

Giving Molecules an Identity. On the Interplay Between QSARs and Partial Order Ranking

Lars Carlsen *

Awareness Center, Hyldeholm 4, Veddelev, DK-4000 Roskilde, Denmark

* To whom correspondence should be addressed: e-mail: LC@AwarenessCenter.dk

Received: 2 June 2004 / Accepted: 30 June 2004 / Published: 31 December 2004

Abstract: The interplay between ‘noise-deficient’ QSAR and Partial Order Ranking, including analysis of average linear ranks, constitutes an effective tool in giving substances which have not been investigated experimentally an identity by comparison with experimentally well-characterized, structurally similar compounds. It is disclosed that experimentally well-characterized compounds may serve as substitutes for highly toxic compounds in experimental studies without exhibiting the same extreme toxicity, while from an overall viewpoint they exhibit analogous environmental characteristics.

Keywords: ‘Noise-deficient QSARs; Partial Order Ranking; Hasse Diagrams; Organophosphates; Nerve agents.

Introduction

The lack of data for the vast majority of existing chemicals is well known and constitutes obviously a significant problem in relation to e.g., risk assessment. Thus, according to the European Commission only in the case of approximately 14% of the HPV (High Production Volume) chemicals on the EINECS list, comprising 100,116 entries, the minimum required data for evaluating the chemicals were available. For approximately 21% of the compounds no data at all concerning their potential impact on the environment and human health were found [1]. In a study by the Danish EPA [2] it was concluded that even in major sources of test data, information on selected ecotoxicological effects could only be found for very limited number of the compounds on the EINECS list (acute toxic effect: 10.5%, reproductive damage: 2.2%, genetic damage: 3.2%, carcinogenic effect: 1.6%, effect on the aquatic environment: 3.5%). Since intensive and experimental evaluations of chemicals are rather costly [3, and references therein], QSAR derived data for physico-chemical as well as toxicological

endpoints appear as an attractive alternative. However, although the lack of data can be remedied to a certain extent through QSAR modeling, this will leave us with the possibility of characterizing the single molecules based on single parameters, such as solubility, octanol-water partitioning, vapor pressure, biodegradation - and bioaccumulation potential. However, to establish an identity for a given molecule, e.g., as a potential PBT substance requires taking several parameters into account simultaneously, i.e., Persistence, Bioaccumulation and Toxicity.

In the present study the advantageous use of so-called “noise-deficient” QSARs, developed using data from experimentally well-characterized compounds as the training set, as a preprocessing tool to derive the desired endpoints for substances where experimental data are not available. Subsequently, these endpoints will be applied as descriptors in establishing a partial ordering of combined sets of compounds, hereby giving the experimentally not investigated compounds an identity by comparing to structurally related, experimentally well-characterized compounds [4,5].

Methods

QSAR

In the present study the end-points are generated through QSAR modeling, the EPI Suite being the primary tool [6]. To generate new linear “noise-deficient” QSAR models, EPI generated values for, e.g., $\log Sol$, $\log K_{OW}$, $\log VP$ and $\log HLC$ are further treated by estimating the relationships between the EPI generated data and available experimental data [7] for the a series of experimentally well-characterized compounds in the training set, the general formula for the end-points, D_i , to be used being

$$D_i = a_i \times D_{EPI} + b_i \quad (1)$$

D_{EPI} is the EPI generated end-point value and a_i and b_i being constants. The $\log K_{OW}$ values generated in this way are subsequently used to generate $\log BCF$ values according to the Connell formula [8]

$$\log BCF = 6.9 \times 10^{-3} \times (\log K_{ow})^4 - 1.85 \times 10^{-1} \times (\log K_{ow})^3 + 1.55 \times (\log K_{ow})^2 - 4.18 \times \log K_{ow} + 4.72 \quad (2)$$

The model was somewhat modified. Thus, a linear decrease of $\log BCF$ with $\log K_{OW}$ was assumed in the range $1 < \log K_{OW} < 2.33$, the $\log BCF = 0.5$ for $\log K_{OW} \leq 1$, the latter value being in accordance with BCFWin [6]. Subsequently data for not characterized compounds are calculated based on these formulae and the appropriate EPI generated data.

In the present study a training set consisting of up to 65 organo phosphorus (OP) insecticides are applied. Due to the lack of experimental data for the training set compounds with regards to their biodegradation, the above procedure was not applicable to the biodegradation potential, $BDP3$. Thus, data on $BDP3$ are used as estimated by the appropriate modules in the EPI Suite.

Partial Order Ranking

The theory of partial order ranking is presented elsewhere [9] and its application in relation to QSAR is presented in previous papers [10-13]. In brief, Partial Order Ranking is a simple principle, which a priori includes “ \leq ” as the only mathematical relation. If a system is considered, which can be described by a series of descriptors p_i , a given compound A, characterized by the descriptors $p_i(A)$ can be compared to another compound B, characterized by the descriptors $p_i(B)$, through comparison of the single descriptors, respectively. Thus, compound A will be ranked higher than compound B, i.e., $B \leq A$, if at least one descriptor for A is higher than the corresponding descriptor for B and no descriptor for A is lower than the corresponding descriptor for B. If, on the other hand, $p_i(A) > p_i(B)$ for descriptor i and $p_j(A) < p_j(B)$ for descriptor j, A and B will be denoted incomparable. In mathematical terms this can be expressed as

$$B \leq A \Leftrightarrow p_i(B) \leq p_i(A) \text{ for all } i \quad (3)$$

Obviously, if all descriptors for A are equal to the corresponding descriptors for B, i.e., $p_i(B) = p_i(A)$ for all i, the two compounds will have identical rank and will be considered as equivalent. It further follows that if $A \leq B$ and $B \leq C$ then $A \leq C$. If no rank can be established between A and B these compounds are denoted as incomparable, i.e. they cannot be assigned a mutual order.

In partial order ranking – in contrast to standard multidimensional statistical analysis - neither assumptions about linearity nor any assumptions about distribution properties are made. In this way the partial order ranking can be considered as a non-parametric method. Thus, there is no preference among the descriptors. However, due to the simple mathematics outlined above, it is obvious that the method *a priori* is rather sensitive to noise, since even minor fluctuations in the descriptor values may lead to non-comparability or reversed ordering. The graphical representation of the partial ordering is often given in a so-called Hasse diagram [14-17]. In practice the partial order rankings are done using the WHasse software [17].

Linear extensions

The number of incomparable elements in the partial ordering may obviously constitute a limitation in the attempt to rank e.g. a series of chemical substances based on their potential environmental or human health hazard. To a certain extent this problem can be remedied through the application of the so-called linear extensions of the partial order ranking [18,19]. A linear extension is a total order, where all comparabilities of the partial order are reproduced [9,16]. Due to the incomparisons in the partial order ranking, a number of possible linear extensions corresponds to one partial order. If all possible linear extensions are found, a ranking probability can be calculated, i.e., based on the linear extensions the probability that a certain compound have a certain absolute rank can be derived. If all possible linear extensions are found it is possible to calculate the average ranks of the single elements in a partially ordered set [20,21]. The average rank is simply the average of the ranks in all the linear extensions. On this basis the most probably rank for each element can be obtained leading to the most probably linear rank of the substances studied.

The generation of the average rank of the single compounds in the Hasse diagram is obtained applying the simple empirical relation recently reported by Brüggemann et al [22]. The average rank of a specific compound, c_i , can be obtained by the simple relation

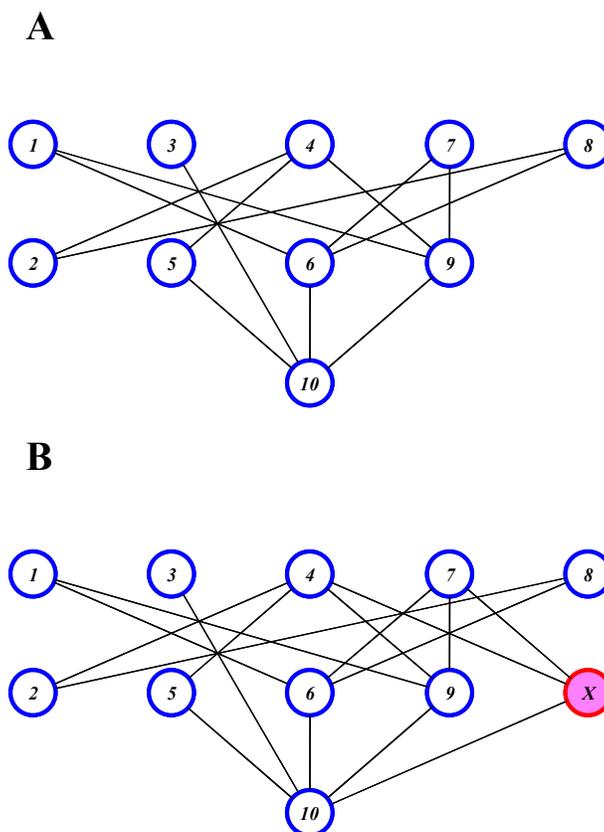
$$Rk_{av}(c_i) = (N+1) - (S(c_i)+1) \times (N+1) / (N+1-U(c_i))$$

(4) where N is the number of elements in the diagram, $S(c_i)$ the number of successors to c_i and $U(c_i)$ the number of elements being incomparable to c_i [22].

Results and Discussion

The basic idea of using partial order ranking for giving molecules an identity is illustrated in Figure 1. Thus, let us assume that a suite of 10 compounds has to be evaluated and that the evaluation should be based on three pre-selected criteria, e.g., persistence, bioaccumulation and toxicity. Let the resulting Hasse diagram be the one depicted in Figure 1A. If we apply the three descriptors representing biodegradation, bioaccumulation and toxicity, respectively, so the more persistent, the more bioaccumulating and the more toxic a substance would be the higher in the diagram it would be found, Figure 1A discloses that the compounds in the top level, i.e., compounds 1, 3, 4, 7 and 8 on a cumulative basis can be classified as the environmentally more problematic of the 10 compounds studied with respect to their PBT characteristics, whereas compound 10 that a found in the bottom of the diagram is the less hazardous.

Figure 1. Illustrative Hasse diagram of A: 10 compounds using three descriptors and B: the same 10 compounds plus one new compound X.



Subsequently we can introduce compounds solely characterized by QSAR derived data in order to give this new compound, X, an identity, e.g., in an attempt to elucidate the environmental impact of X. Adopting the above discussed 10 compounds and the corresponding Hasse diagram (Figure 1A) we then introduced the compound X. The revised Hasse diagram, now including 11 compounds is visualized in Figure 1B. It is immediately disclosed that compound X has now obtained an identity in comparison to the originally well-characterized compounds, as it is evaluated as less environmentally harmful than compounds 4 and 7, but more harmful than compound 10. Thus, through the partial order ranking the compound, X, has obtained an identity in the scenario with regard to its potential environmental impact.

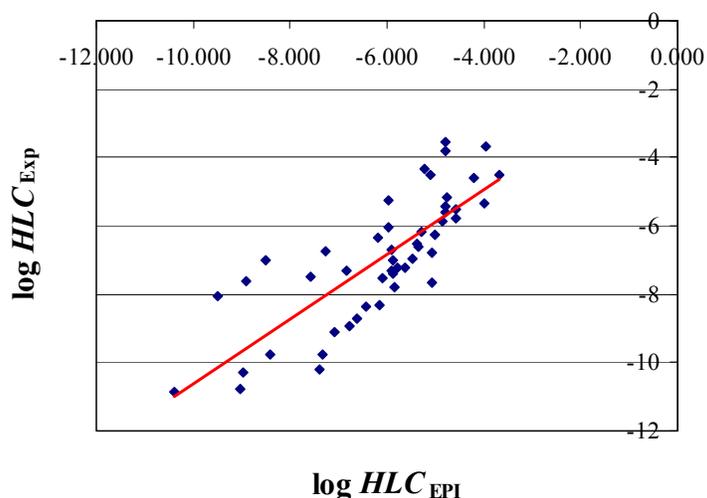
To illustrate the above an example from our current study on the physico-chemical characteristics of OP compounds with special emphasis on chemical warfare nerve agents as the G-agents, like Tabun, Sarin and Soman, and V-agents, like VX, shall be used [4,5]. In the present study we shall focus on the aqueous persistence of OP insecticides and know and potential nerve agents as expressed through the solubility (*Sol*), the biodegradation potential (*BDP*) and the Henry's Law Constants (*HLC*), the latter being derived based on the EPI values as given by HenryWin [6].

As mentioned the EPI Suite [6] has been the primary tool for QSAR modeling, the single EPI generated values for $\log Sol$, $\log K_{OW}$, $\log VP$ and $\log HLC$ being further treated to generate new linear "noise-deficient" QSAR models, cf. eqn. 1 [4].

As an example the new 'noise-deficient' QSAR model for $\log HLC$ is depicted in Figure 2, the corresponding model being expressed through eqn. 5 [4].

$$\log HLC = 0.946 \times \log HLC_{EPI} - 1.168; r^2 = 0.636 \quad (5)$$

Figure 2. Visualization of the EPI-based modified QSAR modeling of $\log HLC$ based on 49 OP insecticides

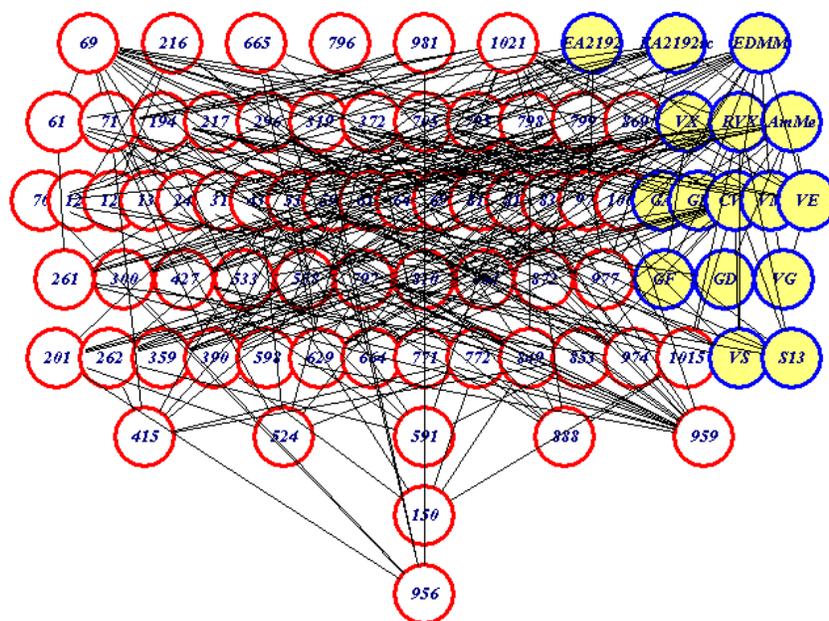


The 'noise-deficient' QSAR for the solubility was derived analogously, the resulting model being described through eqn. 6 [4].

$$\log Sol = 0.983 \times \log Sol(EPI) + 0.625; n = 64, r^2 = 0.830 \quad (6)$$

The generated end-point are subsequently used to generate partial order rankings of the the 65 OP insecticides together with the 16 known potential nerve agents taking two or more descriptors simultaneously into account. Thus, as in total 81 compounds are included in the subsequent ranking procedure, the resulting Hasse diagrams may seem somewhat confusing. Figure 3 depicts the Hasse diagram disclosing the mutual ranking of the compounds due to their aqueous persistence, i.e., bringing simultaneously the solubility ($\log Sol$), the biodegradation potential for ultimate biodegradation ($BDP3$) and Henry's Law Constant ($\log HLC$) into play.

Figure 3. Hasse diagram displaying the aqueous persistence of the 65 OP insecticides (white/red) and 16 nerve agent (yellow/blue), The numbers corresponds to the numbering of the OP insecticides in the FADINAP database [7]



From the above figure it can be seen that the nerve agent VX is located at the same level as the compounds 61 (Anilofos), 71 (Azinphos methyl), 194 (Chlorfenvinphos), 217 (Chlorpyriphos methyl), 296 (Dialifos), 319 (Dicrotophos), 372 (Ditalimfos), 705 (Monocrotophos), 795 (Phosalone), 798 (Phosmet), 799 (Phosphamidon) and 869 (Pyraclofos), in addition to the Russian version of VX (RVX) and the potential nerve agent AmMe (Amiton methyl).

A priori the location of the compounds on the same level in the Hasse diagram suggests these compounds to be close in their overall characteristics based on the set of descriptors used, i.e. solubility, biodegradation potential and Henry's Law Constant. However, a further analysis appears to be necessary in order eventually to disclose how close these compounds actually are. For this analysis the concept of average rank [4,5,22,23] was adopted. Thus, it is assumed that if the average ranks, Rk_{av} , of two compounds are close, the two compounds will on an average basis display similar characteristics as being determined by the set of descriptors applied. In Table 1 the average ranks for the above-mentioned OPs are given together with minimum acute oral toxicity and acute percutaneous toxicity, respectively, in both cases for rats [7].

Table 1. Average ranks for the aqueous persistence as determined by the solubility, the biodegradation potential and the Henry's Law Constants for a series of OP insecticides and VX (the compound ID refers to the FADINAP database, cf. the above text; na: not available)

| <i>Compound</i> | <i>Average Rank</i> <i>Rk_{av}</i> | <i>Acute Oral</i> <i>Toxicity (mg/kg)</i> | <i>Acute Percutaneous</i> <i>Toxicity (mg/kg)</i> |
|---------------------|---|--|--|
| Anilofos | 20.5 | 472 | >2000 |
| Azinphos methyl | 25.6 | 4 | 220 |
| Chlorfenvinphos | 9.6 | 24 | 31 |
| Chlorpyrifos methyl | 18.2 | 1630 | >3700 |
| Dialifos | 41 | 5 | na |
| Dicrotophos | 9.1 | 17 | 110 |
| Ditalimfos | 19.3 | 5660 | >2000 |
| Monocrotophos | 10.3 | 20 | 112 |
| Phosalone | 35.1 | 135 | >1500 |
| Phosmet | 21.9 | 160 | na |
| Phosphamidon | 6.2 | 17.9 | 374 |
| Pyraclufos | 18.9 | 237 | >2000 |
| VX | 5.3 | 0.088 | 0.1 |

It is immediately seen that although the compounds were placed on the same level in the Hasse diagram, only through the analysis of average linear rank the true identity of the single compounds are disclosed. Thus, in the present case it is obvious that VX ($Rk_{av} = 5.3$) that in the present context is the unknown compound achieves an identity that can be compared to Phosphamidon ($Rk_{av} = 6.2$) as the closest counterpart. Thus, with regard to aqueous persistence, the above combined QSAR and partial order ranking analysis indicates that VX and Phosphamidon will display close to identical behavior. This further means that Phosphamidon, within the present set of compounds included in the investigation, appears as the optimal substitute for VX in experimental studies where aqueous persistence is a crucial parameter. It is noted that the acute oral toxicity associated with Phosphamidon is approximately 200 times lower than that of VX and in the case of acute percutaneous toxicity, Phosphamidon appears to be nearly 4000 times less toxic than VX.

Conclusions

The present study has demonstrated how 'unknown' compounds may obtain an identity by comparing to structurally related, experimentally well-characterized structurally similar compounds. The identity can be established by a close interplay between so-called "noise-deficient" QSARs, in the present study generated using the EPI Suite as the modeling onset. Subsequently, the generated physico-chemical end-points are used as descriptors in a partial order based ranking and the subsequent analysis of the average linear rank. It is suggested that experimentally well-characterized compounds may serve as substitutes for highly toxic compounds, such as the nerve agent in experimental studies without exhibiting the same extreme toxicity, however from an overall viewpoint exhibit analogous environmental characteristics.

References and Notes

1. EINECS (European Inventory of Existing Commercial Chemical Substances). cf. European Commission 1967: Directive 67/548/EEC on the application of laws, regulations and administrative provisions relating to the classification, packaging and labeling of dangerous substances and the 6th amendment: Directive 79/831/EEC; art. 13
2. Niëmela, J. Working document on the availability of data for classification and labelling of chemical substances at the European market, 1994
3. Walker, J.D.; Carlsen, L.; Hulzebos, E.; Simon-Hettich, B. Government Applications of Analogues, SARs and QSARs to Predict Aquatic Toxicity, Chemical or Physical Properties, Environmental Fate Parameters and Health Effects of Organic Chemicals, *SAR QSAR Environ. Res.* **2002**, *13*, 607-619
4. Carlsen, L. A QSAR Approach to Physico-Chemical Data for Organophosphates with Special Focus on Known and Potential Nerve Agents. *Submitted for publication*
5. Carlsen, L. Partial Order Ranking of Organophosphates with Special Emphasis on Nerve Agents. *Commun. Math. Comp. Chem.-MATCH*, in press
6. Pollution Prevention (P2) Framework, EPA-758-B-00-001; may be obtained through the link 'P2 Manual 6-00.pdf' found at <http://www.epa.gov/pbt/framework.htm>, US EPA
7. FADINAP, *Database on pesticide and the environment*, <http://www.fadinap.org/pesticide/>
8. Connell, D.W.; Hawker, D.W. Use of polynomial expressions to describe the bioconcentration of hydrophobic chemicals in fish. *Ecotox. Environ. Safety* **1988**, *16*, 242-257
9. Davey, B.A.; Priestley, H.A. *Introduction to Lattices and Order*; Cambridge University Press: Cambridge, UK, 1990
10. Carlsen, L.; Sørensen, P.B.; Thomsen, M. Partial order ranking based QSAR's: Estimation of solubilities and octanol-water partitioning. *Chemosphere* **2001**, *43*, 295-302
11. Brüggemann, R.; Pudenz, S.; Carlsen, L.; Sørensen, P.B.; Thomsen, M.; Mishra R.K. The use of Hasse diagrams as a potential approach for inverse. QSAR, *SAR QSAR Environ. Res.* **2001**, *11*, 473-487
12. Carlsen, L.; Sørensen, P.B.; Thomsen, M.; Brüggemann, R. QSAR's Based on Partial Order Ranking. *SAR and QSAR Environ. Res.* **2002**, *13*, 153-165
13. Carlsen, L.; Walker, J.D. QSARs for Prioritizing PBT Substances to Promote Pollution Prevention. *QSAR Comb. Sci.* **2003**, *22*, 49-57
14. Hasse, H. *Über die Klassenzahl abelscher Zahlkörper*; Akademie Verlag: Berlin, 1952
15. Halfon, E.; Reggiani, M.G. On the ranking of chemicals for environmental hazard. *Environ. Sci. Technol.* **1986**, *20*, 1173-1179
16. Brüggemann, R.; Halfon, E.; Welzl, G.; Voigt, K.; Steinberg, C.E.W. Applying the concept of partially ordered sets on the ranking of near-shore sediments by a battery of tests. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 918-925
17. Brüggemann, R.; Halfon, E.; Bücherl, C. Theoretical base of the program "Hasse", GSF-Bericht 20/95, Neuherberg, 1995; the software may be obtained by contacting Dr. R. Brüggemann, Institute of Freshwater Ecology and Inland Fisheries, Berlin
18. Fishburn, P.C. On the family of linear extensions of a partial order. *J. Combinat. Theory* **1974**, *17*, 240-243

19. Graham, R.L. Linear Extensions of Partial Orders and the FKG Inequality. In *Ordered Sets*; Rival, I (ed.); D. Reidel Publishing Company: Dordrecht (The Netherlands), **1982**; pp. 213-236
20. Winkler, P.M. Average height in a partially ordered set. *Discrete Mathematic.* **1982**, *39*, 337-341
21. Winkler, P.M. Correlation among partial orders. *Siam J Alg Disc Meth.* **1983**, *4*, 1-7
22. Brüggemann, R.; Lerche, D.; Sørensen, P.B.; Carlsen, L. Estimation of average ranks by a local partial order model, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 618-625
23. Lerche, D.; Brüggemann, R.; Sørensen, P.; Carlsen, L.; Nielsen O.J. A comparison of partial order technique with three methods of multi-criteria analysis for ranking of chemical substances, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1086-1098

© 2004 by MDPI (<http://www.mdpi.org>). Reproduction is permitted for noncommercial purposes.