ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“ – БЪЛГАРИЯ
НАУЧНИ ТРУДОВЕ, ТОМ 33, КН. 5, 2005 – ХИМИЯ
UNIVERSITY OF PLOVDIV „PAISII HILENDARSKI“ – BULGARIA
SCIENTIFIC PAPERS, VOL. 33, BOOK 5, 2005 – CHEMISTRY

MODELING OF LOGP FOR HYDROCARBON COMPOUNDS

Nikolay T. Kochev, Ognyan Pukalov, George N. Andreev
Bulgaria, Plovdiv 4000, Tsar Assen St. 24, University of Plovdiv, Department of Analytical Chemistry

ABSTRACT

A method for calculation of partition coefficient LogP is introduced. The method is based on an atomic additive scheme. LogP value is obtained as a sum of individual contributions of each atom of the molecule. The atom contributions called increments correspond to atom classes derived on the base of local atomic properties. Increment values are calculated by means of linear regression applied for a set of hydrocarbon compounds. Test results are presented and discussed.

Keywords: QSPR, QSAR, additive scheme, linear regression

INTRODUCTION

One of the main problems chemists try to solve is to create chemical compounds with particular properties. This problem implies the fundamental tasks to find relationships between structures and their properties. Chemoinformatics brought these methods to a new level where quantitative characterization of these relationships is made by approaches known as QSPR/QSAR (Quantitative Structure Property/Activity Relationship). QSPR/QSAR models are generally based on the abstract equation:

$$\text{Biological Activity} = F(\text{Structure, parameters}) \qquad (1)$$

LogP is one of the widely used parameters in the QSAR modeling [1]. Knowing logP value for a particular compound is a must for calculating other important molecular characteristics. LogP is defined as a decimal logarithm of the ratio of equilibrium concentrations of particular compound in heterogenic system n-octanol/water:

$$\log P = \ln_{10}(C_{\text{n-octanol}}/C_{\text{water}}) \qquad (2)$$

The experimental methods for LogP determination have some disadvantages: significant technological time, sensitivity to polluted samples, complications with surface-active compounds. Another important argument is the fact that in many cases LogP should be determined for a compound which is not synthesized yet. Having in mind that there are about 30 000 compounds with known logP values and a number of 20000000 potential target compounds the need of a theoretical method for LogP calculation is obvious. There are several approaches for logP prediction. Most of them present logP as a linear or non-linear model of a set of different molecular descriptors: topological descriptors, 3D-descriptors, electron descriptors, quantum-mechanical descriptors. Additive modeling methods [2] are particular case of the linear ones. The modeled property is obtained additively by adding the contributions of each compound fragment (in this case having the role of a descriptor). An appropriate scheme should be created where the compound is fragmented and each fragment is assigned an increment value. In this paper we present an additive scheme successfully applied for modeling of logP.

LOGP ADDITIVE SCHEME

The molecule fragmentation scheme is based on atomic fragments. The latter are represented by categories of atoms where each atom category A[n,b2,b3,π] is defined by 4 parameters: A – atom type, n-number of neighbors, b2-number or double bonds, b3-number of triple bonds, π - number of pi neighbor atoms (i.e. atoms which participate in a pi-electron system). The logP model is obtained as follows:

$$\log P_{o/w} = \sum m_{A[n,b2,b3,\pi]} I_{A[n,b2,b3,\pi]} \qquad (3)$$

$I_{A[n,b2,b3,\pi]}$ is the increment (contribution) for category A[n,b2,b3,π]. $m_{A[n,b2,b3,\pi]}$ is the number of atoms of type A[n,b2,b3,π]. $m_{A[n,b2,b3,\pi]}$ is derived as a descriptor for each compound (it is directly obtained from the structure). $I_{A[n,b2,b3,\pi]}$ values determine the model and they are to be calculated by means of a linear regression applied for the training data set.

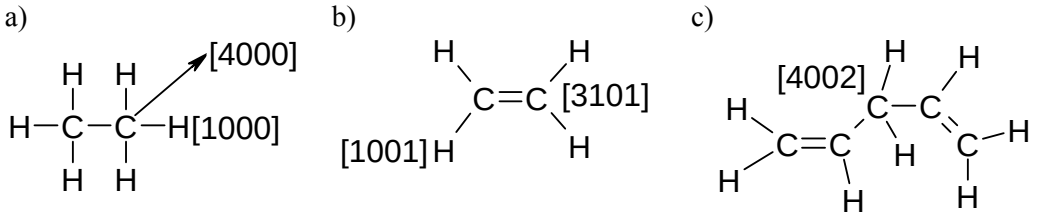

Figure 1. *Different types of atom categories (atomic fragments).*

Figure 1 illustrates some atom categories. In case a) there are only two types of atom fragments which are characteristic for all alkanes: H[1000] and C[4000]. For example C[4000] stands for a carbon atoms which has 4 neighbors (i.e. sp^3 hybridized carbon), 0 double bonds, 0 triple bonds and 0 pi neighbors. Case b) shows

the other possible category of hydrogen atom H[1001] i.e. hydrogen atom connected to a pi atom. Case c) demonstrates that atomic code describes the first atom layer of the atom and partially the second one. C[4002] stands for a sp^3 hybridized carbon which is connected with two other sp^2 or sp hybridized carbon atoms (i.e. in the second atom layer there are double or triple bonds). Carbons participating in triple bonds can be described by the code C[2010]. A special simplified case of eq. (3) can be obtained for all alkanes having n carbon atoms (C_nH_{2n+2}):

$$\log P = n.I_{C[4000]} + (2n+2).I_{H[1000]} \qquad (4)$$

MODEL CREATION

The model was created with software JBSMM (Java Based System for Molecular Modelling) developed in our laboratory. JBSMM is specially designed for representation of structures, for calculation of molecular descriptors and creation of models of the type of eq. (1). To obtain the logP model from eq. (3) the values of increments $I_{A[n,b2,b3,\pi]}$ have to be determined. Next figure summarises the model creation process.

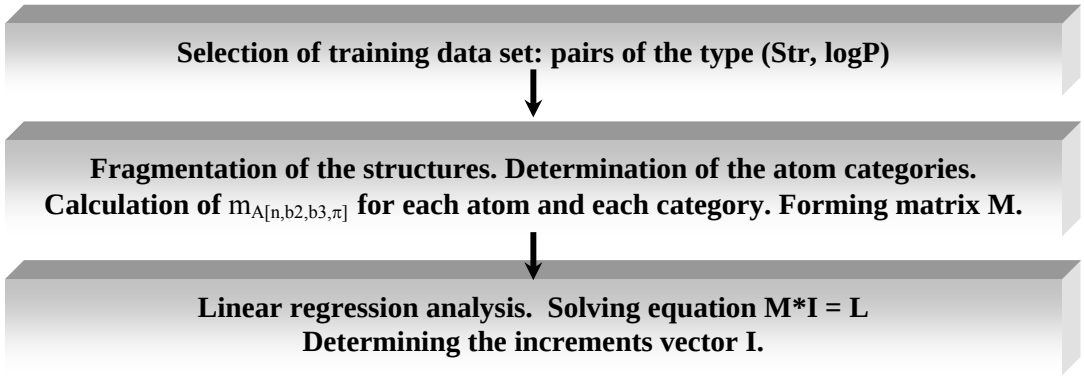

Figure 2. *Flow chart of the model creation with the aid of software JBSMM.*

Training data set was formed by 40 hydrocarbon compounds (alkanes, alkenes, alkynes and aromatics). The logP experimental values were taken from SRC PhysProp online database [3].

RESULTS AND DISCUSSION

Table 1. *Increment values for the logP model*

Atom	H[1000]	H[1001]	C[2010]	C[3101]	C[3102]	C[3103]	C[4000]	C[4001]	C[4002]
Incr.	0.317	-0.225	-0.132	1.081	0.578	0.175	-0.108	-0.667	-1.153

Table 1 shows the obtained increment values. They can be used directly to predict logP values. For example the model can be applied for ethene (see Fig. 1b) as follows: ethene structure is fragmented to 4 H[1001] atoms and 2 C[3101] atoms hence $\log P = 4I_{H[1001]} + 2I_{C[3101]} = 4*(-0.225) + 2*1.081 = 1.262$.

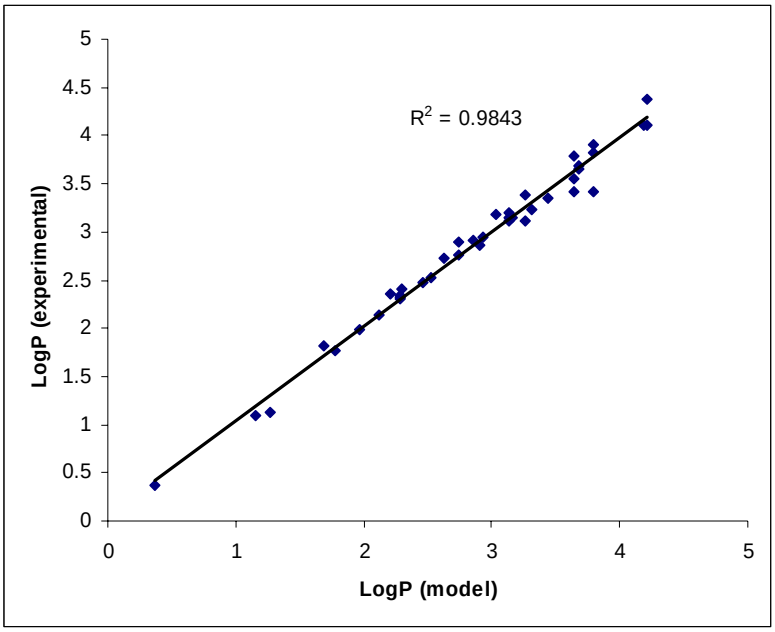

Figure 3. *Comparison of logP experimental and modelled values.*

Figure 3 represents the comparison of the logP modeled values and the experimental values. The square of the correlation coefficient is very high $R^2 = 0.98$. The root mean squire error of the model is $RMS_{Error} = 0.11$. Both statistical parameters demonstrate the efficiency of the model which can be easily applied even without a computer. The model predicts logP values close enough to experimental ones. As it can be seen from equation (4) the additive scheme does not recognize the different alkane isomers. The latter should not be considered as a big disadvantage since in most of the cases the experimental logP values of the isomers differ each other with values smaller than the mean model error.

As a future development, the presented logP model could be extended with fragment categories describing hetero atoms. Then logP values could be predicted for a wider set of organic compounds.

ACKNOWLEDGEMENTS

We would like to thank Plovdiv University scientific Fund (project 05-X-64) for supporting this scientific work.

REFERENCES

1. Gombar V., Enslein K. Assessment Of N-Octanol/Water Partition Coefficient: When The Assessment Relieble J. Chem. Inf. Comput. Sci. 1996, 36, 1127 – 1134.
2. Miller K. Additivity Methods In Molecular Polarizabiliti J. Am. Chem. Soc. 1990, 112, 8533–8542.
3. Syracuse Research Corporation: PhysProp online data base (http://www.syrres.com).