Predicting Boiling Points of Aliphatic Alcohols through Multivariate Image Analysis Applied to Quantitative Structure−Property Relationships

Mohammad Goodarzi[†,‡] **and Matheus P. Freitas***[*,§]

Department of Chemistry, Faculty of Sciences, Azad University, Arak, Iran, Young Researchers Club, Azad University, Arak, Iran, Departamento de Química, Universidade Federal de Lavras, CP 3037, 37200-000, Lavras, MG, Brazil

Received: July 4, 2008; Revised Manuscript Received: August 22, 2008

The boiling points of a set of 58 aliphatic alcohols have been modeled through an image-based approach, in which descriptors are pixels (binaries) of 2D chemical structures. While some simple descriptors, such as molecular weight, do not account for some structural influences (e.g., in chain and position isomerism) on the studied property, the MIA-QSPR (multivariate image analysis applied to quantitative structure−property relationship) method, coupled to multilinear partial least-squares regression, correlated the chemical structures with the corresponding boiling points satisfactorily well.

1. Introduction

Various quantitative structure−property relationship (QSPR) models have been proposed for estimating the boiling points of a series of aliphatic alcohols.[1−4] Although predictive, such methods require descriptors which are sometimes difficult to compute. On the other hand, simple parameters that correlate with such a physicochemical property, like molecular weight, are often inefficient to differentiate chain and/or position constitutional isomers. An image-based approach, which is both operationally simple and selective, has been successfully applied to give highly predictive QSPR models.[5−11] Thus, a corresponding QSPR analysis has been performed in this work to predicting boiling points of 58 aliphatic alcohols by using MIA (multivariate image analysis) descriptors and multilinear partial least-squares (N-PLS) regression. Application of this regression method has been found in just a few QSAR studies,[12−17] although the known advantages over bilinear (traditional) PLS. MIA treatment allows the building of 3D arrays, which can be analyzed through multimode methods, such as N-PLS; alternatively, the 3D arrays may be unfolded to X-matrices (two-way arrays) and then carried out using PLS.

2. MIA-QSPR Method

MIA descriptors are binaries obtained from pixels of 2D chemical structures, which must be drawn by using any appropriate drawing program. In this work, the set of 58 aliphatic alcohols is described in Table 1, and the corresponding experimental data (boiling points at 1 atm) were obtained from the literature,[1−4] which are in agreement with a standard source.[18] Since the whole series of alcohols contain similarity centers, for instance the hydroxyl group and C-1, the 2D images (drawn through using the ChemDraw program[19]) were saved as 58 bitmap files and then superimposed by taking a pixel on C-1 at the 340,115 coordinate of a 250 × 400 window size (Figure 1), giving a 58 × 250 × 400 three-way array. Alternatively, a pixel on the hydroxyl group could be taken. Both procedures make the common skeleton of the whole series congruent. Regression of the three-way array with the dependent variables block was carried out using N-PLS. Model validation was achieved through leave-one-out cross-validation (LOO CV) and external validation (for a test set), and the predictive ability was statistically evaluated through the root-mean-square errors of calibration (RMSEC) and validation (RMSECV), as well as by the squared correlation coefficients of the regression line of experimental vs fitted/predicted boiling points.

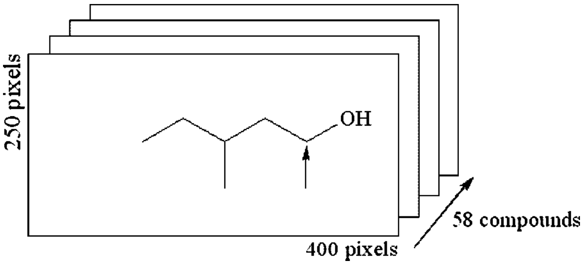

Figure 1. Generic example of how chemical structures for the series of alcohols were drawn to derive the QSPR model. The arrow indicates the pixel fixed at the 340,115 coordinate of a 250 × 400 bitmap workspace.

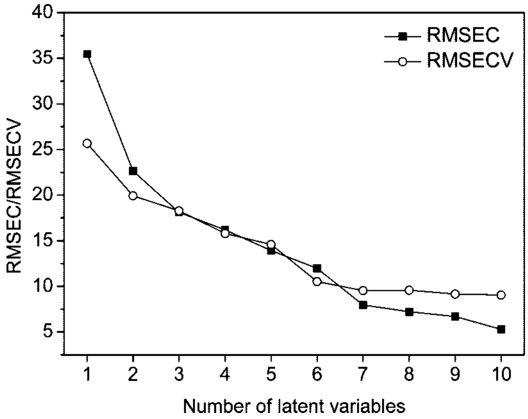

Figure 2. Plot of number of latent variables vs RMSEC/RMSECV.

[*] Corresponding author: matheus@ufla.br; Phone: +55 35 3829-1891; Fax: +55 35 3829-1271

[†] Department of Chemistry, Azad University.

[‡] Young Researchers Club, Azad University.

[§] Universidade Federal de Lavras.

10.1021/jp8059085 CCC: $40.75 © 2008 American Chemical Society
Published on Web 10/09/2008

TABLE 1: Experimental, Fitted (Calibrated), and Predicted Boiling Points (°C) of 58 Aliphatic Alcohols Used in the MIA-QSPR Model[a]

alcohol	exptl	fitted	predicted (LOO CV)	fitted (training set)	predicted (test set)
methanol	64.7	80.6	85.2	74.5	
ethanol	78.3	84.4	78.6	74.4	
1-propanol	97.2	95.9	93.4	104.4	
1-butanol	117.0	118.9	116.7	125.1	
1-pentanol*	137.8	145.4	142.6		139.0
1-hexanol	157.0	169.1	166.5	166.4	
1-heptanol	176.3	185.5	183.2	181.5	
1-octanol	195.2	197.7	200.4	192.9	
1-nonanol*	213.1	203.3	206.4		212.7
1-decanol	230.2	204.8	201.8	226.0	
2-propanol	82.3	81.0	79.8	71.6	
2-butanol	99.6	92.6	93.5	108.5	
2-pentanol*	119.0	116.1	116.4		125.1
2-hexanol	139.9	142.1	142.2	141.7	
2-octanol	179.8	182.2	181.8	185.5	
2-nonanol	198.5	194.2	193.0	189.5	
3-pentanol*	115.3	115.6	115.0		122.4
3-hexanol	135.4	138.1	137.7	130.9	
3-heptanol	156.8	164.6	164.3	156.2	
4-heptanol	155.0	157.7	157.2	166.4	
3-nonanol*	194.7	204.3	202.8		184.0
4-nonanol	193.0	207.1	211.3	191.8	
5-nonanol	195.1	183.5	164.2	185.9	
2-me-1-propanol	107.9	111.7	114.1	113.8	
2-me-2-propanol	82.4	90.7	91.7	65.9	
2-me-1-butanol	128.7	135.2	135.8	137.3	
2-me-2-butanol*	102.0	102.6	102.0		103.5
3-me-1-butanol	131.2	131.1	132.7	131.7	
3-me-2-butanol	111.5	108.6	110.7	111.4	
2-me-1-pentanol	148.0	161.2	163.5	153.7	
3-me-1-pentanol*	152.4	155.6	158.1		154.5
4-me-1-pentanol	151.8	145.3	141.4	137.1	
2-me-2-pentanol	121.4	126.0	125.9	130.2	
3-me-2-pentanol	134.2	131.6	132.5	141.6	
4-me-2-pentanol	131.7	128.4	134.1	137.5	
2-me-3-pentanol	126.6	128.0	128.4	127.2	
3-me-3-pentanol*	122.4	124.8	124.1		113.5
2-me-2-hexanol	142.5	151.8	151.6	142.7	
3-me-3-hexanol	142.4	148.5	150.0	140.1	
7-me-1-octanol	206.0	200.7	198.5	202.7	
2-et-1-butanol	146.5	134.5	133.5	145.9	
3-et-3-pentanol*	142.5	127.4	120.9		110.8
2-et-1-hexanol	184.6	181.6	173.3	180.9	
2,2-dime-1-propanol	113.1	111.6	110.8	110.7	
2,2-dime-1-butanol	136.8	134.8	132.0	133.4	
2,3-dime-1-butanol	149.0	146.3	149.2	147.7	
3,3-dime-1-butanol	143.0	127.7	142.5	136.2	
2,3-dime-2-butanol*	118.6	118.1	123.1		112.2
3,3-dime-2-butanol	120.0	108.3	109.3	115.7	
2,3-dime-2-pentanol	139.7	141.2	144.3	141.3	
3,3-dime-2-pentanol	133.0	131.5	133.9	138.5	
2,2-dime-3-pentanol	136.0	131.2	129.7	131.8	
2,4-dime-3-pentanol	138.8	143.8	147.3	137.6	
2,6-dime-4-heptanol	178.0	167.2	168.3	174.2	
2,3-dime-3-pentanol*	139.0	137.7	136.6		123.2
3,5-dime-4-heptanol	187.0	195.4	193.5	190.9	
2,2,3-trime-3-pentanol*	152.2	140.8	135.7		127.0
3,5,5-trime-1-hexanol	193.0	199.3	187.1	185.2	

[a] Compounds marked with an asterisk pertain to the test set.

3. Results and Discussion

A 58 × 250 × 400 three-way array was built by grouping the 58 2D images (the shapes corresponding to the alcohol structures) in such a way that only variable portions (the chain size and branching) corresponded to the data variance. The three-way array may be unfolded to a 58 × 100000 two-way array, which can by regressed against the Y block (the boiling points column vector) through bilinear (traditional) PLS. However, N-PLS is supposed to be superior to the unfolding PLS due to its simplicity (the number of variables can be effectively reduced) and predictive ability.[20] Thus, this regression method was used in the calibration step directly to the 58 × 250 × 400 three-way array. The optimum number of latent variables (LV) used was seven, since both RMSEC and RMSECV (the root-mean-square errors of calibration and cross-validation) did not decrease significantly after this number (Figure 2). The calibration for the entire data set gave the fitted boiling points of Table 1, correlating with the experimental data by $r^2 = 0.950$ (RMSEC = 7.9), as illustrated in Figure 3. To show that the good correlation was not by chance and to assess the model robustness, the Y block was randomized and the regression step carried out again. The average r^2 obtained after this procedure (10 repetitions) was 0.30 ± 0.08, which is significantly worse than the true calibration, confirming that it was not a fortuitous correlation.

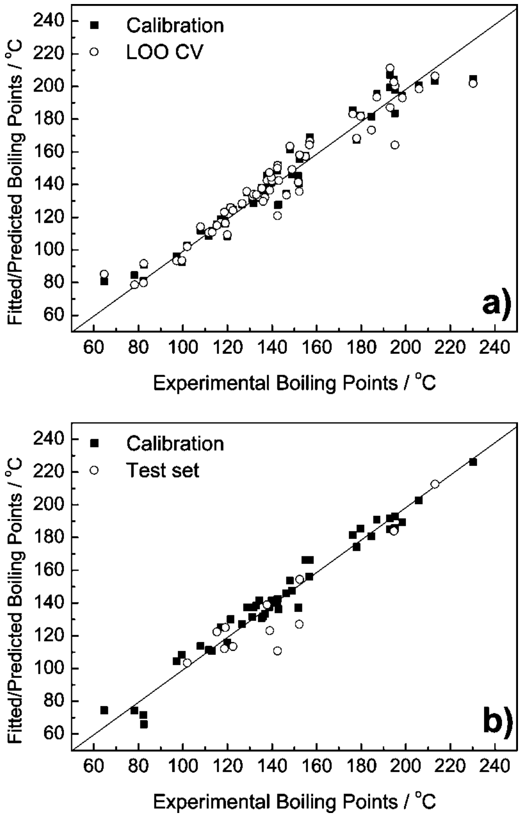

Figure 3. Plots of experimental vs fitted/predicted boiling points for (a) the MIA-QSPR model with the entire data set and (b) the MIA-QSPR model split into training (calibration) and test sets.

The model was validated through LOO CV. The predicted values found when using this procedure are shown in Table 1, and the correlation with experimental data was very good ($q^2 = 0.927$, RMSECV = 9.5), Figure 3. LOO CV has often been considered to be an incomplete validation method; external validation has been strongly recommended instead.[21] Thus, the data set was split into 46 training samples and 12 test samples (Table 1). In this case, the model also demonstrated to be highly predictive, with r^2 of 0.968 and q^2_{test} of 0.864 (Figure 3). These results are comparable to QSPR models previously established, in which r^2 between 0.94 and 1.00 were obtained (Table 2).[1−4]

TABLE 2: Statistics of Theoretical Methods for the Prediction of Alcohol Boiling Points

method	r^2	$q^2_{\text{LOO-CV}}$	q^2_{test}
MIA-QSPR, whole series	0.950	0.927	
MIA-QSPR, split into training and test sets	0.968		0.864
variable connectivity index [1]	0.988−0.994		
CWLIMG [2]	0.990		0.990
CROMRsel [3]	0.943−0.992	0.939−0.990	
RBF/NN [4]	0.996	0.996	0.982

4. Conclusions

Overall, the model built showed to be highly predictive, and the MIA-QSPR method was supposed to be a potential tool for the prediction of other physicochemical parameters, bioactivities, pharmacokinetic data, etc.

Acknowledgment. FAPEMIG (Grant No. CEX 415/06) is gratefully acknowledged for the support of this research, as is CNPq for the fellowship (M.P.F.).

References and Notes

(1) Randic, M.; Pompe, M.; Mills, D.; Basak, S. C. *Molecules* **2004**, *9*, 1177.

(2) Krenkel, G.; Castro, E.; Toropov, A. A. *J. Mol. Struct.* **2001**, *542*, 107.

(3) Janežic, D.; Lucic, B.; Nikolic, S.; Milicevic, A.; Trinajstic, N. *Int. Electron. J. Mol. Des.* **2006**, *5*, 192.

(4) Lia, Q.; Chen, X.; Hu, Z. *Chemom. Intell. Lab. Sys.* **2004**, *72*, 93.

(5) Freitas, M. P.; Brown, S. D.; Martins, J. A. *J. Mol. Struct.* **2005**, *738*, 149.

(6) Freitas, M. P. *Org. Biomol. Chem.* **2006**, *4*, 1154.

(7) Freitas, M. P. *Curr. Comput.-Aid. Drug Des.* **2007**, *3*, 235.

(8) Pinheiro, J. R.; Bitencourt, M.; da Cunha, E. F. F.; Ramalho, T. C.; Freitas, M. P. *Bioorg. Med. Chem.* **2008**, *16*, 1683.

(9) Freitas, M. P. *Chemom. Intell. Lab. Sys.* **2008**, *91*, 173.

(10) Freitas, M. P.; Rittner, R. *QSAR Comb. Sci.* **2008**, *27*, 582.

(11) Freitas, M. P. *Med. Chem. Res.* **2008**, *16*, 461.

(12) Nilsson, J.; de Jong, S.; Smilde, A. K. *J. Chemom.* **1997**, *11*, 511.

(13) Nilsson, J.; Homan, E. J.; Smilde, A. K.; Grol, C. J.; Wikström, H. *J. Comput.-Aid. Mol. Des.* **1998**, *12*, 81.

(14) Hasegawa, K.; Arakawa, M.; Funatso, K. *Chemom. Intell. Lab. Sys.* **2000**, *50*, 253.

(15) Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatso, K. *Comput. Chem.* **2002**, *26*, 583.

(16) Hasegawa, K.; Arakawa, M.; Funatso, K. *Comput. Biol. Chem.* **2003**, *27*, 211.

(17) Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatso, K. *Comput. Biol. Chem.* **2003**, *27*, 381.

(18) Lide, D. R. *CRC Handbook of Chemistry and Physics*; CRC Press: New York, 2007−2008.

(19) *ChemDraw Ultra version 7.0*; CambridgeSoft: Cambridge, MA, 2001.

(20) Ferreira, M. M. C. *J. Braz. Chem. Soc.* **2002**, *13*, 742.

(21) Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269.

JP8059085