TOPS-MODE Versus DRAGON Descriptors in QSAR. 2. Chromatographic Properties.

Maykel Pérez González [1, 2, *], Aliuska Morales Helguera [3, 2].

[1] Unit of Services, Experimental Sugar Cane Station "Villa Clara-Cienfuegos", Ranchuelo, Cuba.

[2] Department of Drug Design, Chemical Bioactives Center, Central University of Las Villas, Santa Clara, Villa Clara. Cuba.

[3] Department of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara, Villa Clara, Cuba.

Abstract

The TOPological Sub-Structural MOlecular DEsign (TOPS-MODE) approach has been applied to the study of the retention times in liquid chromatographic of various polyhalogenated biphenyls. A model able to describe more than 96% of the variance in the experimental retention times of 56 organic compounds was developed with the use of the mentioned approach. In contrast, four different approaches, including the use of Constitutional, Charges indices, Randic Molecular profiles and RDF descriptors was able to explain more than 91% of the variance in the mentioned property with the same number of descriptors. In addition, the TOPS MODE model was compared with the model obtained with Hasan and Jurs where the superior of the first was clearly demonstrated.

Keywords: **Molecular descriptors, Retention times, QSPR, Dragon, TOPS-MODE.**

***Corresponding Author:** mpgonzalez76@yahoo.es
Tel: 53-42-281473; Fax: 53-42-281130

1 INTRODUCTION

Halogenated biphenyls are a group of chemical compounds formed by substituting hydrogen in biphenyl with one of the halogens. In theory, there are 209 possible congeners for each series of halogenated biphenyl. However, the compounds, which have received the most attention from environmental chemist and toxicologist, are the chlorinated biphenyls (PCBs). Their chemical inertness and other desirable physical properties have made the PCBs a versatile chemical product with many industrial applications [1]. Nevertheless due to indiscriminate use and improper disposal, PCBs have become one of the most widely spread environmental contaminants. Because of their persistence, they are still considered a major threat to the environment, despite the fact that their production has long been banned.

On the other hand, the other halogenated biphenyls such as polybrominated biphenyls (PBBs) are not as persistent as PCBs. Thus, their toxicology and environmental impact have not been as extensively studied as with PCBs.

One of the most commonly used methods of analysis for polyhalogenated biphenyls in environmental samples is Liquid Chromatography. This method is still considered especially for sample cleanup or fractionation prior to GC analysis. Erickson [2] has given a discussion on analytical method for analisys of PBCs.

On the other hand, the prediction of physicochemical and biological properties of organic molecules is one of the main objectives of the methods based on quantitative structure-property relationships (QSPRs) [1]. Among the most important parameters that have been extensively studied by using these approaches are the chromatographic ones [2].

In the context of *in silico* methods for modeling physicochemical and biological properties of chemicals, the topological sub-structural molecular design (TOPS-MODE) approach has been introduced. The TOPS-MODE has been applied to the description of physicochemical and biological properties of organic compounds [3-8].

The successful application of this theoretical approach to the modeling of Chromatography properties [5] has inspired us to perform a study in order to test and/or validate TOPS MODE applicability.

We will show here how TOPS MODE is able to produce good QSAR models that permit easy structural interpretation and their comparison with other methodologies.

2 MATERIALS AND METHODS

2.1 The Tops-Mode Approach

TOPS-MODE is based on the computation of the spectral moments of the bond matrix, the mathematical basis of which has been described previously [9–12]. The TOPS-MODE approach has been recently reviewed in the literature [13], and both the methodology and its software implementation have been described [14].

According to the authors, the application of the TOPS-MODE approach to the study of quantitative structure – property relationships (QSPR) can be summarized in the following steps:

1. To draw the hydrogen-depleted molecular graphs for each molecule of the data set,
2. To use appropriate bond weights in order to differentiate the molecular bonds, e.g., bond distance, bond dipoles, bond polarizabilities, etc.,
3. To compute the spectral moments of the bond matrix with the appropriate weights for each molecule in the data set, generating a table in which rows correspond to the compounds and columns correspond to the spectral moments of the bond matrix. Spectral moments are defined as the trace of the different powers of the bond matrix [15],
4. To find QSPR by using a suitable linear or non-linear multivariate statistical technique, such as multi-linear regression analysis (MRA), etc. to obtain an equation of the form:

   $$P = a_0\mu_0 + a_1\mu_1 + a_2\mu_2 + a_3\mu_3 \ldots\ldots\ldots\ldots a_k\mu_k + b \qquad \text{(Eq. 1)}$$

   where P is the property measurement, μ_k is the *k*th spectral moment, and a_k's are the coefficients obtained by the MRA,

5. To test the predictive capability of the QSPR model by using cross-validation techniques.

2.2 Data Sets and Computational Strategies.

A data set of 56 polyhalogenated biphenyls including 13 fluorinated, 22 chlorinated, 18 brominated, and three iodinated biphenyls for which the Chromatographic retention data were reported in the literature was selected [16]. The parameter studied is the retention time in Liquid Chromatography (RT) where retention for all 56 compounds in the data set were obtained on an ODS (C_{18}) column using 100% methanol as the mobile phase. The names of the compounds, as well as the calculated and experimental values of RT are shown in Table 1.

Table 1 comes about here

TOPS-MODE [14] and DRAGON [17] computer softwares were employed to calculate the molecular descriptors. In the case of TOPS-MODE software, the standard distance, the Van Der Waals radii and hidrophobicity were used to weigh the bond adjacency matrix. The selection of only these three types of descriptors from the whole pool of ten types included in TOPS-MODE methodology was carried out for the sake of simplicity and on the belief that steric and polarity; parameters influence the retention times of compounds. The total number of descriptors used for obtaining this model was 48 spectral moments. On the other hand, we carry out geometry optimization calculations for each compound used in this study using the quantum chemical semiempirical method AM1 [18] included in MOPAC 6.0 [19]. Four other models were developed using the computer software Dragon [17], and calculating the

Constitutional, Charges Indices, Randic Molecular Profiles and RDF descriptors [20]. The statistical processing to obtain the QSAR models was carried out by using the forward stepwise regression methods.

The statistical significance of the models was determined by examining the regression coefficient, the standard deviation, the number of variables, the cross validation leave-one-out statistics and the proportion between the cases and variables in the equation.

3 RESULTS AND DISCUSSION

3.1 Quantitative Structure Property Relations

The best QSPR model obtained with the TOPS-MODE descriptors is given below together with the statistical parameters of the regression.

$$RT = -78.96 + 1158.00 \cdot \mu_2^{H} - 0.01 \cdot \mu_{11}^{H} + 3.13 \cdot 10^{-4} \cdot \mu_{11}^{VDW} + 3.27 \cdot 10^{-7} \cdot \mu_{15}^{VDW} - 1.36 \cdot 10^{-4} \cdot \mu_{11}^{SD} \qquad \textbf{(Eq. 2)}$$

N = 56 S = 56.513 $R^2 = 0.965$ F = 272.09 p = 0.000 $q^2 = 0.955$ $S_{cv} = 59.35$

where N is the number of compounds included in the model, R^2 is the correlation coefficient, S the standard deviation of the regression, F the Fisher ratio, q^2 the correlation coefficient of the cross – validation, p is the significance of the variables in the model and S_{cv} is the standard deviation of the cross – validation.

The variables included in the model are designated as follows: the sub-index represents the order of the spectral moment and the super-index the type of bond weight used, i.e., SD for Standard distance, VDW for Van Der Waals radii and H for hidrophobicity. From the statistical point of view, this model is a robust one as can be seen from the statistical parameters of the cross-validation.

As we previously mentioned, one of the objectives of the current work was to compare the reliability of the TOPS-MODE approach to describe the property under study as compared to other different descriptors and methods. Consequently, four other models were developed using the same data set and the same number of variables that was included in the TOPS-MODE QSPR model. The results obtained with the use of Constitutional, Charges Indices, Randič Molecular Profiles, RDF descriptors are given in Table 2.

Table 2 comes about here.

3.2 Comparison with other Approaches

As can be seen there are remarkable differences concerning the explanation of the experimental variance given by these models compared to the TOPS-MODE one. While the TOPS-MODE QSPR model explains more than 96% of Retention times the rest of the models are unable to explain beyond 91% of such variance.

The TOPS-MODE model is superior to the other four models not only in the statistical parameters of the regression but also, and more importantly, in its stability upon inclusion/exclusion of compounds as measured by the correlation coefficient and standard deviation of the cross-validation. Because of the structural variability of the compounds in the data set these statistics of the leave-one-out cross validation might be considered as a good measurement of the predictive ability of the models. As can be seen in Table 2, the value of the determination coefficient of leave-one-out cross-validation for the model obtained with the spectral moments ($q^2 = 0.955$) was the highest of all.

On the other hand, we established a comparison with the work of Hasan and Jurs [16] in order to demonstrate the potentiality of the TOPS MODE approach in this topic.

In the interesting paper, Hasan and Jurs used a combination of five molecular descriptors for modeling this chromatographic property but some of these descriptors are based on experimental properties. Therefore, prediction of the retention times of the three iodinated biphenyl derivates present in the model is impossible because parameters for the calculation of partial charges for iodine were not available, problem that not present the TOPS-MODE approach. In addition, the variable used (fraction of positively charge, fraction of negatively charge and relative positive charge) in the model should present a high intercorelation due to that this properties encode the same information in the molecule and the authors not shown a absolute presumption of orthogonalization of this according to the Randic methods for example. Therefore, this bring serious problem when the model is interpreted because the sign of the coefficient in general change and even some variable should be not significant in the new model [21].

Finally, the authors obtained an excellent $R^2 = 0.968$ vs 0.965 obtained by our model. This correlation coefficient not presents great differences. Nevertheless, we used 56 compounds, when we eliminated three compounds of our model (51, 52, 53) the $R^2 = 0.981$ whereas overcome at the model reported by Hasan and Jurs.

This compounds present a big symmetric. This interesting behavior was reported by Hasan and Jurs but in the compounds 34 and 53. This anomaly is probably caused by interactions at the molecule in a specific site with the methanol solvent.

Therefore, we in this work have demonstrated the potential of the TOPS MODE in the modeling of physico - chemical properties in special the Chromatographic properties.

4 CONCLUDING REMARKS

We have shown that the TOPS-MODE approach is able to describe the retention times of different compounds in liquid chromatographic. In fact, we have developed a model for predicting the retention time coefficient of a data set of 56 polyhalogenated biphenyls, which is both statistically and chemically sound. This model explains more than 96 % of the variance in the experimental retention times coefficients and shows good predictive ability in cross-validation. These features are significantly better than those obtained for other four different methodologies used to predict this property. Therefore, the spectral moments show a better performance than other kind of descriptors, which suggests that they can be used in new QSPR applications.

Finally, the present results were compared to others obtained in previous works and evidence was obtained on the similarity of the properties that explain the phenomenon.

Acknowledgments

Maykel Pérez thanks the collaboration of the owner of the software Modeslab 1.0 for the donation of this valuable tool and his valuable contributions to the completion of this work.

5 REFERENCES

1. O. Hutzinger, S. Safe, V. Zitko. The Chemistry of PCBs, CRC Press: Boca Raton, Fl, 1980.
2. M. D. Erickson. Analytical Chemistry of PCBs, Butterworth: Stoneham, MA, 1986.
3. E. Estrada. Modeling the diamagnetic susceptibilities of organic compounds by a substructural graph theoretical approach. *J. Chem. Soc. Faraday Trans*., **1998.** 94, 1407-1411.

4. E. Estrada. Generalized Spectral Moments of the interacted line graph sequence. A novel approach to QSPR studies. *J. Chem. Inf. Comput. Sci*., **1998**. 39, 90-95.
5. E. Estrada. and Y. Gutiérrez. Modeling chromatographic parameters by a novel graph theoretical sub-structural approach. *J. Chromatogr. A*, **1999,** 858, 187-199.
6. E. Estrada, Y. Gutiérrez and H. González. Modeling diamagnetic and magnetooptic properties of organic compounds with the TOSS-MODE approach. *J. Chem. Inf. Comput. Sci*., **2000**, 40, 1386-1399.
7. M. P. Gonzalez and A. M. Helguera. TOPS-MODE versus DRAGON descriptors to predict permeability coefficients through Low-density polyethylene. *J. Comp. Aided Mol. Design.* (*in press*)
8. M. P. Gonzalez, H. Gonzalez, R. Molina, M. A. Cabrera and Ramos R. TOPS-MODE Based QSARs Derived from Heterogeneous Series of Compounds. Applications to the Design of New Herbicides. *J. Chem. Inf. Comput. Sci*. **2003**, 43, 1192-1199.
9. E. Estrada. Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. *J. Chem. Inf. Comput. Sci*., **1996**. 36, 844-849.
10. E. Estrada. Spectral moments of the edge adjacency matrix in molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. *J. Chem. Inf. Comput. Sci*., **1997**. 37, 320-328.
11. E. Estrada. Spectral moments of the edge adjacency matrix in molecular graphs. 3. Molecules containing cycles. *J. Chem. Inf. Comput. Sci*., **1998**. 38, 23-27.
12. E. Estrada. Modeling the diamagnetic susceptibilities of organic compounds by a substructural graph theoretical approach. *J. Chem. Soc. Faraday Trans*., **1998.** 94, 1407-1411.
13. Y. Gutierrez and E. Estrada. TOSS-MODE (1997) (Topological Sub-Structural Molecular Design) for Windows Version 4.0, Universidad de Santiago de Compostela, Spain.
14. Y. Gutierrez and E. Estrada. Modes Lab ®, **2002**, version 1.0 b.
15. E. Estrada, Edge Adjacency relationship and a novel topological index related to molecular volume, *Inf. Comput. Sci*., **1995**, 35, 31-33.

16. M. N. Hasan and P.C. Jurs. Prediction of gas and liquid Chromatographic retention indices of polyhalogenated biphenyls. *Anal. Chem.* **1990,** 62, 2318-2323.
17. R. Todeschini, V. Consonni and M. Pavan. 2002 Dragon. Software version 2.1.
18. J.S. Michael, E. Dewar, G. Zoebisch, F. Eamonn, and J.P. Stewart. *J. Am. Chem. Soc.*, **1985,** 107, 3902.
19. J. J. P. Stewar. **1990** MOPAC manual, 6th ed., p 189, Frank J. Seiler Research Laboratory, U. S. Air Force academy, Colorado Springs, CO.
20. R. Todeschini and V. Consonni. Handbook of molecular descriptors. Wiley-VCH, Weinheim, Germany, **2000**.
21. E. Estrada, I. Perdomo-Lopez and J.J. Torres-Labandeira. Combination of 2D-, 3D-Connectivity and Quantum chemical descriptors in QSPR. Complexation of α- and β-cyclodextin with benzene derivates. *J. Chem Inf. Comput. Sci.* **2001**, 41, 1561-1568.

Table 1. Names of compounds; observed, predicted, and residual values of retention times in liquid chromatographic for the 56-polyhalogenated biphenyls used to derive the QSPR [2].

Number	**Compounds**	**Observed**	**Predicted**	**Deleted Residuals**
1	2-fluorobiphenyl	367.000	351.368	16.933
2	3-fluorobiphenyl	382.000	362.384	20.942
3	4-fluorobiphenyl	370.000	322.916	51.162
4	2,3-difluorobiphenyl	355.000	321.403	35.494
5	2,4-difluorobiphenyl	304.000	287.604	17.482
6	2,5-difluorobiphenyl	295.000	300.234	-5.542
7	2,6-difluorobiphenyl	223.000	279.812	-62.275
8	3,4-difluorobiphenyl	347.000	329.024	20.158
9	3,5-difluorobiphenyl	382.000	342.036	44.475
10	4,4’-difluorobiphenyl	227.000	210.541	23.113
11	2,2’,4-trifluorobiphenyl	228.000	205.077	26.476
12	2,3,5,6-tetrafluorobiphenyl	212.000	302.903	-102.591
13	decafluorobiphenyl	278.000	263.946	45.671
14	2-chlorobiphenyl	442.000	435.706	6.849
15	3-chlorobiphenyl	578.000	562.573	16.592
16	4-chlorobiphenyl	521.000	531.620	-11.247
17	2,3-dichlorobiphenyl	540.000	561.229	-22.444

18	2,4-dichlorobiphenyl	605.000	578.861	26.803
19	2,5-dichlorobiphenyl	559.000	589.550	-31.376
20	2,6-dichlorobiphenyl	415.000	450.893	-39.198
21	3,4-dichlorobiphenyl	706.000	690.780	16.183
22	4,4’-dichlorobiphenyl	583.000	627.949	-47.302
23	2,3,4-trichlorobiphenyl	732.000	709.760	23.585
24	2,4,5-trichlorobiphenyl	817.000	761.610	57.092
25	2,4,6-trichlorobiphenyl	700.000	641.570	60.244
26	3,4,5-trichlorobiphenyl	891.000	874.440	18.419
27	2,3,4,5-tetrachlorobiphenyl	978.000	913.903	70.875
28	2,3,4,6-tetrachlorobiphenyl	832.000	793.271	41.043
29	2,3,5,6-tetrachlorobiphenyl	794.000	794.085	-0.090
30	2,2’,4,4’-tetrachlorobiphenyl	682.000	724.701	-46.499
31	3,3’,4,4’-tetrachlorobiphenyl	871.000	946.263	-88.650
32	2,3,3’,4,5-pentachlorobiphenyl	982.000	1041.261	-65.095
33	2,2’,4,4’,5,5’-hexachlorobiphenyl	977.000	1090.368	-132.664
34	2,2’,4,4’,6,6’-hexachlorobiphenyl	916.000	856.249	74.259
35	decachlorobiphenyl	1520.000	1499.669	41.430
36	2-bromobiphenyl	480.000	478.813	1.277
37	3-bromobiphenyl	633.000	657.506	-26.646
38	4-bromobiphenyl	606.000	629.571	-24.955
39	2,3-dibromobiphenyl	624.000	643.754	-20.831
40	2,4-dibromobiphenyl	728.000	715.529	13.043
41	2,5-dibromobiphenyl	669.000	729.175	-63.797
42	2,6-dibromobiphenyl	494.000	534.375	-46.171
43	3,4-dibromobiphenyl	761.000	827.257	-68.984
44	3,5-dibromobiphenyl	882.000	926.129	-52.640
45	4,4’-dibromobiphenyl	703.000	823.851	-131.548
46	2,4,5-tribromobiphenyl	923.000	938.579	-16.438
47	2,4,6-tribromobiphenyl	832.000	814.409	19.985
48	3,4,5-tribromobiphenyl	1000.000	1047.164	-50.538
49	2,3,4,5-tetrabromobiphenyl	1050.000	1073.716	-26.137
50	2,3,4,6-tetrabromobiphenyl	939.000	953.058	-16.217
51	3,3’,4,4’-tetrabromobiphenyl	996.000	879.713	132.736
52	3,3’,5,5’-tetrabromobiphenyl	1030.000	879.713	171.545
53	3,3’,4,4’,5,5’-hexabromobiphenyl	1370.000	1219.216	187.133
54	2-iodobiphenyl	483.000	432.448	61.730
55	3-iodobiphenyl	663.000	691.102	-30.036
56	4-iodobiphenyl	640.000	666.362	-28.019

Table 2. Statistical parameters of the lineal regressions models obtained for the ten kinds of descriptors.

Descriptors	**Variables**[a]	**S**	**R^2**	**F**	**q^2**
Spectral moments	μ_2H, μ_{11}H, μ_{11}VDW, μ_{15}VDW μ_{11}D	56.51	0.965	272.09	0.955
Constitutional	Sp, Me, Mp, Ms, nI	106.50	0.861	69.432	
Charges Indices	GGI5, GGI8, JGI1, JGI7, JGI8	221.63	0.455	8.343	
Randic Molecular Profiles	SP03, DP02, SP07, DP01, SP08	131.65	0.789	45.295	
RDF	RDF095m, RDF040v, RDF040p, RDF085p, RDF100p,	88.417	0.913	105.24	

[a] The definition of the terms appears largely explained in reference 20.

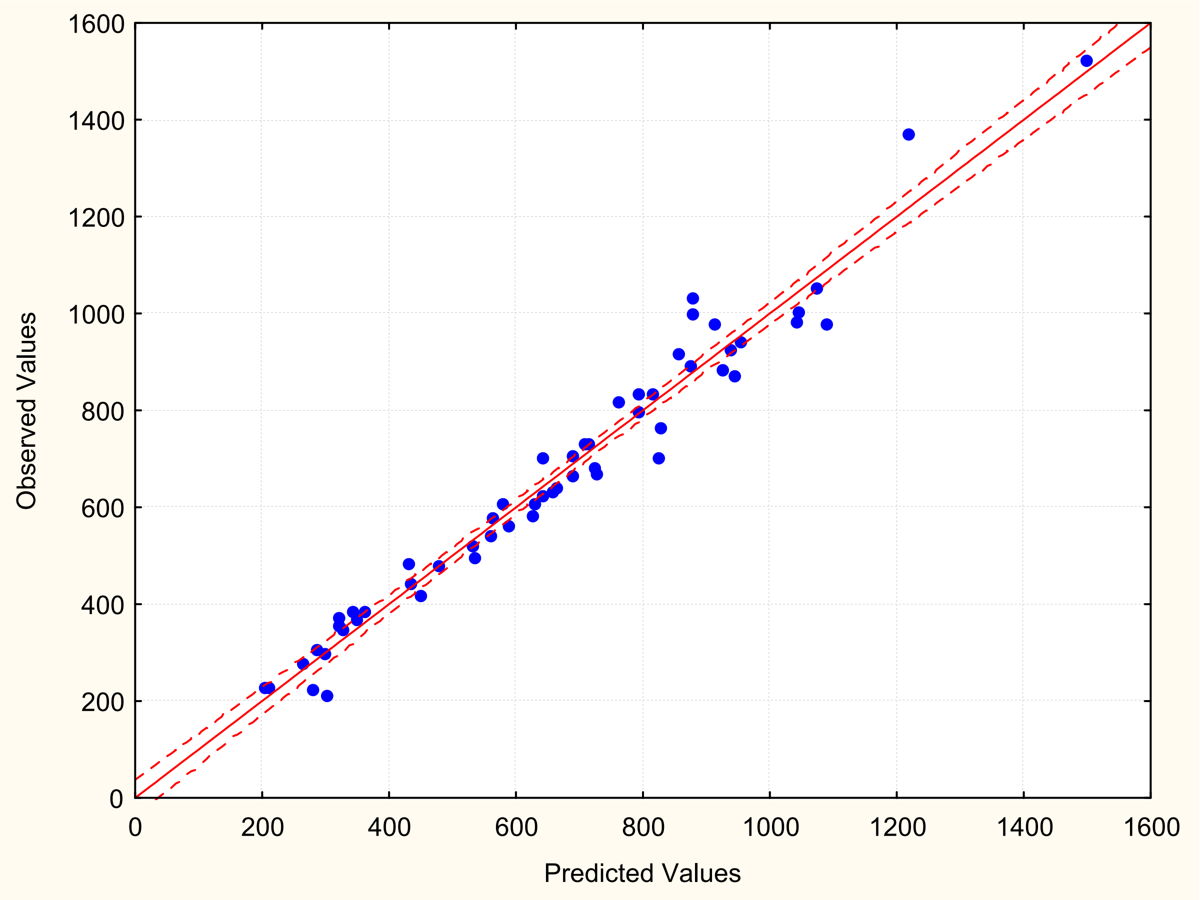

Figure 1. The linear relation between observed and predicted retention times for the compounds of the training set.