Study on QSPR Method for Theoretical Calculation of Boiling Point of Some Organic Compounds

Kamal Raj Sapkota

Department of Chemistry, Prithvi Narayan Campus, Tribhuvan University, Nepal

Correspondence: kamalrajsapkota@yahoo.com

Abstract

Quantitative structure-property relationship (QSPR) models based on molecular descriptors derived from molecular structures have been developed for the prediction of boiling point using a set of 25 organic compounds. The molecular descriptors used to represent molecular structure include topological indices and constitutional descriptors. Forward stepwise regression was used to construct the QSPR models. Multiple linear regressions is utilized to construct the linear prediction model. The prediction result agrees well with the experimental value of these properties.

Keywords: *QSPR, Boiling point, Organic compounds.*

Introduction

The study of the quantitative relationship between property/activity and molecular structure (QSPR/QSAR) is an important research area in computational chemistry and has been widely used in the predication of physicochemical properties and biological activities of organic compound (Katritzky et.al; 2000 and Katritzky et.al 2001). This kind of study develops a method for the predication of the property under investigation of new compounds that have not been synthesized. It can also identify and describe important structure features of molecules that are relevant to variations in molecular properties, thus gain some insight into structural factors affecting molecular properties. To develop a QSPR model, the following step are usually involved i.e. data collection, molecular geometry optimization, molecular descriptors generation, descriptors selection, model development and finally model performance evaluation. One of the important problems in QSPR is the description of molecular structure using molecular descriptors which can include structural information as much as possible. At present, there exist a great number of molecular descriptors that encode constitutional, topological, geometry and electronic features of organic compounds (Karelson, 2000; Devillers & Balaban, 1999). Among various structure descriptors, those derived from molecular structure alone have a particular advantage of the possibility to calculate them based only in molecular structural feature and to be applicable to different families of compounds (Etrada and Uriarte; 2001, Karelson et.al 1996). After the calculation of molecular descriptors, linear methods, such as multiple linear regressions (MLR), can be used in the development of a mathematical relationship between the structural descriptors and the property to be predicted. Physical and thermodynamic property data of organic compounds such as boiling point are important in the engineering design and operation of industrial chemical processes. Since the experimental determination of boiling point is both time-consuming and expensive and there is increased need of reliable physical and thermodynamic data for the optimization of chemical processes, it would be very useful to develop predictive models that can be used to predict these properties of organic compounds that are not synthesized or their properties are unknown. The goal of the present study is to extend our previous investigation in order to, for the first time, establish a QSPR model that can predict the boiling point for set of organic compounds dependent only upon their molecular structures. Multiple linear regressions is applied to establish quantitative linear relationship between boiling point and molecular descriptors.

Experiment

Data set: All boiling point data in the present investigation were obtained from the CRC. The compounds include a diverse set of substituted alcohols. A complete list of the compounds names and corresponding experimental boiling point is shown in Table-1.

Molecular descriptor generation

The calculation of molecular descriptor is described as below: all molecular were drawn in to Hyperchem[8] and pre-optimized using MM+ molecular mechanics force field. A more precise optimization is done with semi-empirical PM3 method is Hyperchem and thereafter quantum chemical descriptors were obtained. The resulted geometry was then transferred into software Dragon[9] to calculated constitutional and topological descriptors. Constitutional descriptors are basically related to the number of atoms and bonds in each molecule topological descriptors include valence and non-valence molecular connectivity indices calculated from the hydrogen suppressed formula of the molecule encoding information about the size, composition and the degree of branching of a molecule. The quantum chemical descriptors include information about binding and formation energies, partial atom charge, dipole moment and molecular orbital energy levels.

Feature Selection

Once descriptors were garneted, descriptor-screening methods are used to select the most relevant descriptor to establish the models to predict the molecular property. Here the forward stepwise regression method was used to choose the subset of molecular descriptors. Forward stepwise regression stats with no model terms and at each step it adds the most statistically significant term (the one with the highest F- statistic or lowest p value) until there are none left. It was determined to be the best model when adding a descriptor no longer improved the cross- validation model.

Table 1 Boiling Point for Some Organic Compounds

Molecule	T_b) (exp)	T_b (MLR)	Residual
1- Propanol	97.20	90.05	7.15
2- Propanol	82.30	79.92	2.38
1- Butanol	117.73	115.86	1.87
2- Butanol	99.51	97.78	1.73
2- methaly1-1-propanol	107.89	109.21	-1.32
2- methaly1-2-propanol	82.40	86.87	-4.47
Cyclopentanol	140.20	138.25	2.17
1-Pentnol	137.98	137.25	0.17
2-Pentnol	119.30	128.38	-9.08
3-Pentnol	116.25	123.55	-7.30
2-Methy 1-1-butanol	128.00	118.10	9.90
3-Methy1-1-butanol	131.10	108.52	22.58
2-Methy1-2-butanol	102.40	96.47	5.93
3-Methy1-2-butanol	112.90	99.90	13.00
Pyclohexanol	160.84	173.26	-12.42
Phenol	181.81	194.80	-12.99
1-Hexanol	157.60	155.54	2.06
2-Hexanol	140.00	133.70	7.30
m-Cresol	202.27	203.93	-1.66
1-Heptanol	176.45	174.91	1.54
2,6-Xylenol	201.07	214.85	-13.78
3,5-Xylenol	221.74	212.10	9.64
1-Octanol	195.16	197.09	-1.93
1-Nonanol	213.37	218.20	-4.83
1-Decanol	231.10	240.33	-9.23

MLR= Multiple linear regression.

Regression analysis: After the descriptor was selected, multiple linear regressions was employed to develop the linear model of the property of interest, which takes the form:

$$Y=b_0+b_1x_1+b_2x_2+.....+b_nx_n$$

In this equation, Y is the property, that is, the dependent variable, $x_1$to x_n represent the specific descriptor, while b_1 to b_n represent the coefficient of those descriptor; b_0 is the intercept of this equation.

Results and Discussion

A great number of molecular descriptor, which encodes the electronic, geometric and topological features of the molecules, was calculated to describe the molecular structure. Forward stepwise regression routine implemented in SPSS is used to develop the linear model for the prediction of boiling point using calculated molecular descriptor. The best linear model contains six molecular descriptor. They are listed in Table-2.

Table 2 Specification of the Multiple Linear Regression Models

Descriptor	Notation	Coefficient	Mean effect
Narumi simple topological index (log)	SNAP	12.512	5.113
Number of ring quatemary c(sp^3)	$NCRH_2$	-2.094	-2.803
Maximal elector topological negative variation	MAXDN	-29.496	-10.384
Randic indices of different orders	X_4	11.177	7.652
Kier symmetry index	SOK	3.1	4.606
Path/walk 5-randic shape index	PW_5	113.154	2.957

Results of MLR model: R=0.999,S_F=2.08634, F=1868.157

The boiling point is determined by different interactions between molecules. These descriptors determined by one constitutional descriptor ($NCRH^2$) and 5 topological descriptors (SNAP,MAXDN,X4,SOK PW5).

In order to access the accuracy are predictability of the proposed model the cross-validation test was employed.

Results of cross-validation test : R =0.986,S_E=7.81053, F=778.07

Conclusion

QSPR models for the prediction of boiling point for some organic compounds using MLR based on descriptors calculated from molecular structure have been developed. Fig 1 shows the plot of the MLR calculated versus the ability of the MLR model in predication boiling point for some organic compounds.

QSPR Calculation of Boiling Point of Organic Compounds

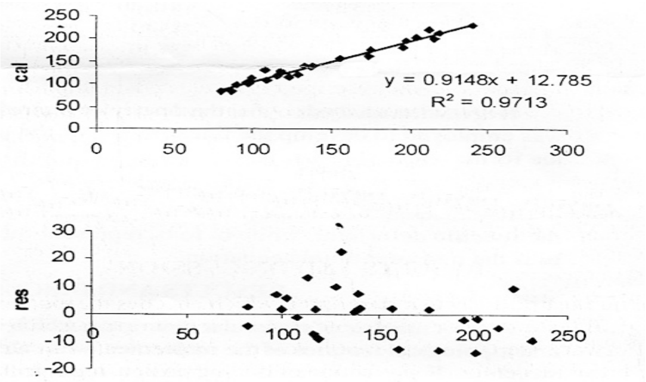

Fig 1. Plot of MLR calculated versus experimental values

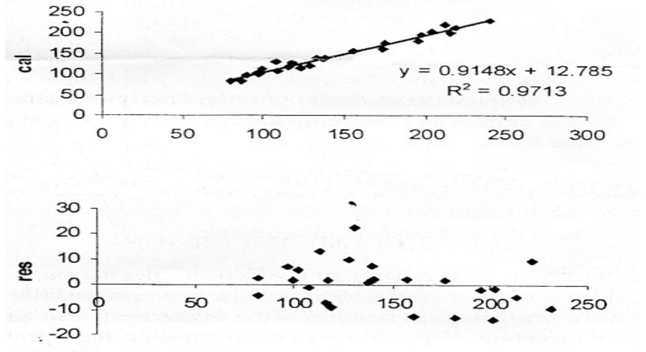

Fig. 2. Plot of the residuals versus experimental values

The propagation of residuals on both sides of zero indicates that no systematic error exists in the development of the multiple linear regression models.

In general, the good agreement between the experimental results and the predicated values using the multiple linear regression models confirms its validity.

References

Devillers, Y. and Balaban, A.T. (1999). in eds: Gordon and Breach, Topological Indices and Related descriptors in QSAR and QSPR, Amsterdam. The Netherlands (1999).

Estrda E. and Uriare, E. (2001). Curr. Med. **8**. 1699.

Hyprechem 4.0, Hypercubem Inc (1994).

Karelson, M., (2002). Molecular Descriptors in (QSAR/ QSP; John Wiley &Sons, New York.

Karelson, M., Lobanov V.S., and Katritzky, A.R. (1996). Chem Rev, **96**, 1027.

Katrizky, A.R., Maran,U., V.S. Lobanov and Karelson M., Vhem Y. (2000). Inf.Comput. Sci, **40**.I .

Katrizky, A.R., Petrukhin, R., Tatham, D., Basak, S., Benfenatim, E. Karelaon, M., and Maran, U., (2001). Y., Chem. Inf. Comput. Sci,**41**, 679.

Todeschichini, R., and Consonni, V., (2000). Handbook of Molecular Descriptors, Wiley-VCH: Eeinheim, Germany.

Todeschini, R. (2000). Dragon Software for the Calculation of the molecular Descriptors, Rel. 1.1 for Windows, Milano.