Chemosphere 79 (2010) 887–890

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

New QSPR equations for prediction of aqueous solubility for military compounds

Eugene N. Muratov[a], Victor E. Kuz'min[a], Anatoly G. Artemenko[a], Nikolay A. Kovdienko[b], Leonid Gorb[c], Frances Hill[d], Jerzy Leszczynski[b,d,*]

[a] Laboratory of Theoretical Chemistry, Department of Molecular Structure, A.V. Bogatsky Physical–Chemical Institute National Academy of Sciences of Ukraine, Lustdorfskaya Doroga 86, Odessa 65080, Ukraine
[b] Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Jackson State University, Jackson, MS 39217, USA
[c] SpecPro, Inc., Vicksburg, MS 39180, USA
[d] US Army ERDC, Vicksburg, MS 39180, USA

ARTICLE INFO

Article history:
Received 4 February 2010
Accepted 11 February 2010
Available online 16 March 2010

Keywords:
QSPR
SiRMS
EPI Suite™
Hazardous nitrocompounds

ABSTRACT

The development of a new quantitative structure–property relationship (QSPR) model to predict aqueous solubility (S_w) accurately for compounds of military interest is presented. The ability of the new model to predict solubility is assessed and compared to available experimental data. A large set of structurally diverse organic compounds was used in this analysis. SiRMS methodology was employed to develop PLS models based on 135 training compounds and predictive accuracy was tested for 155 compounds selected for that purpose. The use of descriptors calculated only from the 2D level of representation of molecular structure produces a well-fitted and robust QSPR model ($R^2 = 0.90$; $Q^2 = 0.87$). Predictive ability for the model produced in this study on external test set ($R^2_{test} = 0.81$) is comparable to the predictive ability of EPI Suite™ 4.0. Consensus solubility predictions using SiRMS and EPI models for 25 compounds of military interest (not included into the training set) have been completed.

Published by Elsevier Ltd.

1. Introduction

Information on the solubility of new and emerging compounds is an important factor for environmental risk assessment, providing data for the modeling of transport and fate of chemical compounds and for understanding the pharmacokinetic behavior of contaminants in living organisms. The manufacturing, storage, transportation and utilization of munitions can lead to the release of nitro- and nitroso-compounds into the environment. These compounds and their metabolites may have long-term environmental impact. In many cases aqueous solubility (S_w) of new and emerging chemicals is the determinative property for the estimation of the environmental impact of these compounds and drives research into remediation techniques. However, S_w experimental data, particularly on military crucial contaminants, often are not available and are predicted based on existing QSPR methodologies.

Many attempts have been made to estimate S_w values using QSPR techniques beginning from late 1970s (Dearden, 2006). Since that time, many new methods have been proposed and the prediction accuracy and coverage have been significantly improved. Methods for estimation of the S_w values can be divided into two classes (Dearden, 2006):

- The first type of methods is classified as a 'substructure' approach, where a molecular structure is represented by atoms (atom contribution methods) or fragments (group contribution methods) and S_w values are obtained as additive sum of contributions from each atom or fragment of the molecules.
- The 'whole structure' approaches on the other hand, apply descriptors like molecular lipophilicity potentials (MLP), topological indices and/or global molecular features to the calculation of S_w. Lesser amounts of descriptors are used in 'whole structure' approaches in comparison with 'substructure' approaches. However, these kind of methods need experimental correction parameters (like melting point or boiling point), or complex correction terms.

Thus, 'substructure' approaches are generally more practical because they allow S_w to be calculated directly from the chemical structure. Often these approaches are based on the assumption that the properties of a molecule (or compound) can be represented as an additive sum of the properties of their structural

[*] Corresponding author. Address: Interdisciplinary Center of Nanotoxicity, Jackson State University, Department of Chemistry, 1400 J.R. Lynch Street, Jackson, MS 39217, USA. Tel.: +1 601 9797824; fax: +1 601 9797823.
E-mail address: jerzy@ccmsi.us (J. Leszczynski).

0045-6535/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.chemosphere.2010.02.030

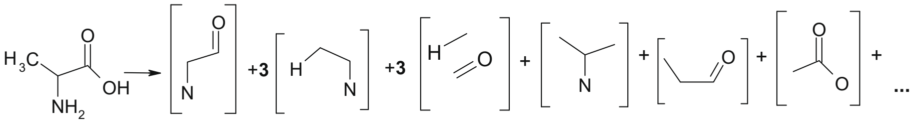

Fig. 1. An example of the Simplex Representation of Molecular Structure for alanine.

fragments. This technique is utilized by the Environmental Protection Agency[1] in the EPI Suite™ 4.0 (Estimation Programs Interface) to estimate S_W. However, in fact solubility does not obey an additive scheme. This is the main reason why the accuracy of estimations of this parameter for selected molecules produced by different currently-implemented QSPR methods is frequently insufficient.

We expect that the Simplex Representation of Molecular Structure (SiRMS), which avoids additive contributions of structural fragments owing to usage of non-bonded simplexes (Kuz'min et al., 2008a), will be able to overcome the above described limitations and provide accurate QSPR predictions of S_W.

The main purpose of this study is to develop new QSPR models to accurately predict S_W values for compounds of military interest (explosives and their degradation intermediates) using the SiRMS approach, with subsequent validation of results using a broad spectrum of available experimentally determined data.

2. Materials and methods

Aqueous solubility[2] in pure water (S_W) (Kholod et al., 2009) (represented in mol L^{-1}, and log S_W) was the characteristic property investigated in this study. All compounds' names, their CAS registry number and corresponding observed log S_W values are shown in Supplementary materials. The work set used in this study consisted of: (1) a training set of 135 diverse compounds, (2) a test set of 155 compounds (Klampt et al., 2002) for external validation of obtained "structure–solubility" models, and (3) a set of compounds of military interest for aqueous solubility prediction. In both the training and the test sets a number of nitrocompounds (including some compounds of military interest) were included (12 and 9, respectively). This addition will give the authors an opportunity to develop well-fitted and robust models of aqueous solubility for militarily important nitrocompounds.

In the framework of SiRMS (Kuz'min et al., 2008a) any molecule can be represented as a system of different simplexes (tetratomic fragments of fixed composition, structure, chirality and symmetry). At the 2D level, the connectivity of the atoms in a simplex, atom type and bond nature (single, double, triple, aromatic) have been considered. The usage of several variants of simplex vertices (atoms) differentiation represents an important aspect of SiRMS (Fig. 1). We hypothesize that the specification of atoms by their chemical identity alone (i.e., C, N or O), that is implemented in many traditional QSPR approaches, limits the possibilities of pharmacophore fragment selection (Kuz'min et al., 2008a).

Many simplex descriptors have been generated in the SiRMS model. The PLS-method proved to be efficient for models with a great number of parameters (Rannar et al., 1994). The PLS regression model may be written as (Rannar et al., 1994)

$$Y = b_0 + \sum_{i=1}^{N} b_i x_i, \tag{1}$$

where Y is an appropriate activity, b_i is PLS regression coefficients, x_i is the ith descriptor value, N is the total number of descriptors. PLS can analyze any number of x-variables (K) regardless of the number of objects (N) (Rannar et al., 1994).

Workflow of the most relevant descriptors selection in PLS modeling (Artemenko et al., 2009) consists of the following: elimination of non-significant and highly correlated descriptors → trend-vector procedure → automatic variable selection ↔ genetic algorithm → partial or complete enumeration methods → best QSAR model. Selection of the best QSAR model at every stage of the scheme was carried out by maximizing the fitness function (FF) criterion, where FF = $R^2 + 2Q^2$ and FF → max, i.e., the best selected QSAR model is the model with the maximum FF value. Additionally, in order to avoid chance correlations which are possible because of large number of generated descriptors, y-scrambling test with 1000 randomization rounds was applied by the same scheme.

Each QSAR model has its own "domain applicability" (DA) in the space of structural features (descriptors). It's evident that predictions for new compounds which are structurally very different from the training set structures are not very reliable. Leverage procedure and ellipsoid DA approach (Kuzmin et al., 2008a) were used in the given work for DA estimation. Predictions for new compounds were taken into account only in the case where it belongs to DA estimated both approaches.

Though the SiRMS method is novel, it has been employed successfully in several studies to differentiate "structure–activity" relationships (Kuz'min et al., 2005, 2007, 2008a,b; Artemenko et al., 2007; Artemenko et al., 2009). SiRMS methodology does not have many of the restrictions encountered in well-known and widely used approaches such as CoMFA (Cramer et al., 1988), CoMSIA (Klebe et al., 1994), and HASL (Doweyko, 1988); the applications of the latter are limited only to the structurally homogeneous set of molecules. SiRMS approach is similar to HQSAR (Seel et al., 1999) but has not its restrictions (consideration of atom type only) and deficiencies (an ambiguity of descriptor formation during the hashing of molecular holograms). Furthermore, as compared to HQSAR, different physicochemical properties of atoms can be taken into account in SiRMS.

3. Results and discussion

In the present study, 2D simplex descriptors for representations of molecular structure were utilized in the prediction of S_W. At the initial step, 9701 simplex descriptors were generated. Various physicochemical atomic characteristics were used for atom differentiation: atom types, partial charge (Jolly and Perry, 1973), lipophilicity (Wang et al., 1997), refraction (Ioffe, 1983), and the ability of an atom to act as a donor/acceptor in hydrogen-bond formation. In this study all atoms were divided into groups corresponding to their partial charge $A \leqslant -0.05 < B \leqslant 0 < C \leqslant 0.05 < D$, lipophilicity $A \leqslant -0.5 < B \leqslant 0 < C \leqslant 0.5 < D$ and refraction $A \leqslant 1.5 < B \leqslant 3 < C \leqslant 8 < D$.

The QSPR model was generated using a training set of 135 molecules (Table S1). Y-scrambling test repeated 1000 times revealed the absence of chance correlations ($Q^2_{YS} = 0.27 \pm 0.01$). After PLS analysis, a well-fitted and robust 2D QSPR model ($R^2 = 0.90$; $Q^2 = 0.87$) was obtained for more suitable simplex descriptors.

The validation of the model was performed by exploring the associated external test set that consists of 155 compounds presented in Table S2. This validation indicates that a high level of confidence in the prediction ($R^2_{test} = 0.81$) has been established. It

[1] http://www.epa.gov/oppt/exposure/pubs/episuite.htm.

[2] Kindly provided by Prof. A. Klamt, Dr. F. Eckert, from COSMO*logic* GmbH & Co. KG.