmospheric half-lives for nonhalogenated compounds were 0.25 and 3.4 d, respectively, compared to 1.9 and 5.6 d, respectively for the halogenated compounds. Even with these shifts in location and spread, the frequency distributions of the two chemical groups are continuous and overlapping.

**Hazard indicator**

Any demarcation limit for toxicity or persistence will be arbitrary, and both aspects need to be considered simultaneously. One approach to create a continuous scale for risk assessment of chemicals would be to combine the two characteristics of toxicity and persistence used in the present study into a joint toxic persistence rating (TPR). The reciprocal LC50 values can be used as a convenient scale to compare additive toxic potencies, similar to the toxic equivalence factors used for dioxin-like compounds [35]. It also, however, is necessary to consider the persistence, and multiplying with the estimated atmospheric half-life can achieve this. The more toxic and persistent a chemical, the greater the TPR, and vice versa.

**Toxic persistence rating = atmospheric half-life (d)/LC50 (mg/L)**

A chemical with an atmospheric half-life of 2 d and an acute toxicity of 10 mg/L would have a TPR value of 2/10 = 0.2. Thirteen percent (6,447) of compounds for which both QSARs were applicable had a TPR of greater than 0.2. This, of course, is a much wider and more relaxed definition of a persistent organic pollutant compared with the current standard, but a significant impact on health and environment accumulating from compounds outside the Stockholm convention is not unlikely.
Virtual screening for environmental pollutants

The environmental half-lives are proportional to the time-integrated exposure for the amounts emitted, and baseline toxicity is assumed to be an additive property. Thus, the amount or concentration of each compound considered can be multiplied by its TPR to form a quantity that can be added together to estimate the total environmental burden from mixture exposure to organic environmental pollutants. To facilitate comparison, this estimated quantity can, without difficulty, be normalized and expressed as a suitable compound equivalent.

**Well-known environmental pollutants**

Many environmental pollutants of concern are halogenated compounds, and many contain an aromatic carbon skeleton as well. In the evaluated SMILESCAS database, 62% of the compounds are aromatics, 23% halogenated, and 16% halogenated aromatics. Among the 6,447 selected compounds with a TPR greater than 0.2, 69% were aromatics, 55% halogenated, and 40% halogenated aromatics. The current focus by various regulatory authorities on halogenated aromatics as a major compound group of concern therefore is supported by these findings, but obviously, a large group of potentially toxic and persistent compounds also have other structural features and composition.

If, however, we limit ourselves only to the 100 most toxic and persistent compounds—those with the highest TPR values—a familiar pattern is revealed. All 100 of these compounds are halogenated and have comparably high molecular weights (median, 292 amu), and most of them are well-known environmental pollutants, such as DDT metabolites and derivatives, polychlorinated biphenyls and contaminants, chlorinated paraffins, chlorinated benzenes and derivatives, hydrochlorofluorocarbons, and dichlorononylphenol (Table 1). The constraints imposed by the domain of applicability did not permit evaluation of some related pollutants like polychlorinated dibenzop-dioxins and polybrominated diphenyl ethers.

**Additional information needed**

The effect of the global mixture exposure to small anthropogenic chemicals can be evaluated only by simultaneously...
considering the exposure for all these compounds, and QSAR models are then the only viable approach. The suggested TPR concept can extend the estimates of additive toxic effects with an indication of the potential for a persistent environmental exposure. The additional data needed are production data and estimates of the losses into environmental media through the entire product cycle, from “cradle to grave.” It can be argued that these data will not be easy to obtain and that the TPR concept has several weaknesses, such as considering only one type of toxicity in one species, not considering other removal pathways, and not including sufficient detail about the environmental distribution.

Of course, it is difficult to obtain detailed data regarding production and importation of chemical goods and products. However, national product registers are already in operation, and their use could be expanded. The alternative, to test and monitor every chemical, certainly will not be any easier to achieve. Specific toxicity and nonadditive behavior often are a concern, but several investigations have indicated that additive toxic effects are dominating in mixture exposure at low concentrations [2,18,20,23,36–39]. Species differences also are likely to be less important, because all living organisms experience baseline toxicity. More detailed information regarding environmental distribution and removal is of great value in refining the exposure estimates, but the gas-phase reaction with OH radicals in the lower atmosphere probably will remain the dominant degradation pathway for the most persistent organic compounds. In addition, atmospheric half-lives are determined more readily than those in other environmental media.

Opportunity for green chemistry

The chemical industry and chemical products are vital for the economic growth and welfare in modern industrial society. Therefore, perceived risks must always be weighed against the benefits coupled to the production and use of these products. An index such as the TPR outlined here could assist in a continuous selection process of environmentally friendly substitutes for compounds with a potential to cause harm. The systematic application of QSAR models and virtual screening, as the pharmaceutical industry does in drug discovery, also could assist with “green chemistry” in designing safe and degradable products [40,41].

If we examine the 100 compounds with the lowest TPR values, then a pattern of structural properties is revealed that is equally consistent as the patterns for those with the highest values. None of these 100 compounds is halogenated, and most are hydroxylated: Diols, sugars, sugar derivatives, and other hydroxylated compounds (Table 2). All of these environmentally friendly compounds can easily participate in hydrogen bonding and be expected to have high water solubility. Therefore, they will not bioaccumulate and will be excreted easily by biological organisms. However, some compounds with low TPR values may still express specific toxicity (e.g., by interfering with metabolic pathways).

The main factors of variation for the two QSAR models have been identified previously [14,15]. Baseline toxicity is influenced mainly by size and polar interactions, atmospheric persistence by halogenation, and the presence of reactive sites [14,15]. The suggested TPR ranking therefore is influenced by the same factors, some of which are shown in Figure 4. It is interesting to note the nearly monotonic relationships; a gradual increase in the TPR value occurs with each added chlorine or fluorine atom. Likewise, a steady decrease in the TPR value occurs with the addition of hydrogen atoms bonded to oxygen and nitrogen (i.e., hydrogen donors).

CONCLUSION

Chemoinformatics and computational chemistry can provide rational tools for design of products that are easily degradable and have low toxicity. Still, the most important feature with the suggested virtual screening approach is that it can encourage alternative thinking about risks from exposure to chemicals. Previous screening approaches have aimed to isolate a limited number of hazardous environmental pollutants from the majority of organic chemicals [28]. A tiered system with screening as the first step has been suggested, but separate assessment of each chemical has remained the final objective. An example of this compound-specific risk management approach is the use of ranking and scoring to create various

Table 2. Estimated toxic persistence rating (TPR), median lethal concentration (96-h LC50 for Pimephales promelas), and atmospheric half-lives (based on the hydroxyl radical reaction rate) for the 100 compounds rated as the least toxic and persistent

<table>
<thead>
<tr>
<th>Compound</th>
<th>TPR (dL/mg)</th>
<th>LC50 (mg/L)</th>
<th>Half-life (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diols</td>
<td>13</td>
<td>2.0E-6–8.5E-6</td>
<td>8,200–96,000</td>
</tr>
<tr>
<td>Sugars*</td>
<td>28</td>
<td>1.9E-6–8.8E-6</td>
<td>4,200–28,000</td>
</tr>
<tr>
<td>Sugar derivatives*</td>
<td>15</td>
<td>3.8E-6–8.6E-6</td>
<td>4,700–67,000</td>
</tr>
<tr>
<td>Other hydroxylated</td>
<td>37</td>
<td>1.7E-6–8.5E-6</td>
<td>2,700–90,000</td>
</tr>
<tr>
<td>Other</td>
<td>7</td>
<td>6.2E-6–8.8E-6</td>
<td>1,900–50,000</td>
</tr>
</tbody>
</table>

* Several sugars and sugar derivatives are stereoisomers.
Virtual screening for environmental pollutants

Environ. Toxicol. Chem. 2006 1183

priority lists [42]. Here, instead, it is suggested to use toxicity and persistence estimates to get an indication of the total environmental impact. By recalculating the mass flow of each individual compound to an equal scale using the TPR concept, these quantities subsequently may be added together to estimate the environmental burden imposed by society as a whole. Instead of looking at the incremental effects one by one, it may then be possible to ask what the environment can endure in a sustainable society and if we are heading in the right direction. Attaching models that also estimate the environmental burden imposed by society as a whole. These quantities subsequently may be added together to estimate the environmental burden imposed by society as a whole. Attaching models that also estimate the environmental burden imposed by society as a whole.

Acknowledgement—The present work was funded by the Faculty of Science at the University of Kalmar. I would like to thank Jan Bergström, Pasi Peltola, Karin Oberg, and two anonymous reviewers for helpful comments and suggestions that improved the manuscript.

REFERENCES


