J Mol Model (2001) 7:65–69
DOI 10.1007/s008940100012

ORIGINAL PAPER

Mohamed Nohair · Driss Zakarya · Hamid Nyassi

Structure–boiling point relationships of alkanes using the multifunctional autocorrelation method

Received: 28 September 2000 / Accepted: 1 February 2001 / Published online: 10 April 2001

Abstract The concept of the multifunctional autocorrelation method (MAM), governing the global description of molecules, has been changed to take into account the structural environment of each atom in all fragments of the molecule. New topological parameters are generated as possible descriptors in QSAR (quantitative structure–activity relationships) and can be useful for database characterization and encoding a variety of physicochemical properties. The well-known component P_k was modified as shown in the following formula: $P_k=\Sigma[f(i)(\Sigma(f(kij))f(j)]^{1/n}$ (n is the number of atoms in the fragment considered). The efficiency of the approach is demonstrated through a case study dealing with the design of a model structure–property relationship for boiling points of alkanes. The data set was analysed by multiple linear regression.

Keywords Boiling points · Molecular structure–property · Multifunctional autocorrelation method · Alkanes · Linear modeling method

Introduction

Recently, the establishment of structure–activity relationships has become a very interesting research field. Elaboration of such relationships requires a set of molecules and their activities or properties. It is currently believed that the activity or the property of the molecule generally depends on its chemical structure [1, 2]. Thus the reliability of structure–activity relationships is highly dependent on the parameters chosen to account for the chemical structure.

M. Nohair (✉) · D. Zakarya
Faculty of Sciences and Technology, Department of Chemistry,
BP 146, 20650 Mohammedia, Morocco
e-mail: nohairm@hotmail.com
e-mail: nohair@mail.uh2m.ac.ma
Tel.: +212-3-314705/314708, Fax: +212-3-315353

H. Nyassi
Faculty of Sciences, El Jadida, Morocco

Numerous methods have been introduced to describe the chemical structure of a given set of molecules; several are based on the molecular graph in conjunction with structure and properties of molecules. Among the large variety of descriptions, many are based on topological indices (TI) [3, 4, 5, 6, 7] contained in molecular graphs, usually hydrogen-depleted graphs. The topology of a chemical structure can be coded by the adjoining matrix A=(aij), where aij is the weight of the edge (i, j), aij=0 if the vertices i and j are not connected by an edge. The weight of aij is chosen to take into account the differences between the types of atoms and bonds. Another matrix Dij, called the distance matrix, will be defined; the entries of this matrix, dij, are equal to the number of edges connecting vertices i and j on the shortest path between them.

Different numbers characterizing the chemical structure of the molecule are calculated from its graph. Such numbers are called topological indices (TI). Topological indices have found a wide variety of applications in structural chemistry. In particular, they can be used to code chemical information, in the design of a chemical experiment, in the theory of the atomic structure and reactivity of molecules, and for quantitative description of chemical structures in the analysis of the relationships between the structure of a molecule and its properties.

Wiener [8] first proposed an index based on path distances between carbon atoms. This index was defined as the sum of the chemical bonds existing between all pairs of carbon atoms in the molecule under consideration. The Wiener index has been used to model a wide range of the physicochemical properties including a variety of thermodynamic properties. Although (TI) have been used successfully in many studies, several cannot describe chemical structure with high chromatism.

In this paper we give the general procedure for the use of the autocorrelation method in structure–property relationships. The instructive example was directed to design of the structure–property relationship for the boiling points of alkanes. Properties such as boiling point, heat of vaporization, and critical temperature are determined by

intermolecular forces. For neutral molecules these forces consist of the polar van der Waals' forces of attraction. Molecules considered in this study (alkanes) are non-polar and do not contain functional groups, so a number of complexities that arise with more polar compounds are avoided. By examining the normal alkanes, it is apparent that the boiling points are related to molecular size. However, consideration of the isomers demonstrates that boiling points are also related to the molecular shape.

Methods

The autocorrelation method was introduced into the field of structure–activity relationships by Moreau and Broto [9]. The general relationship used to calculate the autocorrelation component is defined in Eq. (1):

$$P_k=\sum[f(i)f(j)]^x \quad d=k \tag{1}$$

- Pk is the autocorrelation component corresponding to the topological distance k (the smallest number of bonds).
- x is equal to 1 for the classical method and 0.5 for the modified autocorrelation method [10]. This modification was needed to enable satisfactory physical interpretation of the component. As an example, for a roughly additive molecular property, P_0 can be considered as a first approximation and the other component, Pi as measures of interactions between atoms. This corresponds to the modified autocorrelation method ($^{0.5}$MAM).
- The atomic contribution f(i) depends on the property under study. These properties can be based on the van der Waals' volume (V) and surface (S) to account for the size and the shape of the molecule, connectivity (number of non-hydrogen neighbors or vertex degree of atom i), or electronegativity and charge [11], to take into account electronic aspects.

We compute P_k by fixing the considered distance d (d=k), P_k is defined as the sum of (f(i)×f(j)) for all the chemical bonds existing between all pairs of carbon atoms i and j separated by a topological distance equal to d.

Example of computation of P_k components

Taking 2,3,4-trimethylhexane, shown in Fig. 1, for example, carbon atoms in the molecule can be classified as four basic types – types 1, 2, 3, and 4 for primary (CH_3–), secondary (CH_2<), ternary (–CH<), and quaternary (>C<) carbons. Their properties are given in Table 1.

The procedure for computing the components of autocorrelation vectors, using the MAM method, is illustrated as follows:

f(i)=van der Waals' volume (V)

There are nine carbon atoms in the molecule considered – five primary, one secondary, and three tertiary carbon atoms. For k=0 we obtain:

V_0 (k=0)=5×(13.67)+1×(10.23)+3×(6.78)=92.92

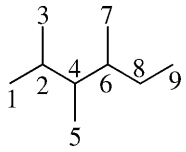

Fig. 1 A hydrogen-depleted molecular graph corresponding to the skeleton of 2,3,4-trimethylhexane

For a topological distance equal to 1 (k=1), there are eight pairs (i, j) of atoms ((1, 2); (2, 3); (2, 4); (4, 5); (4, 6); (6, 7); (6, 8); (8, 9)) and we compute V_1 as follows:

$$V_1(k=1)=[4(13.67\times6.78)^{1/2}+2(6.78\times6.78)^{1/2}+(6.78\times10.23)^{1/2}+(10.23\times13.78)^{1/2}]=72.22$$

etc.

Materials

In this study we considered 300 substituted alkanes. The boiling points are taken from the CRC Handbook of Chemistry and Physics [12]. The range of experimental normal boiling points (the number of carbons per alkanes spanned from six through eleven carbons) was from 68.74 °C for hexane up to 195.89 °C for undecane. To simplify the computation of the components, molecules were coded by means of the Smiles system [13] and stored as input files. The computer program used to compute the components of the autocorrelation method represents an algorithm for the construction of both the connectivity and the distance matrix of any molecule from its Smiles code.

Results

A linear modeling method was employed to investigate the behavior of each component. Generally, models obtained by using descriptors based on connectivity and surface are good enough for practical purposes. To derive more significant models to estimate boiling points, however, it seemed logical to consider components of an autocorrelation vector based on van der Waals' volumes. It was necessary to consider several components of a volume van der Waals' vector to obtain better regressions. Monoparametric correlation with boiling points (BP) for 300 alkanes with n=6–11 carbons leads to Eq. (2):

$$\begin{aligned}BP(°C) = {} & (-148.81\pm6.46)+(5.21\pm0.34)\times V_0 \\ & -(1.60\pm0.22)\times V_1-(0.94\pm0.09)\times V_2 \\ & -(0.11\pm0.02)\times V_3-(0.09\pm0.01)\times V_4 \\ & -(0.12\pm0.01)\times V_5\end{aligned} \tag{2}$$

(n=300, r=0.99, s=3.1 °C)

Table 1 Contributions of atoms to some molecular properties

Property	Number of atoms in molecule (Fig. 1)								
	1	2	3	4	5	6	7	8	9
Connectivity	1	3	1	3	1	3	1	2	1
Van der Waals' volume, V ($cm^3 \cdot mol^{-1}$)	13.67	6.78	13.67	6.78	13.67	6.78	13.67	10.23	13.67

Table 2 Intercorrelation matrix of the components (V_0–V_5) for 300 alkanes

	V_0	V_1	V_2	V_3	V_4	V_5
V_0	1.00					
V_1	0.76	1.00				
V_2	0.69	0.11	1.00			
V_3	0.59	0.22	0.43	1.00		
V_4	0.68	0.42	0.50	0.58	1.00	
V_5	0.64	0.68	0.35	0.07	0.19	1.00

Despite the general difficulty in obtaining a satisfactory interpretation of the autocorrelation components, it has been shown recently [14, 15] that the components with a small index are essentially correlated with the size of the molecule whereas those with a higher index account for the molecular branching (shape). According to Gore's method [16], the contributions of components were 79%, 15%, 4%, and 1% for V_0, V_1, V_2, and V_3, respectively. The component V_0, which is correlated with the size of the molecule, seems sufficient to explain the

Table 3 Values of components V_i for the isomers of heptane

Compound	V_0	V_1	V_2	V_3	V_4	V_5	BP (°C)
Heptane	78.49	64.57	54.34	44.11	33.88	23.65	98.43
2-Methylhexane	78.48	59.87	67.7	43.8	33.28	27.34	90.05
3-Methylhexane	78.48	60.16	63.66	57.18	37.32	13.67	91.85
3-Ethylpentane	78.48	60.46	59.57	70.95	41.01	0	93.47
2,2-Dimethylpentane	78.47	48.13	94.15	42.22	41.01	0	79.2
2,3-Dimethylpentane	78.47	55.82	72.33	74.29	27.34	0	89.78
2,4-Dimethylpentane	78.47	55.17	81.42	38.51	54.68	0	80.5
3,3-Dimethylpentane	78.47	48.82	84.7	78.33	13.67	0	86.06
2,2,3-Trimethylbutane	78.46	44.25	97.06	82.02	0	0	80.88

Table 4 Observed boiling points of 74 alkanes and values predicted by use of Eq. (3) (data set used by Seybold)

Observation	Name	Observed values	Predicted values	Difference	Observation	Name	Observed values	Predicted values	Difference
1	2	–88.63	–87.39	–1.24	38	234MMM5	113.46	112.66	0.81
2	3	–42.07	–43.76	1.69	39	2233MMMM4	106.47	105.98	0.49
3	4	–0.50	–2.21	1.71	40	9	150.79	151.90	–1.10
4	2M3	–11.73	–11.81	0.08	41	2M8	143.26	143.17	0.09
5	5	36.07	34.07	2.01	42	3M8	144.18	144.28	–0.10
6	2M4	27.85	28.50	–0.65	43	4M8	142.48	142.35	0.13
7	22MM3	9.50	10.67	–1.17	44	3E7	143.00	142.59	0.41
8	6	68.7	66.14	2.60	45	4E7	141.20	140.33	0.87
9	2M5	60.27	60.41	–0.14	46	22MM7	132.69	133.36	–0.67
10	3M5	63.28	63.93	–0.64	47	23MM7	140.50	139.09	1.41
11	22MM4	49.74	51.28	–1.53	48	24MM7	133.50	135.20	–1.70
12	23MM4	57.98	58.55	–0.56	49	25MM7	136.00	136.21	–0.21
13	7	98.42	95.79	2.63	50	26MM7	135.21	135.80	–0.59
14	2M6	90.05	88.96	1.09	51	33MM7	137.30	137.76	–0.46
15	3M6	91.85	91.91	–0.06	52	34MM7	140.60	140.28	0.32
16	3E5	93.47	94.51	–1.03	53	35MM7	136.00	137.44	–1.44
17	22MM5	79.19	80.54	–1.34	54	44MM7	135.20	137.16	–1.96
18	23MM5	89.78	89.97	–0.19	55	23ME6	138.00	137.25	0.75
19	24MM5	80.50	83.14	–2.64	56	24ME6	133.80	134.67	–0.87
20	33MM5	86.06	87.67	–1.60	57	33ME6	140.60	142.04	–1.44
21	223MMM4	80.88	81.83	–0.95	58	34ME6	140.40	140.56	–0.16
22	8	125.66	123.76	1.91	59	223MMM6	133.60	133.40	0.20
23	2M7	117.64	116.62	1.02	60	224MMM6	126.54	127.15	–0.61
24	3M7	118.92	118.20	0.73	61	225MMM6	124.08	124.90	–0.81
25	4M7	117.70	117.63	0.07	62	233MMM6	137.68	137.51	0.17
26	3E6	118.53	118.55	–0.01	63	234MMM6	139.00	137.30	1.70
27	22MM6	106.84	106.88	–0.04	64	235MMM6	131.34	131.30	0.04
28	23MM6	115.60	114.65	0.95	65	244MMM6	130.64	130.95	–0.30
29	24MM6	109.42	111.33	–1.90	66	334MMM6	140.46	140.39	0.07
30	25MM6	109.10	108.83	0.27	67	33EE5	146.16	147.72	–1.56
31	33MM6	111.96	113.41	–1.44	68	223MME5	133.83	134.88	–1.05
32	34MM6	117.72	117.70	0.03	69	233MME5	142.00	143.35	–1.35
33	23ME5	115.65	116.51	–0.86	70	234MEM5	136.73	135.17	1.56
34	33ME5	118.25	119.81	–1.55	71	2233(M)5	140.27	136.73	3.54
35	223MMM5	109.84	110.78	–0.94	72	2234(M)5	133.01	131.39	1.62
36	224MMM5	99.23	101.00	–1.77	73	2244(M)5	122.28	117.38	4.91
37	233MMM5	114.70	114.70	0.06	74	2334(M)5	141.55	138.85	2.70

physical meaning of the boiling points of alkanes. This corresponds to the well known high correlation between the molecular size and the boiling points.

Table 2 presents the intercorrelation matrix of components (V_0–V_5) for the 300 alkanes. The high intercorrelation coefficient between V_0 and V_1 indicates that these components express approximately the same type of structural information. If we consider that V_1 represents molecular size, its negative contribution is the opposite of this hypothesis. The component V_1 is defined as the sum of f(i)×f(j) for all chemical bonds existing between all pairs of adjoining carbon atoms in the molecule under consideration. Generally, V_1 increases with the number of atoms. When isomers were considered separately (Table 3), the component V_1 varies irregularly with the branching of the molecule. The component V_1 is considered as taking into account both the molecular size and the shape. V_2 and V_3 contribute poorly and negatively, suggesting that the branching makes a minor contribution to the determination of boiling points.

Although components were used successfully as predictor variables in many studies, they cannot describe the chemical structure with high chromatism because only atomic properties f(i) and f(j) are considered in each interatomic distance d(i,j). It will be interesting to see how much is gained by the introduction of new components based on the concept of chromatism developed in the DARC system [17] by including the property of the intermediate atom for topological distance greater than two, as shown by the modified Eq. (3):

$$Pk=\sum[f(i)(\sum (f(k_{ij}))f(j)]^{1/n} \quad (3)$$

where k_{ij} represents the intermediate carbon atom between atoms i and j and n the number of atoms in the selected fragment.

The model was improved (r=0.99, s=2.8 °C) when we consider as structural descriptor a vector of components computed by Eq. (4). A plot of predicted boiling points against the experimental boiling points is given in Fig. 2. The new predicted model was:

$$\begin{aligned} BP(°C) = {} & (-178.45 \pm 5.45) + (3.20 \pm 0.07) \times V_0 \\ & -(1.16 \pm 0.07) \times V_1 - (1.16 \pm 0.07) \times V_2 \\ & -(0.46 \pm 0.04) \times V_3 - (0.17 \pm 0.02) \times V_4 \\ & -(0.39 \pm 0.02) \times V_5 \end{aligned} \quad (4)$$

(n=300, r=0.99, s=2.8 °C)

The boiling points of alkanes have been predicted many times [18, 19, 20, 21], but the set used for these models is usually limited to alkanes with up to nine carbon atoms. The most accurate QSPR models based on TI for predicting boiling points (all alkanes with up to nine carbon atoms have been considered, but not methane) is taken by Seybold [22]. Their model was built for a set containing 74 alkanes (n=74, r=0.99, s=1.86 °C). Considering the same set, the fit of the modeling using the MAM method with components of Eq. (3) gives a better prediction (n=74, r=0.99, s=1.47 °C) (Table 4).

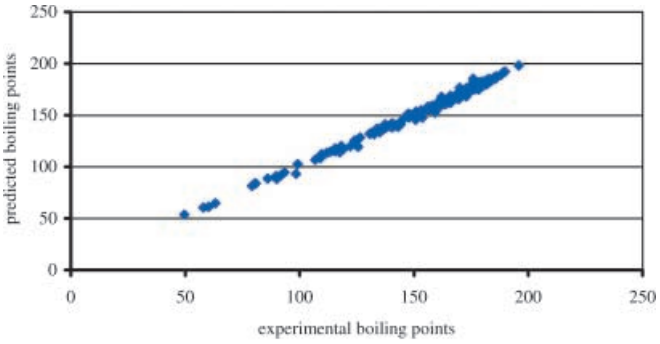

Fig. 2 Plot of predicted against experimental boiling points for the set of 300 alkanes

Table 5 The observed boiling points for the test set (28 alkanes) and the values predicted from the linear model (LRA)

Molecule	Observed	Calculated (LRA)	Deviation
3-Methylpentane	63.28	65.04	1.76
3-Ethylpentane	93.47	94.90	1.43
2,5-Dimethylhexane	109.10	108.42	–0.68
2,3,3-Trimethylpentane	114.76	116.19	1.43
2,6-Dimethylheptane	135.21	133.22	–1.99
2,2,3-Trimethylhexane	133.60	134.52	0.92
2,4-Dimethyl-3-ethylpentane	136.69	138.02	1.33
5-Methylnonane	165.10	163.02	–2.08
3,4-Dimethyloctane	163.40	162.13	–1.27
2-Methyl-5-ethylheptane	159.70	157.55	–2.15
2,2,5-Trimethylheptane	150.80	150.67	–0.13
3,3,4-Trimethylheptane	161.90	162.14	0.24
2,3-Dimethyl-3-ethylhexane	163.70	164.24	0.54
2,2,4,4-Tetramethylhexane	153.80	147.37	–6.43
2,2,4-Tetramethyl-3-ethylpentane	155.30	155.40	0.10
5-Ethylnonane	183.00	181.93	–1.07
3,3-Dimethylnonane	182.00	184.11	2.11
4-Isopropyloctane	178.00	176.12	–1.88
4-Methyl-3-ethyloctane	183.00	182.23	–0.77
2,3,4-Trimethyloctane	180.00	180.16	0.16
2,5,6-Trimethyloctane	178.00	177.94	–0.06
4,4,5-Trimethyloctane	180.00	181.39	1.39
2,4 Dimethyl-4-ethylheptane	180.00	178.94	–1.06
2,2,4,4-Tetramethylheptane	185.00	186.34	1.34
2,3,4,4-Tetramethylheptane	175.00	175.78	0.78
3,3,4,5-Tetramethylheptane	181.00	182.65	1.65
3-Methyl-3,4-diethylhexane	187.00	187.66	0.66
2,4-Dimethyl-3,3-diethylpenthane	189.00	189.66	0.66

To validate the stability of the model seen previously, a new model, obtained by using 272 compounds in the training set from the initial set (300 alkanes), was developed to predict the boiling points for the 28 compounds in the external prediction set. The standard error was 2.8 °C for the training set and 1.8 °C for the test set. Predicted boiling points of the test set are listed in Table 5. Four compounds show a deviation larger than 2 °C and one compound a high deviation (>6 °C).

This study demonstrates that the modified autocorrelation method and multiple linear regression enable successful estimation of boiling points with a small number of topological parameters (six components of autocorrela-

tion method). The efficiency of the topological descriptors generated from the multifunctional autocorrelation method is not well demonstrated here in comparison with other models. However, our aim is only to illustrate how we have adapted it to generate new components for a correct description of the molecular structure in alkanes. The success of this method depends on the correct description brought by each component when it is modified by taking into account all carbon atoms existing in a given fragment with a topological distance greater than two.

Conclusion

In this study, limited to the boiling point of a class of non-polar compounds (alkanes), a single major factor was dominant. This factor was interpreted as being associated with polar van der Waals forces of attraction between molecules. So the model obtained is a function of the structural environment of this pattern and the boiling points. The correlation coefficient is high ($r>0.99$) and the standard deviation error is very low when component V_i accounts for properties of all atoms included in fragments with a topological distance greater than two.

References

1. Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334–343.
2. Seybold, P. G.; May; M.; Bagal, U. A. *J. Chem. Educ.* **1987**, *64*, 575–581.
3. Rouvary, D. H. *J. Comput. Chem.* **1987**, *8 (4)*, 470–480.
4. Labanowski, J. K.; Motoc, I.; Dammkoehler. R. A. *Computers Chem.* **1991**, *15 (1)*, 47–53.
5. Radic. M, *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
6. Kier; L. B.; Hall, L. H. *Molecular-Connectivity in Structure–Activity Analysis. Wiley*, New York, 1986.
7. Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*. Cordon and Breach Science Publishers. 1999.
8. Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
9. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359–360.
10. Chastrette, M.; Tiyal, F.; Peyraud, J. F. *C.R. Acad. Sci. Paris, Ser II.* **1990**, *310*, 514–515.
11. Pauling, L. *The Nature of the Chemical Bond*; 3rd edn; Cornell University Press: Ithaca, NY, 1960.
12. *Handbook of Chemistry and Physics*; Weast, R. C., Shelbey, S. M., Eds.; 70th edn; Chemical Rubber Company: Cleveland, 1990.
13. Weininger, D. *J. Chem. Inf. Comput. Sci*. **1988**, *28*, 31–36.
14. Zakarya, D. *New J. Chem.* **1992**, *16*, 1039–1042.
15. Zakarya, D.; Tiyal, F.; Chastrette, M. *J. Phys. Org. Chem*. **1993**, *6*, 574–582.
16. Gore, W. L. *Statistical Methods for Chemical Experimentation*; Interscience: New York, 1952.
17. Dubois, J. E.; Mercier, C.; Panaye, A. *Acta. Pharm. Jugosl*. **1986**, *36*, 135–169.
18. Liu, S.; Cao. C.; Li, Z. *J. Chem. Inf. Sci.* **1998**, *38*, 387–394.
19. Estrada, E.; Rodriguez., L. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1037–1041.
20. Ren, B. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 139–143.
21. Cao, C.; Liu, S.; Li, Z. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1105–1111.
22. Needham, E. D.; Wei, I-Chien; Seybold, P. G. *J. Am. Chem. Soc.* **1988**, *110*, 4186–4194.