International Journal of Pure and Applied Mathematics
Volume 47 No. 1 2008, 23-30

Invited Lecture Delivered at
Forth International Conference of Applied Mathematics
and Computing (Plovdiv, Bulgaria, August 12–18, 2007)

STRUCTURE-PROPERTY BASED MODEL FOR ALKANES BOILING POINTS

Sorana D. Bolboacă[1] §, Lorentz Jäntschi[2]

[1]Medical Informatics and Biostatistics
"Iuliu Hatieganu" University of Medicine and Pharmacy
Cluj-Napoca, 400349, ROMANIA
e-mail: sorana@j.academicdirect.ro

[2]Technical University of Cluj-Napoca
Cluj-Napoca, 400641, ROMANIA
e-mail: lori@academicdirect.org

Abstract: This study discusses the abilities of the family of molecular descriptors on the structure-property relationships ($MDFSPR$) approach in modelling of the alkanes boiling points based on chemical structure information. All alkanes from C3 to C9 were included in the analysis. A $MDFSPR$ model with two descriptors, were constructed. The MDF SPR model was validated, and its correlation coefficient was compared with the best previously reported model. The results of this study revealed that the MDF SPR approach is a useful method to model the boiling points of alkanes, providing valid and stable models.

AMS Subject Classification: 03H05, 62P35, 93E24, 93E35
Key Words: molecular descriptor family on structure-property relationships (MDF SPR), models assessment, boiling point, alkanes

1. Introduction

Quantitative structure-property relationship (QSPR), the method in which the properties of compounds are quantitatively correlated with their structure, has been used since 1868, when Crum-Brown and Fraser studied the physiological

Received: August 17, 2007

§Correspondence author

action of the ammonium salts (see [1]). Twenty-five years latter, Richet studied the relationship between chemical structure and oil-water partition coefficient [2]. Since then, many properties were modelled using quantum chemical descriptors [3], topological indices [4], artificial neural networks [5], and molecular descriptors [6]. The boiling points of alkanes has been previous studied by many researchers. Several models were reported with one, two, three, and four variables, respectively [7]. The equation for the best model (a model with three variables) had the following formula:

$$Bp(^{\circ}C) = 727.26(\pm 20.76)\cdot 3D0_{\chi} - 19.46(\pm 0.9)\cdot 3DSRW2 + 7.99(\pm 0.39)\cdot M2 - 779.42(\pm 20.08),\ n = 73; r = 0.9986; s = 2.17; F = 8340\,, \quad (1)$$

where $3D0_{\chi}$ and $3DSRW2$ are MIS (Method of Ideal Symmetry) indices, and $M2$ is a $3D$ modification of the Zagreb index; Bp(°C) is the boiling point. A new method called molecular descriptor family on the structure-property relationships (MDF SPR) has been introduced [8] and its prediction and estimation abilities has been proved (see [9]). The current study attempted to find the relationships between the structure of alkanes with three, four, five, six, seven, eight and nine carbons and boiling points, identifying the best MDF SPR model and to analyzing their estimation and prediction abilities.

2. Material and Method

2.1. Alkanes Set and Boiling Points

A sample of seventy-three compounds was studied, containing all alkanes isomers with three to nine carbons (one C3 alkane, two C4 compounds, three C5 compounds, five C6 compounds, nine C7 compounds, eighteen C8 compounds and thirty-five C9 compounds, respectively). The names of the compounds included in the study are presented in Table 1. The experimental boiling points data were taken from a previously reported paper [10].

2.2. Mathematical Model

The methodology of MDF SPR has been developed and is described in details in [8]. The molecules were drawn and optimized by using the HyperChem software. All seventy-three compounds were used in the construction and generation of the molecular descriptors family, resulting thus a large class

of molecular descriptors. A seven-letter name was assigned to each descriptor to identify its modality of construction [8]. The best performing multivariate model was identified based on the value of the correlation coefficient. To demonstrate the absence of chance correlation in the obtained models, two validation analyses were conducted. In the first method, one compound was randomly extracted from the sample; the MDF SPR model was rebuilt and based on this new model the boiling point of the excluded compound was estimated. In the second method, the whole set of 73 compounds was randomly divided into two groups: training and test sets. The MDF SPR model were rebuilt in training set and the boiling points of the compounds from test sets were predicted using the model obtained by corresponding training set. This action has been done twenty-four times, for sample sizes in training sets that vary from 40 to 63 and corresponding sample sizes in test sets from 33 to 10. In both validation procedures, the same molecular descriptors were used in the MDF SPR model and just the coefficients were allowed to vary. The correlation coefficients obtained in training and test sets were compared by using the Fishers Z-test [11] at a level of significance of 5% (see also [12, 13]). The Steigers Z test was applied to test the significance of the difference between the correlation coefficient of the best performing MDF SPR model and that of the previously reported model [11].

3. Results and Discussion

The MDF SPR model with the ability to estimate boiling points for studied alkanes was identified, having two descriptors (see equation (2)), where lGDrtGt, and IbDrfHt are molecular descriptors (MDF members).

$$\hat{Y} = -129.20 - 67.45 \cdot lGDrtGt + 4.89 \cdot IbDrfHt\,. \tag{2}$$

According to equation (2), the boiling points of alkanes are strongly dependent on the topology of compounds (lGDrtG***t***, IbDrfH***t***, ***t*** from topology), being related with the group electronegativity (lGDrt***G***t) and with the number of directly bonded hydrogens (IbDrf***H***t), respectively. Thus, the boiling point is directly related to the descriptor called IbDrfHt and inverse related to the second descriptor. Analyzing the absolute value of residuals obtained from the model equation (2) it can be concluded that in fifty out of seventy-three cases, the model described by equation (2) resulted in better values. The values of molecular descriptors used by the MDF SPR models, the estimated boiling points obtained the model (equation (2)), and the residuals are presented in

Short name	Bp (°C)	lGDrtGt	IbDrfHt	$\hat{Y}$ (°C)	Bp-$\hat{Y}$(°C)
C3	-42.50	0.703	27.34	-42.97	0.47
2M-C3	-0.50	0.053	27.36	1.0131	-1.51
n-C4	-11.73	0.318	28.35	-12.05	0.32
2,2-MMC3	36.07	-0.510	27.39	39.125	-3.06
2M-C4	27.85	-0.094	30.68	27.171	0.68
n-C5	9.50	0.114	30.06	10.069	-0.57
2,2-MMC4	68.74	-0.954	27.81	71.123	-2.38
2,3-MMC4	60.27	-0.612	30.49	61.144	-0.87
2M-C5	63.28	-0.462	32.76	62.149	1.13
3M-C5	49.74	-0.162	34.20	48.942	0.80
n-C6	57.99	-0.193	34.99	54.919	3.07
2,2,3-MMMC4	98.43	-1.350	28.14	99.399	-0.97
2,2-MMC5	90.05	-1.030	30.81	91.012	-0.96
3,3-MMC5	91.85	-0.881	32.96	91.412	0.44
2,3-MMC5	93.48	-0.719	35.29	91.814	1.67
2,4-MMC5	79.20	-0.622	34.02	79.127	0.07
2M-C6	89.78	-0.549	36.67	87.183	2.60
3M-C6	80.50	-0.738	33.18	82.785	-2.29
3E-C5	86.06	-0.426	37.83	84.528	1.53
n-C7	80.88	-0.230	39.53	79.581	1.30
2,2,3,3-MMMMC4	125.66	-1.690	28.54	124.29	1.37
2,2,3-MMMC5	117.65	-1.430	30.80	117.73	-0.08
2,3,3-MMMC5	118.93	-1.270	33.04	117.8	1.13
2,2,4-MMMC5	117.71	-1.240	33.35	117.41	0.30
2,2-MMC6	118.53	-1.050	36.08	117.77	0.76
3,3-MMC6	106.84	-1.020	34.28	107.07	-0.23
3,3-MEC5	115.61	-0.935	36.92	114.39	1.22
2,3,4-MMMC5	109.43	-0.992	35.60	111.75	-2.32
2,3-MMC6	109.10	-1.140	33.53	111.5	-2.40
2,3-MEC5	111.97	-0.810	38.00	111.23	0.74
2,4-MMC6	117.73	-0.829	38.70	115.93	1.80
2,5-MMC6	115.65	-0.780	39.14	114.78	0.87
2-MC7	118.26	-0.633	41.51	116.51	1.75
3-MC7	109.84	-0.564	41.01	109.33	0.51
4-MC7	99.24	-0.759	36.51	100.54	-1.30
3-EC6	114.76	-0.497	42.67	112.94	1.82
n-C8	113.47	-0.655	40.21	111.57	1.90
2,2,3,3-MMMMC5	106.47	-0.249	45.08	108.01	-1.54
2,2,3,4-MMMMC5	150.80	-2.000	28.87	146.78	4.02
2,2,3-MMMC6	143.26	-1.760	31.01	141.44	1.82
2,2,3-MMEC5	144.18	-1.630	32.94	141.48	2.70
2,3,3,4-MMMMC5	142.48	-1.560	33.67	140.94	1.54
2,3,3-MMMC6	143.00	-1.390	36.10	141.21	1.79
2,3,3-MMEC5	142.10	-1.320	37.14	141.38	0.72
2,2,4,4-MMMMC5	132.69	-1.410	33.99	132.19	0.50
2,2,4-MMMC6	140.50	-1.310	36.67	138.75	1.75
2,4,4-MMMC6	133.50	-1.340	35.85	136.32	-2.82
2,2,5-MMMC6	136.00	-1.380	35.38	136.87	-0.87
4,4-MMC7	135.21	-1.540	33.04	136.32	-1.11
3,3-EEC5	137.30	-1.180	37.93	135.68	1.62
2,3,4-MEMC5	140.10	-1.170	38.92	140.21	-0.11

Table 1: Optimized parameters values and its statistics

Short name	Bp (°C)	lGDrtGt	IbDrfHt	$\hat{Y}$ (°C)	Bp-$\hat{Y}$(°C)
2,3,5-MMMC6	136.00	-1.230	37.65	137.72	-1.72
2,3-MEC6	135.20	-1.140	38.48	135.56	-0.36
3,4-EMC6	138.00	-1.070	40.19	139.59	-1.59
2,4-MEC6	133.80	-1.140	38.84	137.57	-3.77
3,4-MMC6	140.60	-0.944	42.22	140.87	-0.27
n-C9	140.40	-1.000	41.57	141.51	-1.11
2-MC8	133.60	-0.924	41.27	134.92	-1.32
3-MC8	126.54	-1.010	38.87	128.68	-2.14
4-MC8	124.08	-1.150	36.81	128.13	-4.05
3-EC7	137.68	-0.849	42.93	137.96	-0.28
4-EC7	139.00	-0.910	42.30	139.05	-0.05
2,2-MMC7	131.34	-1.060	39.42	134.94	-3.60
2,3-MMC7	130.65	-0.934	40.47	131.64	-0.99
2,4-MMC7	140.46	-0.769	44.46	140.07	0.39
2,5-MMC7	146.17	-0.796	45.27	145.80	0.37
2,6-MMC7	133.83	-0.778	43.42	135.58	-1.75
3,3-MMC7	142.00	-0.690	46.19	143.17	-1.17
3,4-MMC7	136.72	-0.846	42.84	137.36	-0.64
3,5-MMC7	140.27	-0.509	47.90	139.33	0.94
3,3-MEC6	133.01	-0.679	44.32	133.31	-0.30
3,3,4-MMMC6	122.28	-0.794	39.74	118.70	3.58
2,3,4-MMMC6	141.55	-0.585	47.10	140.54	1.01

$M = methyl$; $E = ethyl$; $C = carbon$

Cx - a linear (normal) alkane with x carbon atoms

Table 1: Continuation

Table 1. Statistical characteristics in terms of squared correlation coefficients, Fisher parameters and associated significance, standard error of the MDF SPR models are:

$$^{95\%}CI_{intercept}^{[-132.23,-126.16]}{}_{,lGDrtGt}^{[-68.3,-66.6]}{}_{,IbDrfHt}^{[4.81,4.97]};\ r^2 = 0.9982,\ F = 19361,\ s = 1.74$$

$$p_F < 1\%\%;\ leave - one - out:\ r^2 = 0.998,\ F = 17837,\ s = 1.82,\ p_F < 1\%\%$$

The correlation coefficients of the descriptors from equation (2) shown that the descriptors did not correlate one to each other, the squared correlation coefficient being equal with 0.0024. The MDF SPR model with two descriptors (equation (2)) was validated in training vs. test analysis. The results of this analysis are presented in Table 2. In seventy-five percent of cases, no significant differences were identified between the correlation coefficients obtained in training and test sets. More, with a single exception, the values of the correlation coefficients obtained in training sets were included in the 95% confidence intervals of the MDF SPR model from equation (2), observation which is valid for seventy percent of the cases in test sets. The hypothesis that there are not significant differences between the model described by equation (2) and previously reported model (1) was tested using the Steigers Z test. The results showed that the statistical significance of the correlation coefficient obtained

N_{tr}	r	$^{95\%}CI_r$	N_{ts}	r	$^{95\%}CI_r$
40	0.9992	0.9987-0.9994	33	0.9990	0.9984-0.9993
41	0.9993	0.9988-0.9995	32	0.9986	0.9977-0.9991
42	0.9989	0.9982-0.9993	31	0.9993	0.9988-0.9995
43	0.9988	0.9980-0.9992	30	0.9994	0.9990-0.9996
44	0.9987	0.9979-0.9991	29	0.9994	0.9990-0.9996
45	0.9993	0.9988-0.9995	28	0.9986	0.9977-0.9991
46	0.9991	0.9985-0.9994	27	0.9993	0.9988-0.9995
47	0.9990	0.9984-0.9993	26	0.9992	0.9987-0.9994
48	0.9993	0.9988-0.9995	25	0.9979	0.9966-0.9986
49	0.9994	0.9990-0.9996	24	0.9985	0.9976-0.9990
50	0.9984	0.9974-0.9989	23	0.9995	0.9992-0.9996
51	0.9992	0.9987-0.9994	22	0.9987	0.9979-0.9991
52	0.9991	0.9985-0.9994	21	0.9992	0.9987-0.9994
53	0.9991	0.9985-0.9994	20	0.9992	0.9987-0.9994
54	0.9991	0.9985-0.9994	19	0.9993	0.9988-0.9995
55	0.9990	0.9984-0.9993	18	0.9994	0.9990-0.9996
56	0.9992	0.9987-0.9994	17	0.9985	0.9976-0.9990
57	0.9991	0.9985-0.9994	16	0.9992	0.9987-0.9994
58	0.9991	0.9985-0.9994	15	0.9992	0.9987-0.9994
59	0.9993	0.9988-0.9995	14	0.9965	0.9944-0.9978
60	0.9990	0.9984-0.9993	13	0.9995	0.9992-0.9996
61	0.9992	0.9987-0.9994	12	0.9962	0.9939-0.9976
62	0.9992	0.9987-0.9994	11	0.9920	0.9872-0.9949
63	0.9992	0.9987-0.9994	10	0.9971	0.9953-0.9981

Table 2: Training (tr) versus Test (ts) experiment results

from equation (2) is greater than the one obtained from equation (1) (Steigers Z parameter = 2.8, p = $2.6 \cdot 10^{-3}$). Thus, the MDF SPR model with two descriptors resulted in better data than the previously reported model. Based on these findings, we propose that the MDF SPR[1] model with two descriptors can be used to predict the boiling point of other alkanes.

4. Conclusions

A $MDFSPR$ model with good statistical parameters proved to be able to estimate and predict the boiling points of the alkanes with variable number of atoms (from 3 to 9). The descriptors involved in the MDF SPR model were calculated solely from the chemical structure and showed that the boiling points of the studied alkanes depend on the topology of the compounds and correlate with the group electronegativity and with the number of directly bonded hydrogens. The internal validation of the MDF SPR model with two descriptors demonstrates the stability and reliability of the model.

[1] http://vl.academicdirect.org/molecular_topology/mdf_findings/sar

Acknowledgments

This research was partly funded by UEFISCSU Romania through projects ET36/2005&108/2006.

References

[1] A. Crum-Brown, T.R. Fraser, On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the salts of the ammonium bases, derived from Strychnia, Brucia, Thebia, Codeia, Morphia, and Nicotia, *Trans. R. Soc. Edinbours*, **25** (1868), 151-203.

[2] M.C. Richet, *Compt. Rend. Soc. Biol.*, **45** (1893), 775-776.

[3] W. Liu, P. Yi, Z. Tang, QSPR models for various properties of polymethacrylates based on quantum chemical descriptors, *QSAR Comb. Sci.*, **25** (2006), 936-943.

[4] A.T. Balaban, Can topological indices transmit information on properties but not on structures? *J. Comput. Aided. Mol. Des.*, **19** (2006), 651-660.

[5] N.J. English, D.G. Carroll, Prediction of Henry's law constants by a quantitative structure property relationship and neural networks, *J. Chem. Inf. Comp. Sci.*, **41** (2001), 1150-1161.

[6] J. Xu, B. Guo, B. Chen, Q. Zhang, A QSPR treatment for the thermal stabilities of second-order NLO chromophore molecules, *J. Mol. Modeling.*, **12** (2005), 65-75.

[7] A. Toropov, A. Toropova, T. Ismailov, D. Bonchev, 3D weighting of molecular descriptors for QSPR/QSAR by the method of ideal symmetry (MIS). 1. Application to boiling points of alkanes, *J. Mol. Struct. Theochem*, **424** (1998), 237-247

[8] L. Jäntschi, Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology, *LEJPT*, **4**, No. 6 (2005), 76-98.

[9] L. Jäntschi, S. Bolboaca, Results from the Use of Molecular Descriptors Family on Structure Property/Activity Relationships, *International Journal of Molecular Sciences*, **8**, No. 3 (2007), 189-203.

[10] S.C. Basak, G.J. Niemi, G.D. Veith, Predicting properties of molecules using graph invariants, *J. Math. Chem.*, **7** (1991), 243-272.

[11] J.H. Steiger, Tests for comparing elements of a correlation matrix, *Psychol. Bull.*, **87** (1980), 245-251.

[12] A.R. Katritzky, A. Lomaka, R. Petrukhin, R. Jain, M. Karelson, A.E. Visser, R.D. Rogers, QSPR Correlation of the Melting Point for Pyridinium Bromides, *Potential Ionic Liquids. J. Chem. Inf. Comput. Sci.*, **42** (2002), 71-74.

[13] A.R. Katritzky, U. Maran, M. Karelson, V.S. Lobanov, Prediction of Melting Points for the Substituted Benzenes: A QSPR Approach, *J. Chem. Inf. Comput. Sci.*, **37** (1997), 913-919.