Novel computational methods for estimation of physical-chemical properties of PCNs

Tomasz Puzyn[1], Jerzy Falandysz[1]

[1]University Of Gdansk

Introduction

Polychlorinated naphthalenes (PCNs) are ubiquitous pollutants which exhibit significant toxicity and diverse physical-chemical properties determining their environmental fate and distribution in various compartments [1-4]. Depending on values of partition coefficients, subcooled liquid vapor pressure, water solubility and some other parameters chloronaphthalene (CN) congeners are long-range transported in the atmosphere [5-8]. Based on that phenomena the multicompartmental models of environmental transport and fate of PCNs could be constructed [9,10]. Nevertheless, to predict their environmental behavior the properties mentioned should be determined for each of 75 CN congeners.

A simple and cost-effective strategy, compared to the robust experimental study, is the computational estimation of the properties by employment of available published laboratory data and novel numerical techniques. Implementation of QSPR (Quantitative Structure-Property Relationships) technique enables predictions based on the molecular information encrypted by means of the structural descriptors, *i.e*. quantum-chemical descriptors, topological descriptors, shape descriptors and others [11-16]. The QSPR approach uses many of chemometrical methods such as simple and multiple regression, component regression, and artificial intelligence techniques such as neural networks and genetic algorithms [17-19].

The aim of this study was to evaluate the most commonly used six chemometrical methods in computational estimation of the four key environmental physical-chemical properties of PCNs.

Materials and Methods

Initially a set of structural descriptors was computed for each of 75 CN congeners and based on the level of the density functional theory using B3LYP hybrid functional and 6-311++G** basis set [20]. In the second step, published experimental data on common logarithm of n-octanol/air partition coefficient (log K_{OA}), logarithm of n-octanol/water partition coefficient (log K_{OW}), subcooled liquid vapor pressure (log P_L) and water solubility (log S_W) of chloronaphthalenes were collected [2,21-23]. Next, the QSPR models were constructed for each of the properties separately by means of six chemometrical methods such as: simple regression method (SRM), principal component regression (PCR), partial least square regression (PLS), partial least square regression with initial elimination of the uninformative variables (UVE-PLS), partial least square regression with variable selection using a genetic algorithm (GA-PLS), and neural networks with variable selection using a genetic algorithm (GA-NN). For each property, a set of congeners for which experimental data are available was divided into training and validation subsets. Each model was developed using the same training set and each validated using the same validation set allowing quantitative comparison of the models. The models were compared taking into account the predictive ability of the model, measured by means of the root mean square error of prediction (RMSEP) for the validation set, and the complexity of the model, expressed as the number of independent variables (descriptors) used.

Results and Discussion

The main two parameters which characterize the models are presented in Table 1 and Table 2.

Table 1. Predictive ability of the models (value of RMSEP)

Method	Parameter			
	Log K_{OA}	Log K_{OW}	Log P_L	Log S_W

SRM	0.273	0.317	0.192	0.260
PCR	0.150	0.216	0.200	0.205
PLS	0.140	0.176	0.289	0.202
UVE-PLS	0.132	0.162	0.132	0.204
GA-PLS	0.106	0.146	0.108	0.222
GA-NN	0.091	0.065	0.078	0.155

Table 2. Complexity of the models (number of independent variables)

Method	Parameter			
	Log K_{OA}	Log K_{OW}	Log P_L	Log S_W
SRM	1	1	1	1
PCR	33	33	26	33
PLS	33	33	26	33
UVE-PLS	20	5	13	29
GA-PLS	9	8	8	7
GA-NN	6	12	17	6

Usually (except for log P_L), the models obtained by means of SRM were characterized by the highest values of RMSEP. The methods based on factor regression (PCR and PLS) shown similar predictive ability, while PLS gave not much lower values of the prediction error than PCR. An implementation of algorithms that reduce the number of independent variables, such as UVE and GA, in PLS significantly improved the results. It is worthy to note, that usually use of GA-PLS leads to better model (lower RMSEP), than UVE-PLS. In all cases, neural networks with variable selection by means of a genetic algorithm (GA-NN) gave the lowest error of prediction.

By definition, the SRM models are characterized by the lowest complexity. The commonly used QSPR methods such as PCR and PLS require a large number of descriptors, and what generates additional time and costs of computation. Use of the techniques, which eliminate uninformative variables in the initial step (UVE-PLS and GA-PLS), results both in decrease of complexity and usually increase the predictive ability of the model. In contrast to UVE, implementation of GA leads to models based on lower number of independent variables (except log K_{OW}).

The optimal number of variables and the lowest error of prediction was for the models obtained from GA-NN, suggesting this method as the most practical and useful for the novel computational prediction of physical-chemical properties of chloronaphthalenes.

Acknowledgements

This study was supported by the Polish Ministry of Science and Informatics under grant no. KBN 1128/T09/2003/24. Computations were carried out using computers in the TASK - Academic Computer Center in Gdańsk.

References

1. Harner T. and Bidleman T.F. (1997) *Atmos. Environ.*, 31/32: 4009-4016.
2. Harner T. and Bidleman T.F. (1998) *J. Chem. Eng. Data*, 43: 40-46.
3. Falandysz J. (2003) *Food Addit. Contam.*, 20: 995-1014.
4. Falandysz J. (1998) *Environ. Pollut.*, 10: 77-90.
5. Harner T., Bidleman T.F., Jantunen L.M.M. and Mackay D. (2001) *Environ. Toxicol. Chem.*, 20: 1612-1621.
6. Wang Y.H. and Wong P.K. (2002) *Water Res.*, 36: 350-355.
7. Wania F. (1999) *Environ Sci. Pollut. Res.*, 6: 11-19.
8. Wania F. (2003) *Environ. Sci. Technol.*, 37: 1344-1351.
9. Wania F. and Mackay D. (1995) *Sci. Total. Environ*, 160/161: 211-232.
10. Wania F. and Mackay D. (1996) *Environ. Sci. Technol.*, 30: 390A-396A.
11. Chen J., Harner T., Schramm K.W., Quan X., Xue X., Wu W.Z. and Kettrup A. (2003) *Comput. Biol. Chem.*, 27: 405-421.
12. Chen J., Xue X., Schramm K.-W., Quan X., Yang F. and Kettrup A. (2003)

Comput. Biol. Chem., 27: 165-171.
13. Gramatica P., Consolaro F. and Pozzi S. (2001) *Chemosphere*, 43: 655-664.
14. Falandysz J., Puzyn T., Szymanowska B., Kawano M., Markuszewski M., Kaliszan R., Skurski P. and Błażejowski J. (2001) *Pol. J. Environ. Std.*, 10: 217-235.
15. Puzyn T. and Falandysz J. (2005) *Atmos. Environ.*, 39: 1439-1446.
16. Puzyn T. and Falandysz J. (2003) *J. Environ. Sci. Health*, 38A: 1761-1780.
17. Sharaf M.A., Illman D.H. and Kowalski.B.R. (1986) Chemometrics, John Wiley & Sons Inc.
18. Mazerski J. (2000) Podstawy chemometrii, Wydawnictwo Politechniki Gdańskiej, Gdańsk.
19. Taskinen J. and Yliruusi J. (2003) *Adv. Drug Deliv. Rev.*, 55: 1163-1183.
20. Frisch M.J., et al. (2003) GAUSSIAN 03, Gaussian Inc., Pittsburgh.
21. Lei Y.D., Wania F. and Shiu W.Y. (1999) *J. Chem. Eng. Data*, 44: 577-582.
22. Opperhuizen A., van der Velde E.W., Gobas F.A.P.C., Liem D.A.K. and van der Steen J.M.D. (1985) *Chemosphere*, 14: 1871-1896.
23. Su Y., Lei Y.D., Daly G. and Wania F. (2002) *J. Chem. Eng. Data*, 47: 449-455.