Chemosphere 90 (2013) 877–880

Contents lists available at SciVerse ScienceDirect

Chemosphere

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

Short Communication

CORAL: QSPR model of water solubility based on local and global SMILES attributes

Andrey A. Toropov [a,*], Alla P. Toropova [a], Emilio Benfenati [a], Giuseppina Gini [b], Danuta Leszczynska [c], Jerzy Leszczynski [d]

[a] Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy
[b] Department of Electronics and Information, Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy
[c] Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch St., Jackson, MS 39217-0510, USA
[d] Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 J.R. Lynch St, P.O. Box 17910, Jackson, MS 39217, USA

HIGHLIGHTS

- ▶ The CORAL software for the building up of QSPR/QSAR models is suggested.
- ▶ The SMILES is used as the representation of the molecular structure.
- ▶ The CORAL model for water solubility is described in detail.

ARTICLE INFO

Article history:
Received 20 April 2012
Received in revised form 16 July 2012
Accepted 19 July 2012
Available online 23 August 2012

Keywords:
Water solubility
QSPR
Monte Carlo technique
CORAL software

ABSTRACT

Water solubility is an important characteristic of a chemical in many aspects. However experimental definition of the endpoint for all substances is impossible. In this study quantitative structure–property relationships (QSPRs) for negative logarithm of water solubility–log S (mol L^{-1}) are built up for five random splits into the sub-training set (≈55%), the calibration set (≈25%), and the test set (≈20%). Simplified molecular input-line entry system (SMILES) is used as the representation of the molecular structure. Optimal SMILES-based descriptors are calculated by means of the Monte Carlo method using the CORAL software (http://www.insilico.eu/coral). These one-variable models for water solubility are characterized by the following average values of the statistical characteristics: $n_{\text{sub\_train}} = 725–763$; $n_{\text{calib}} = 312–343$; $n_{\text{test}} = 231–261$; $r^2_{\text{sub\_train}} = 0.9211 \pm 0.0028$; $r^2_{\text{calib}} = 0.9555 \pm 0.0045$; $r^2_{\text{test}} = 0.9365 \pm 0.0073$; $s_{\text{sub\_train}} = 0.561 \pm 0.0086$; $s_{\text{calib}} = 0.453 \pm 0.0209$; $s_{\text{test}} = 0.520 \pm 0.0205$. Thus, the reproducibility of statistical quality of suggested models for water solubility confirmed for five various splits.

1. Introduction

The solubility of liquids and solids in water is a very important molecular property that affects their biological activity (Huuskonen, 2000; Tetko et al., 2001; Roy and Saha, 2003; Yan and Gasteiger, 2003). Quantitative structure – property/activity relationships (QSPRs/QSARs) based on various molecular descriptors (Furtula and Gutman, 2011; Melagraki and Afantitis, 2011; Mullen et al., 2011; Ojha et al., 2011) are a possible tool to predict physicochemical properties (Huuskonen, 2000; Tetko et al., 2001; Yan and Gasteiger, 2003) as well as biological activity (Marino et al., 2002; Toropov and Toropova, 2002; Peruzzo et al., 2003; Melagraki and Afantitis, 2011; Mullen et al., 2011; Ojha et al., 2011) for substances which have not been examined in the experiment.

Recently, the CORAL software (http://www.insilico.eu/coral) has been suggested as a tool of the QSPR/QSAR analyses of various endpoints (Toropov et al., 2011; Toropova et al., 2011a,b,c). The software is building up models for various endpoints with representation of the molecular structure by simplified molecular input-line entry system (SMILES) (Weininger, 1990). The aim of the present study is the estimation of the software as a tool to build up QSPR models of water solubility.

2. Method

Data on water solubility of 1311 substances, i.e. their CAS number, SMILES, and values of negative logarithm of water solubility – log S (mol L^{-1}) were taken from the web site of Virtual Computational Chemistry Laboratory (http://www.vcclab.org/lab/alogps/). These substances were distributed by means of five random splits into the sub-training set (≈55%), calibration set (≈25%), and test set (≈20%).

[*] Corresponding author.
E-mail address: andrey.toropov@marionegri.it (A.A. Toropov).

0045-6535/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.chemosphere.2012.07.035

Table 1
Definitions of the *BOND*, *NOSP*, and *HALO* attributes.

=	#	@	Comments
Calculation of the BOND index			
0	0	0	There are no double, triple, or stereo chemical bonds
0	0	1	The molecule contains only stereo chemical bonds
0	1	0	The molecule contains only triple bonds
0	1	1	The molecule contains triple and stereo chemical bonds
1	0	0	The molecule contains only double bonds
1	0	1	The molecule contains double bonds and stereo chemical bonds
1	1	0	The molecule contains double and triple bonds
1	1	1	The molecule contains double, triple, and stereo chemical bonds

N	O	S	P	Comments
Calculation of the NOSP index				
0	0	0	0	Nitrogen, oxygen, sulfur, and phosphorus are absent
0	0	0	1	The molecule contains only phosphorus
0	0	1	0	The molecule contains only sulfur
0	0	1	1	The molecule contains sulfur and phosphorus
0	1	0	0	The molecule contains only oxygen
0	1	0	1	The molecule contains oxygen and phosphorus
0	1	1	0	The molecule contains oxygen and sulfur
0	1	1	1	The molecule contains oxygen, sulfur, and phosphorus
1	0	0	0	The molecule contains only nitrogen
1	0	0	1	The molecule contains nitrogen and phosphorus
1	0	1	0	The molecule contains nitrogen and sulfur
1	0	1	1	The molecule contains nitrogen, sulfur, and phosphorus
1	1	0	0	The molecule contains nitrogen and oxygen
1	1	0	1	The molecule contains nitrogen, oxygen and phosphorus
1	1	1	0	The molecule contains nitrogen, oxygen, and sulfur
1	1	1	1	The molecule contains nitrogen, oxygen, sulfur, and phosphorus

F	Cl	Br	Comments
Calculation of the HALO index			
0	0	0	Fluorine, chlorine and bromine are absent
0	0	1	The molecule contains only bromine
0	1	0	The molecule contains only chlorine
0	1	1	The molecule contains chlorine and bromine
1	0	0	The molecule contains only fluorine
1	0	1	The molecule contains fluorine and bromine
1	1	0	The molecule contains fluorine and chlorine
1	1	1	The molecule contains fluorine, chlorine, and bromine

Table 2
Example of calculation DCW(1,35) for SMILES = "CC(N)=O" DCW(1,35) = 14.2270.

Structural attribute (*SA*)	*W*(*SA*)
S_k	
`C..........`	−0.5615
`C..........`	−0.5615
`(..........`	−2.6250
`N..........`	−1.3760
`(..........`	−2.6250
`=..........`	−1.2520
`O..........`	1.0665
SS_k^*	
`C...C.......`	−4.1925
`C...(.......`	−0.4385
`N...(.......`	−0.8165
`N...(.......`	−0.8165
`=...(.......`	−0.0635
`O...=.......`	−0.0675
SSS_k^*	
`C...C...(...`	1.3740
`N...(...C...`	3.2500
`(...N...(...`	4.9395
`N...(...=...`	−3.5605
`O...=...(...`	−0.8770
`NOSP1100`	13.3125
`HAL00000`	6.1845
`BOND100`	3.9335

* The bracket of '(' is inserted instead of ')', since both brackets are representation of the same phenomenon of branching of molecular skeleton; SS_k and SSS_k are ordered according ASCII codes in order to avoid situation where the same attributes are indicated by two various manner (e.g. AB and BA or ABC and CBA), thus instead of '`(...N.......`' and '`N...).......`' we have the only '`N...(.......`'.

The SMILES-based optimal descriptors were calculated with scheme developed for QSAR models of toxicity in rats (Toropov et al., 2011):

$$DCW(T, N_{epoch}) = \sum W(S_k) + \sum W(SS_k) + \sum W(SSS_k) + W(BOND) + W(NOSP) + W(HALO) \quad (1)$$

where S_k, SS_k, and SSS_k are local SMILES attributes (fragments) which are involving one, two, and three SMILES element, respectively. The SMILES element can be one symbol, e.g. '*C*', '*c*', '*N*', '=', '#', etc., or several symbols which cannot be considered separately, e.g. 'Cl', 'Br', '11%', etc. (Weininger, 1990); *BOND*, *NOSP*, and *HALO* are global molecular features which are calculated with SMILES (Toropova et al., 2011d). Table 1 shows the schemes of calculation of *BOND*, *NOSP*, and *HALO*.

The descriptors for each substance are calculated with the correlation weights, i.e. $W(S_k)$, $W(SS_k)$, $W(SSS_k)$, *W*(*BOND*), *W*(*NOSP*), and *W*(*HALO*). The numerical values for the correlation weights are calculated with the Monte Carlo method optimization procedure. The target function (TF) of the procedure is the following (Toropov et al., 2010):

$$TF = R + R' - |R - R'| \cdot R_w - (|C_o - C'_0| + |C_1 - C'_1|) \cdot R_c \quad (2)$$

where *R* and *R*' are correlation coefficients between $DCW(T{,}N_{epoch})$ and an endpoint for sub-training set and calibration set, respectively; C_0, C_1, C'_0, and C'_1 are regression coefficients for the sub-training set and calibration set, respectively; $R_w = 0.1$ and $R_c = 0.01$ are empirical constants; *T* is the threshold in order to classify SMILES attributes into two categories: rare (noise) and active (i.e. not rare). Correlation weights for rare attributes are assumed equal to zero, i.e. they are not involved in the modeling process; N_{epoch} is the number of epochs of the Monte Carlo optimization. In the present study $T = 1$ and $N_{epoch} = 35$ were used. Table 2 contains example of representation of molecular structure by the described local and global attributes extracted from SMILES.

3. Results and discussion

Table 3 contains the statistical quality of models of water solubility for five various splits into the sub-training set, calibration set, and test set. These splits have been selected by taking into account the measure of their identity expressed as percentage (Table 4). The identity of two splits is calculated as ratio of the number of identical substances which have the same status for a couple splits to total number of compounds. Two substances are identical if they have the same status in two splits, i.e. both are in sub-training set (or both are in the calibration set or both are in the test set). Table 4 contains the identity for all pairs of five splits examined in this study. It should be noted there are not pairs of splits with the identity larger than 45%. Studies of groups of various splits into the training and test sets gradually become a general principle of the QSPR/QSAR analyses (Roy et al., 2008; Puzyn et al., 2011). We deem that suggested principle of maximal dissimilarity of splits can be used for the QSPR/QSAR analyses as an alternative of existing

Table 3
Comparison of the statistical quality of the CORAL models for water solubility for five various splits with the statistical quality of four models for water solubility which are described in the literature.

Split	N_{act}	n_{sub_train}	$r^2_{sub\_train}$	s_{sub_train}	F_{sub_train}	n_{calib}	r^2_{calib}	s_{calib}	N_{test}	r^2_{test}	s_{test}	R^2_m
1	712	736	0.9231	0.565	8814	314	0.9543	0.453	261	0.9381	0.511	0.8963
2	703	725	0.9230	0.560	8666	343	0.9493	0.479	243	0.9303	0.530	0.9144
3	731	728	0.9173	0.574	8051	324	0.9614	0.425	259	0.9412	0.508	0.9158
4	722	763	0.9242	0.548	9278	312	0.9526	0.473	236	0.9263	0.554	0.9024
5	724	756	0.9182	0.557	8462	324	0.9600	0.435	231	0.9465	0.495	0.9403

Reference	n_{sub_train}	$r^2_{sub\_train}$	s_{sub_train}	F_{sub_train}	n_{calib}	r^2_{calib}	s_{calib}	N_{test}	r^2_{test}	s_{test}
Huuskonen (2000)	884	0.94	0.47	–	413	0.92	0.60	21	0.91	0.63
Tetko et al. (2001)	879	0.95	0.47	–	412	0.92	0.60	21	0.90	0.64
Liu and So (2001)	1033	0.86	0.70	–	258	0.86	0.71	21	0.79	0.93
Yan and Gasteiger (2003)	797	0.93	0.50	–	496	0.92	0.59	21	0.85	0.77

N_{act} is the number of SMILES attributes which are involved in the modeling process; n is the number of substances in a set; r^2 is square of correlation coefficient; s is mean square error; F is Fischer F-ratio; R^2_m is the metric of predictability: R^2_m should be larger than 0.5 (Ojha et al., 2011).

Table 4
The identity (%) of pairs of splits. The identity is defined as identity (%) = 100 * (number of identical substances/1311).

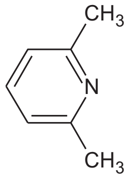
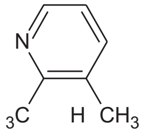
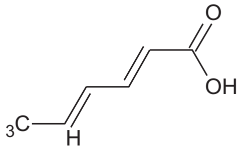
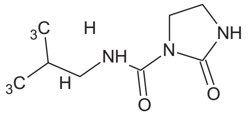
	Split 1	Split 2	Split 3	Split 4	Split 5
Split 1	100	38.8	41.6	43.1	44.4
Split 2	38.8	100	40.7	41.6	42.3
Split 3	41.6	40.7	100	44.2	43.0
Split 4	43.1	41.6	44.2	100	43.7
Split 5	44.4	42.3	43.0	43.7	100

algorithms of the splitting of data into the training and test sets (Roy et al., 2008; Puzyn et al., 2011).

Table 3 contains the statistical characteristics of the models (which are calculated with the same data) for water solubility obtained with the CORAL software together with the statistical characteristics of the models described in the literature (Huuskonen, 2000; Liu and So, 2001; Tetko et al., 2001 and Yan and Gasteiger, 2003). Statistical characteristics of models for other data on water solubility are: (i) various organic compounds, $n = 193$, $r^2 = 0.946$ (Roy and Saha, 2003); (ii) drug-like compounds, $n_{train} = 97$, $r^2_{train} = 0.759$, $n_{test} = 48$, $r^2_{test} = 0.719$ (Duchowicz et al., 2008); and (iii) perfluorinated chemicals: $n = 20$, $r^2 = 0.763$ (Bhhatarai and Gramatica, 2011). Comparison of the statistical quality of models calculated with the CORAL software and the above-mentioned models described in the literature shows that the CORAL models for water solubility are quite good. However, there are substances (Table 5) for which our models give poor prediction. We deem there are two indicators of the poor prediction: (i) symmetry and (ii) the possibility of intramolecular and intermolecular hydrogen bonds.

The CORAL software has been used as a tool of the QSPR/QSAR analyses of several endpoints (Toropov et al., 2010; García et al., 2011; Garro et al., 2011; Mullen et al., 2011; Toropova et al.,

Table 5
Examples of substances for which the CORAL software gives poorest prediction (outliers).

CAS	Structure	Split 1 Δlog S[a]	Split 2 Δlog S	Split 3 Δlog S	Split 4 Δlog S	Split 5 Δlog S
108-48-5		2.505[b]	2.390[b]	2.630[b]	2.413[b]	2.609[b]
583-61-9		2.178[b]	1.935[c]	2.267[b]	2.199[b]	2.356[b]
110-44-1		−1.421[c]	−1.781[b]	−1.807[c]	−1.624[b]	−1.780[b]
30979-48-7		−2.175[c]	−2.183[c]	−2.663[b]	−2.464[b]	−2.532[b]

[a] Δlog S = log S (experiment)−log S(calculated).

[b] Substance is in the sub-training set.

[c] Substance is in the calibration set.

2011a,b,c,d; Ibezim et al., 2012), however water solubility has been examined as target endpoint first time. We believe that presented results (Table 3) indicate that the CORAL can be used as a tool for the QSPR modeling of this endpoint.

4. Conclusions

The CORAL software can be used as a tool for QSPR analysis of the water solubility. We suppose that the reproducibility of the statistical quality of the models for five various splits into the sub-training set, calibration set, and test set is an important advantage of the suggested approach. The suggested measurement of identity for splits (Table 4) can be a criterion for practical definition of group of really different splits for a robust QSPR/QSAR analyses. Four substances are stable outliers for the CORAL models (Table 5).

Acknowledgement

We thank ANTARES (the Project number LIFE08-ENV/IT/00435), and the National Science Foundation (NSF/CREST HRD-0833178, and EPSCoR Award #:362492-190200-01/NSFEPS-090378) for financial support. Also we express our gratitude to Dr. L. Cappellini, Dr. G. Bianchi and Dr. R. Bagnati for valuable consultations on the computer sciences.

Appendix A. Supplementary material

Supplementary materials section contains five splits of examined compounds into the sub-training, calibration, and test sets and technical details of the CORAL method that was used to build up the models. One can check up reproducibility of described approach, using the supplementary materials and the CORAL software available on the Internet (http://www.insilico.eu/coral/). Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.chemosphere.2012.07.035.

References

Bhhatarai, B., Gramatica, P., 2011. Prediction of aqueous solubility, vapor pressure and critical micelle concentration for aquatic partitioning of perfluorinated chemicals environ. Sci. Technol. 45, 8120–8128.

Duchowicz, P.R., Talevi, A., Bruno-Blanch, L.E., Castro, E.A., 2008. New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorg. Med. Chem. 16, 7944–7955.

Furtula, B., Gutman, I., 2011. Relation between second and third geometric–arithmetic indices of trees. J. Chemom. 25, 87–91.

García, J., Duchowicz, P.R., Rozas, M.F., Caram, J.A., Mirífico, M.V., Fernández, F.M., Castro, E.A., 2011. A comparative QSAR on 1,2,5-thiadiazolidin-3-one 1,1-dioxide compounds as selective inhibitors of human serine proteinases. J. Mol. Graph. Model. 31, 10–19.

Garro, C., Martinez, J.C., Duchowicz, P.R., Estrada, M.R., Zamarbide, G.N., Castro, E.A., 2011. QSAR study and molecular design of open-chain enaminones as anticonvulsant agents. Int. J. Mol. Sci. 12, 9354–9368.

Huuskonen, J., 2000. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. J. Chem. Inf. Comput. Sci. 40, 773–777.

Ibezim, E., Duchowicz, P.R., Ortiz, E.V., Castro, E.A., 2012. QSAR on aryl-piperazine derivatives with activity on malaria. Chemom. Intell. Lab. 110, 81–88.

Liu, R.F., So, S.S., 2001. Development of quantitative structure–property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility. J. Chem. Inf. Comput. Sci. 41, 1633–1639.

Marino, D.J.G., Peruzzo, P.J., Castro, E.A., Toropov, A.A., 2002. QSAR carcinogenic study of methylated polycyclic aromatic hydrocarbons based on topological descriptors derived from distance matrices and correlation weights of local graph invariants. Internet Electron. J. Mol. Des. 1, 115–133.

Melagraki, G., Afantitis, A., 2011. Ligand and structure based virtual screening strategies for hit-finding and optimization of Hepatitis C virus (HCV) inhibitors. Curr. Med. Chem. 18, 2612–2619.

Mullen, L.M.A., Duchowicz, P.R., Castro, E.A., 2011. QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents. Chemom. Intell. Lab. 107, 269–275.

Ojha, K., Mitra, I., Das, R.N., Roy, K., 2011. Further exploring r_m^2 metrics for validation of QSPR models. Chemom. Intell. Lab. Syst. 107, 194–205.

Peruzzo, P.J., Marino, D.J.G., Castro, E.A., Toropov, A.A., 2003. QSPR modeling of lipophilicity by means of correlation weights of local graph invariants. Internet Electron. J. Mol. Des. 2, 334–347.

Puzyn, T., Gajewicz, A., Rybacka, A., Haranczyk, M., 2011. Global versus local QSPR models for persistent organic pollutants: balancing between predictivity and economy. Struct. Chem. 22, 873–884.

Roy, K., Saha, A., 2003. QSPR with TAU indices: water solubility of diverse functional acyclic compounds. Internet Electron. J. Mol. Des. 2, 475–491.

Roy, P.P., Leonard, J.T., Roy, K., 2008. Exploring the impact of size of training sets for the development of predictive QSAR models. Chemom. Intell. Lab. 90, 31–42.

Tetko, I.V., Tanchuk, V.Yu., Kasheva, T.N., Villa, A.E.P., 2001. Estimation of aqueous solubility of chemical compounds using E-state indices. J. Chem. Inf. Comput. Sci. 41, 1488–1493.

Toropov, A.A., Toropova, A.P., 2002. QSAR modeling of mutagenicity based on graphs of atomic orbitals. Internet Electron. J. Mol. Des. 1, 108–114.

Toropov, A.A., Toropova, A.P., Benfenati, E., 2010. SMILES-based optimal descriptors: QSAR modelling of carcinogenicity by balance of correlations with ideal slopes. Eur. J. Med. Chem. 45, 3581–3587.

Toropov, A.A., Toropova, A.P., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2011. CORAL: quantitative structure–activity relationship models for estimating toxicity of organic compounds in rats. J. Comput. Chem. 32, 2727–2733.

Toropova, A.P., Toropov, A.A., Diaza, R.G., Benfenati, E., Gini, G., 2011a. Analysis of the co-evolutions of correlations as a tool for QSAR-modeling of carcinogenicity: an unexpected good prediction based on a model that seems untrustworthy. Cent. Eur. J. Chem. 9, 165–174.

Toropova, A.P., Toropov, A.A., Benfenati, E., Gini, G., 2011b. Co-evolutions of correlations for QSAR of toxicity of organometallic and inorganic substances: An unexpected good prediction based on a model that seems untrustworthy. Chemom. Intell. Lab. Sys. 105, 215–219.

Toropova, A.P., Toropov, A.A., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2011c. CORAL: QSPR models for solubility of [C 60] and [C 70] fullerene derivatives. Mol. Divers. 15, 249–256.

Toropova, A.P., Toropov, A.A., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2011d. CORAL: quantitative structure–activity relationship models for estimating toxicity of organic compounds in rats. J. Comput. Chem. 32, 2727–2733.

Weininger, D., 1990. SMILES. 3. Depict. Graphical depiction of chemical structures. J. Chem. Inf. Comput. Sci. 30, 237–243.

Yan, A., Gasteiger, J., 2003. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. J. Chem. Inf. Comput. Sci. 43, 429–434.