Anal Bioanal Chem
DOI 10.1007/s00216-015-9067-6

NOTE

Prediction of retention characteristics of heterocyclic compounds

Karel Nesměrák[1] · Andrey A. Toropov[2] · Alla P. Toropova[2] · Ilkay Yildiz[3] · Ismail Yalcin[3] · Marketa Brozikova[1] · Vera Klimešová[4] · Karel Waisser[4]

Received: 29 July 2015 / Accepted: 18 September 2015

Abstract The CORAL software (http://www.insilico.eu/coral) was used to build up quantitative structure–property relationships (QSPRs) for the retention characteristics of 93 derivatives of three groups of heterocyclic compounds: 2-phenyl-1,3-benzoxazoles, 4-benzylsulfanylpyridines, and benzoxazines. The QSPRs are one-variable models based on the optimal descriptors calculated from the molecular structure represented by simplified molecular input-line entry systems (SMILES). Each symbol (or two undivided symbols) of SMILES is characterized by correlation weight. The optimal descriptor is the sum of the correlation weights. The numerical data on the correlation weights were calculated with the Monte Carlo method by the manner which provides best correlation between endpoint and optimal descriptor for the calibration set. The predictive ability of the model is checked with the validation set (compounds invisible during building up of the model). The approach has been checked with three random splits into the training, calibration, and validation sets: all models have apparent predictive potential. The mechanistic interpretation of the molecular features extracted from SMILES as the promoters of increase or decrease of examined endpoints is suggested.

Keywords QSPR · SMILES · Retention factor · Monte Carlo method · CORAL software

Electronic supplementary material The online version of this article (doi:10.1007/s00216-015-9067-6) contains supplementary material, which is available to authorized users.

✉ Andrey A. Toropov
andrey.toropov@marionegri.it

[1] Faculty of Science, Department of Analytical Chemistry, Charles University in Prague, Hlavova 8, 128 43 Prague 2, Czech Republic

[2] Laboratory of Environmental Chemistry and Toxicology, IRCCS—Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milan, Italy

[3] Faculty of Pharmacy, Department of Pharmaceutical Chemistry, Ankara University, 06100 Ankara, Tandogan, Turkey

[4] Faculty of Pharmacy, Department of Inorganic and Organic Chemistry, Charles University in Prague, Heyrovskeho 1203, 500 05 Hradec Kralove, Czech Republic

Introduction

Reversed-phase high-performance liquid chromatography (RP-HPLC) belongs to the most employed tools of analytical chemistry. Due to a wide range of the combination of stationary and mobile phase defining a given experimental setup, the chromatographic retention prediction methodologies are intensively studied [1]. Although different retention prediction approaches exist, modelling by quantitative structure retention relationships (QSPRs) of solutes is an attractive approach owing to relative simplicity and convenience [2].

The CORAL software [3] has been tested as a tool for development of the predictive models of chromatographic retention characteristics of derivatives of 1-phenylbenzylsulfanyltetrazole. In fact, the CORAL model is a mathematical function of so-called correlation weights of different molecular features extracted from the simplified molecular input-line entry systems (SMILES). In other words, the model is based on descriptors which are the sum of all correlation weights involved in a molecular system. The numerical data on the correlation weights are calculated with the Monte Carlo technique [3]. The SMILES-based optimal descriptors give the possibility to interpret the influence of different molecular features. For instance, a molecular feature that is characterized by positive correlation weight is the promoter of increase of an endpoint, whereas a feature which is characterized by

Published online: 01 October 2015

negative correlation weight is the promoter of decrease of the endpoint.

Often, QSPRs are valid for congeneric series only. Therefore, in the present study, we apply CORAL for varied heterocyclic compounds (Fig. 1). The dataset consists of 21 derivatives of 2-phenyl-1,3-benzoxazoles, 32 derivatives of 4-benzylsulfanylpyridines, and 40 derivatives of benzoxazines. Thus, the aim is estimation of optimal descriptors calculated with the SMILES as a tool to predict retention parameters for the above structurally diverse compounds (n=93).

Materials and methods

Determination of retention data

The studied compounds were synthesized along similar lines as done in previous papers where the synthetic details and analytical data of the studied compounds are described: 2-phenyl-1,3-benzoxazoles [4], 4-(benzylsulfanyl)pyridines [5], and benzoxazines [6]. Acetonitrile Chromasolv (Sigma-Aldrich) with water content below 2×10^{-3} vol% (determined by gas chromatography) was used for the measurements. All other chemicals were of analytical grade.

The measurements were performed using a liquid chromatograph HP 1090 L with a diode array detector (both Hewlett-Packard) working at 235 nm. Reverse-phase RxC-18 ZORBAX 150×4.6 mm column was used. The temperature of the column was held at 25.0 °C. Chromatography was performed with mobile phases containing 80:20, 75:25, 70:30, 65:35, and 60:40 (v/v) acetonitrile to water. The flow rate of the mobile phase was 0.8 mL/min. An aqueous solution of thiourea (c=1 mg/L) was used for the determination of a dead time. The concentration of solutions of the studied compound for HPLC measurements was 1×10^{-4} mol/L in acetonitrile. The Rheodyne loop of 10 μL was used for introduction of the derivative solutions to the HPLC system. All measurements were made at least in triplicate; the average reproducibility of each determination was better than 1.0 % relative.

The retention factor k at a given mobile phase composition was calculated as

$$k = \frac{t_R - t_M}{t_M} \quad (1)$$

where t_R is the retention time of a derivative (s) and t_M is the dead time (s).

The retention characteristics of the linear solvent strength model [7] were calculated using the linear relationship between the logarithm of the retention factor and the volume fraction of the organic modifier in the mobile phase (φ)

$$\log k = \log k_w - S\varphi \quad (2)$$

where k is the solute retention factor at a given φ, $\log k_w$ is $\log k$ extrapolated to a mobile phase composition with 0 % organic modifier (i.e., in pure water), and S is a constant for a given solute in a given chromatographic system (is equal to the slope of the linear regression). The coefficient of determination of linear regressions was for all compounds R^2>0.99 (see Electronic Supplementary Material (ESM) Table S1).

Building up QSPR models

SMILES were used for representation of the molecular structure, and they were generated with ACD/ChemSketch software [8]. The CORAL software [9] was used for the calculations. The SMILES-based optimal descriptors of the correlation weights (DCW), which are calculated with data on the training set, are computed as the following (cf. [10]):

$$DCW(T,N) = CW(HALO) + CW(NOSP) + CW(BOND) + \Sigma CW(S_k) \quad (3)$$

where S_k are one-component SMILES attributes (the component of SMILES represents one symbol, e.g., C, c, N, n, =, F, or two symbols which cannot be separated, e.g., Cl, Br, @@), which are representations of molecular features. CW(S_k) are correlation weights of the SMILES fragments; CW(HALO) are correlation weights of the presence/absence of halogen atoms; CW(NOSP) are correlation weights of the presence/absence of nitrogen, oxygen, sulfur, and phosphorus; and CW(BOND) are correlation weights of the presence/absence of double ("="), triple ("#"), and/or stereo chemical ("@") bonds. Threshold (T) and the number of epochs (N) are

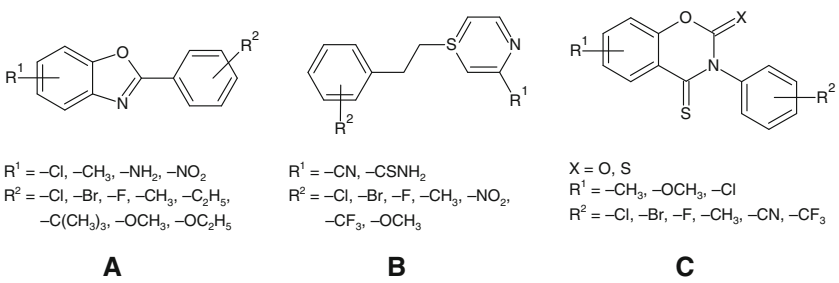

Fig. 1 The structures of studied **A** 2-phenyl-1,3-benzoxazoles, **B** 4-benzylsulfanylpyridines, and **C** benzoxazoles

parameters of the Monte Carlo optimization, used for calculation of the correlation weights. Threshold is criterion for classification of components of the representation of the molecular structure into two classes: rare (noise) and active (not rare). The correlation weight of a rare component is fixed as zero; hence, the rare component is not involved in the building up of the model. N is the number of epochs of the Monte Carlo optimization. The optimal values give the maximum of the correlation coefficient between an endpoint and DCW(T, N) for the training set. The threshold and N were calculated according to the scheme suggested in [11]; the range of the threshold was 1–5; the range of N was 1–100. The preferable T* and N* are values of above-mentioned parameters which give the maximal correlation coefficient for *the calibration set*. The T* and N* are utilized to build up a model.

Having numerical data on the correlation weights, one can calculate DCW(T*, N*) for compounds of training and calibration sets. Using data on the training set, one can calculate by the least squares method model of view

$$\text{endpoint} = C_0 + C_1 \times \text{DCW}(T^*, N^*) \quad (4)$$

where endpoint is log k_w or S. The predictability of the model calculated with Eq. (4) should be checked with the external validation set. It is to be noted that the statistical quality of the model for calibration and validation sets is a mathematical function of the threshold and the number of epochs of the Monte Carlo optimization. Apparently, that statistical quality of the model for external validation set is the most important indicator predictability of an approach. In order to appropriately estimate this approach, we have examined three random splits for the log k_w and S.

The structures, retention characteristics, and SMILES of the examined compounds are represented in the Electronic supplementary material section. Three splits into the training set, calibration set, and validation set of these substances were examined for each endpoint. These splits were prepared according to the following principles: (i) the ranges of the endpoint are comparable for all above-mentioned sets; (ii) these splits are different; and (iii) these splits are random. The histograms for log k_w and S show (see ESM, Table S2) that identical splits for two examined parameters will lead to unbalanced ranges of the log k_w or S for the training, the calibration, and the validation sets. Therefore, utilized splits are balanced for log k_w and S, separately. Moreover, in the case of S, the validation set is the same for three splits, whereas the training and calibration sets are different.

Results and discussion

Table 1 contains one-variable models for log k_w and S together with the statistical characteristics. One can see that different splits have different values of the threshold and the number of epochs of the optimization. The statistical quality of models for log k_w and S considerably varies for different distributions into the visible training and calibration sets and invisible validation set. However, all suggested models can be estimated as quantitative ones.

Table 1 Statistical quality of the QSPR models for half-wave potential of 4-(benzylsulfanyl)pyridines. Here, n, R^2, Q^2, s, and F are the number of substances in set, coefficient of determination, leave-one-out cross-validated coefficient of determination, standard error of estimation, and Fischer F ratio, respectively. The ${}^cR_p{}^2$ is result of Y-scrambling according to [12] model and has predictive potential if ${}^cR_p{}^2$ is greater than 0.5

Split	Set	n	R^2	${}^cR_p{}^2$	Q^2	s	F
log k_w=−3.8635 (±0.0312)+0.1494 (±0.0006)×DCW(2, 39)							
Split 1	Training	39	0.9603	0.9540	0.9568	0.167	894
	Calibration	27	0.9161	0.8969	0.8982	0.167	–
	Validation	27	0.9381	–	0.9296	0.213	–
log k_w=−4.9519 (±0.0620)+0.1654 (±0.0014)×DCW(2, 25)							
Split 2	Training	39	0.9393	0.9145	0.9309	0.227	572
	Calibration	27	0.9281	0.9117	0.9184	0.247	–
	Validation	27	0.9075	–	0.8908	0.263	–
log k_w=−3.1099 (±0.0260)+0.1427 (±0.0006)×DCW(2, 57)							
Split 3	Training	39	0.9714	0.9621	0.9684	0.142	1257
	Calibration	27	0.8799	0.8601	0.8611	0.309	–
	Validation	27	0.9509	–	0.9425	0.186	–
S=−1.1028 (±0.0209)+0.1102 (±0.0005)×DCW(2, 34)							
Split 1	Training	45	0.9705	0.9592	0.9674	0.159	1413
	Calibration	23	0.9127	0.8913	0.8688	0.344	–
	Validation	25	0.9282	–	0.9139	0.256	–
S=−2.1135 (±0.0542)+0.1269 (±0.0013)×DCW(2, 39)							
Split 2	Training	45	0.9308	0.9200	0.9191	0.245	526
	Calibration	23	0.9715	0.9589	0.9654	0.175	–
	Validation	25	0.9094	–	0.9280	0.255	–
S=−2.0726 (±0.0461)+0.1310 (±0.0011)×DCW(2, 37)							
Split 3	Training	45	0.9407	0.9313	0.9318	0.239	682
	Calibration	23	0.9609	0.9372	0.9519	0.251	–
	Validation	23	0.9301	–	0.9172	0.295	–

Having data on a group of runs of the Monte Carlo optimization, one can obtain the following categories of molecular features: the first category: features which have positive correlation weight for all runs; the second category: features which have negative correlation weight for all runs; and the third category: features which have runs where their correlation weight is positive together with runs where their correlation weight is negative. Table 2 contains examples of molecular features of the above three categories for the cases of log k_w and S.

The probabilistic criteria suggested in work [13] for detection of outliers give for log k_w models 5, 3, and 7 outliers for splits 1, 2, and 3, respectively. In the case of the model for S, the number of outliers is 2 for all splits.

Table 2 Correlation weights of SMILES attributes identified as promoters increase or decrease for studied endpoints, and their distributions over training and calibration sets (validation set is invisible during building up of a model)

Split	S_k	CW(S_k) in run 1	CW(S_k) in run 2	CW(S_k) in run 3	Frequency of S_k in the training set	Frequency of S_k in the calibration set
(a) Promoters of increase/decrease for log k_w						
Promoters of increase						
1	1	2.57960	3.29647	3.27353	39	27
2	1	2.67976	3.13591	2.79956	39	27
3	1	2.80598	3.13609	2.49010	39	27
1	2	3.54681	3.30465	3.19329	39	27
2	2	2.87543	3.31334	2.85573	39	27
3	2	3.01584	2.86388	3.18568	39	27
1	3	2.11066	2.32756	2.28935	29	18
2	3	2.78069	2.83531	2.78358	26	21
3	3	2.13136	1.82972	1.73493	27	20
Promoters of decrease						
1	N	−0.31087	−0.34739	−0.20499	36	24
2	N	−2.21736	−2.00452	−2.16118	34	26
3	N	−2.74535	−2.75464	−2.74659	34	26
1	O	−1.23637	−1.22156	−1.21483	27	18
2	O	−1.20212	−1.10913	−1.15371	26	19
3	O	−0.95744	−0.95474	−0.93318	25	20
(b) Promoters of increase/decrease for S						
Promoters of increase						
1	1	2.82676	3.24671	2.84889	45	23
2	1	3.15952	3.07769	2.85971	45	23
3	1	2.92198	3.01125	2.50604	45	23
1	O	1.41491	1.32369	1.40826	29	12
2	O	0.90152	0.83982	0.96456	27	14
3	O	0.86511	0.87234	0.84902	29	12
1	NOSP1110	3.83322	3.63236	3.66282	25	7
2	NOSP1110	3.44023	3.51463	3.29309	19	13
3	NOSP1110	3.22358	3.15348	3.30419	21	11
Promoters of decrease						
1	(	−0.41022	−0.38892	−0.42931	45	23
2	(	−0.02752	−0.04151	−0.04113	45	23
3	(	−0.17956	−0.17837	−0.18890	45	23
1	n	−2.17575	−2.35437	−2.49202	25	18
2	n	−0.57348	−1.11178	−0.79806	31	12
3	n	−0.79155	−0.75441	−0.59832	30	13

1, *2*, and *3* indicators of cycles (rings); *N* indicator for nitrogen; *O* indicator for oxygen; *NOSP1110* indicator of the situation "molecular structure contains nitrogen, oxygen, and sulfur (not phosphorus)"; *bracket (* indicator of branching of molecular skeleton; *n* indicator of nitrogen in aromatic system

Conclusions

The described approach gives quantitative models of retention characteristics for the dataset which involves derivatives of 2-phenyl-1,3-benzoxazoles, derivatives of 4-benzylsulfanylpyridines, and derivatives of benzoxazines. The statistical quality of these models is dependent upon the distribution of the data into the training, calibration, and validation sets. However, for all examined splits into the training, calibration, and validation sets, the models have good predictive potential (Table 1). Thus, the ability of the described approach to be a tool to predict the retention characteristics of non-congeneric series of compounds is demonstrated.

Acknowledgments The authors acknowledge support from the EU project PROSIL funded under the LIFE program (project LIFE12 ENV/IT/000154).

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

1. Put R, Vander Heyden Y (2007) Review on modelling aspects in reversed-phase liquid chromatographic quantitative structure–retention relationships. Anal Chim Acta 602:164–172
2. Kaliszan R (2007) QSPR: quantitative structure-(chromatographic) retention relationships. Chem Rev 107:3212–3246
3. Nesmerak K, Toropov AA, Toropova AP (2014) SMILES-based quantitative structure–retention relationships for RP HPLC of 1-phenyl-5-benzylsulfanyltetrazoles. Struct Chem 25:311–317
4. Pinar A, Yurdakul P, Yildiz I, Temiz-Arpaci O, Acan NL, Aki-Sener E, Yalcin Y (2004) Some fused heterocyclic compounds as eukaryotic topoisomerase II inhibitors. Biochem Biophys Res Commun 317:670–674
5. Klimešová V, Svoboda M, Waisser K, Pour M, Kaustová J (1999) Synthesis and antimicrobial activity of new 4-(benzylsulfanyl)pyridine derivatives. Collect Czech Chem Commun 64:417–434
6. Waisser K, Gregor J, Kubicova L, Klimesova V, Kunes J, Machacek M, Kaustova J (2000) New groups of antimycobacterial agents: 6-chloro-3-phenyl-4-thioxo-2H-1,3-benzoxazine-2(3H)-ones and 6-chloro-3-phenyl-2H-1,3-benzoxazine-2,4(3H)-dithiones. Eur J Med Chem 35:733–741
7. Snyder LR, Dolan JW, Gant JR (1979) Gradient elution in high-performance liquid chromatography: I. Theoretical basis for reversed-phase systems. J Chromatogr A 165:3–30
8. Advanced Chemistry Development, Toronto, Canada, http://www.acdlabs.com/products/draw_nom/draw/chemsketch/. Accessed 1 Feb 2015
9. CORAL, http://www.insilico.eu/CORAL. Accessed 15 Mar 2015
10. Toropov AA, Toropova AP, Benfenati E, Nicolotti O, Carotti A, Nesmerak K, Veselinović AM, Veselinović JB, Duchowicz PR, Bacelo D, Castro EA, Rasulev BF, Leszczynska D, Leszczynski J (2015) QSPR/QSAR analyses by means of the CORAL software: results, challenges, perspectives. In: Roy K (ed) Quantitative structure-activity relationships in drug design, predictive toxicology, and risk assessment. IGI Global, Hershey, pp 560–585
11. Toropova AP, Toropov AA, 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
12. Ojha PK, Roy K (2011) Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection. Chemom Intell Lab 109:146–161
13. Toropov AA, Toropova AP (2015) Quasi-QSAR for mutagenic potential of multi-walled carbon-nanotubes. Chemosphere 124:40–46