Available online at www.pelagiaresearchlibrary.com

Pelagia Research Library

Der Chemica Sinica, 2011, 2(6):351-358

Pelagia Research Library
ISSN: 0976-8505
CODEN (USA) CSHIA5

QSPR study for estimation of Kovats retention indices of some of Adamantane derivatives using genetic algorithm (GA) and multiple linear regression (MLR) analysis by Hartree –Fock method

M. Fakoor Yzdan Abad and Z. Bayat*

Department of Chemistry, Islamic Azad University-Quchan Branch, Iran

---

ABSTRACT

*The Kovats retention indices values are expressed as a important property of Adamantane derivatives.A linear quantitative structure–Property relationship (QSPR) model is presented for the modelling and prediction for the Kovats retention indices of Adamantane derivatives. The model was produced using the multiple linear regression (MLR) technique on a database that consisted of 65 adamantane derivatives compounds. Among the different constitutional, topological, geometrical, electrostatic and quantum-chemical descriptors that were considered as inputs to the model, seven variables were selected using the genetic algorithm subset selection method (GA). A multi-parametric equation containing maximum two descriptors at the Hartree–Fock level with 6-31+G** basis set , with good statistical qualities ($R2_{train}$=0.922 , F_{train}=109.038, $Q2_{LOO}$=0.904 , $R2_{adj}$=0.914, $Q2_{LGO}$=0.863) was obtained by Multiple Linear Regression using stepwise method.The accuracy of the proposed MLR model was illustrated using the following evaluation techniques: cross-validation, validation through an external test set, and Y-randomisation. The predictive ability of the model was found to be satisfactory and could be used for designing a similar group of compounds.*

Keyword: Adamantane derivatives, Kovats retention indices, genetic algorithm, Multiple linear regressions, Hartree–Fock.

INTRODUCTION

Quantitative structure property relationships (QSPR), mathematical equations relating chemical properties such as acidity, electrochemistry, reactivity and chromatographic behavior to a wide variety of structural, topological and electronic features of the molecules [1], have been widely used in the field of chromatographic sciences [2–9]. Quantitative structure–retention relationships (QSSRs) represent statistical models which quantify the relation between the structure of the molecule and chromatographic retention indices of the compound, allowing the prediction of retention indices of novel compounds. QSPR on the Kovats retention indices have been reported for different types of organic compounds [10–14]. The success of a QSAR study

depends on choosing robust statistical methods for producing the predictive model and also the relevant structural parameters for expressing the essential features within those chemical structures. Nowadays, genetic algorithms (GA) are well known as interesting and widely used methods for variable selection [15]. GA are stochastic methods used to solve the optimisation problems defined by the fitness criteria, applying the evolutionary hypothesis of Darwin and also different genetic functions i.e. crossover and mutation. In this paper , we have used a genetic algorithm for the variable selection, and developed an MLR model for the QSPR analysis of the Adamantane derivatives. In a QSPR study the model must be validated for its predictive value before it can be used to predict the response of additiona chemicals. Validating QSPR with external data (i.e. data not used in the model development), although demanding , is the best method for validation [16–17]. However the availability of an independent external validation set of several compounds is rare in QSPR. Thus, the input data set must be adequately split by experimental design or other splitting procedures into representative training and validation/test sets [18–20]. In the present work, the data splitting was performed randomly and was confirmed by the factor spaces of the descriptors. Finally, the accuracy of the proposed model was illustrated using the following: leave one out, bootstrapping and external test set, cross-validations and Y-randomisation techniques.

MATERIALS AND METHODS

Data set

In this study, the data set of 65 Adamantane derivatives were studied .The data set was randomly divided into two subsets: the training set containing 52 compounds (80%) and the test set containing 13 compounds (20%). The training set was used to build a regression model, and the test set was used to evaluate the predictive ability of the model obtained.

Molecular descriptor generation

All of the molecules were drawn into the HyperChem (Version8.03 Hypercube) software and pre-optimised. The molecular structures were optimised using the Gaussian 03. The Gaussian 03 [21] was used for calculating the molecular descriptors.These descriptors could represent a variety of aspects of the compounds, and have been successfully used in various QSAR and QSPR research. Any descriptors with a constant or almost constant value for all the molecules were eliminated. Also, any pairs of variables with a correlation coefficient greater than 0.90 were classified as inter-correlated, and only one of them was considered in developing the model. A total 92 descriptors were considered for further investigations after discarding the descriptors with constant values and the ones that were inter-correlated.

RESULTS AND DISCUSSION

In a QSAR study, generally, the quality of a model is expressed by its fitting ability and prediction ability, and of these the prediction ability is the more important. In order to build and test the model, a data set of 65 compounds was separated into a training set of 52 compounds, which were used to build the model and a test set of 13 compounds, which were applied to test the built model. With the selected descriptors, we have built a linear model using the training set data, and the following equation was obtained:

$RI = -2954.55\ (\pm 1219.061)\ EP_5 - 5.39879\ (\pm 1219.061)\ \sigma_9 - 73.0629\ (\pm 9.99241)\ \Delta G_{CYCLO+}$
$5.362559\ (\pm 0.250731)\ M+ \ 0.048231\ (\pm 0.013282)\ HF- \ 43237.4(\pm 18017.96)(\ HF/6\text{-}31+G^{**})$

R^2_{train}=0.922, F_{train}=109.038 , R^2_{test}=0.848 F_{test}=4.350 , R^2_{adj}=0.914 Q^2_{LOO}=0.90431, Q^2_{LGO}= 0.862 N_{train}= 52, N_{test} = 13

In this equation, N is the number of compounds, R^2 is the squared correlation coefficient, Q^2_{LOO}, Q^2_{BOOT} and Q^2_{ext} are the squared cross-validation coefficients for leave one out, bootstrapping and external test set respectively, RMSE is the root mean square error and F is the Fisher F statistic. In the present study, the QSPR model was generated using a training set of 52 molecules. The test set of 13 molecules was used to assess the predictive ability of the QSPR model produced in the regression.The built model was used to predict the test set data and the prediction results are given in Table 1. and the test results are given in Table 2.

Table 1. the corresponding observed and predicted RI values by the MLR method

Name	EXP.	Pred	Ref
Adamantane	1118	1150.02	22
1 3 dimethyl adamantane	1151	1216.30	22
1-fluoro adamantane	1159	1220.30	22
222-methylene adamantane	1160	1200.18	22
1,3,5 -trimethyl adamantane	1163	1196.94	22
2-methyl adamantane	1196	1228.87	2222
1 2-dimethyl adamantane	1236	1251.24	22
1-ethyl adamantane	1260	1235.70	22
2 2-dimethyl adamantane	1269	1254.24	22
1-ethyl-3,5 di methyl adamantane	1279	1258.32	22
3-ethyl-1-adamantanol	1283	1379.77	22
3-methyl-1-adamantanol	1283	1316.35	22
3 5-dimethyl-1-adamantanol	1295	1292.21	22
1-chloroadamantane	1298	1229.81	22
3,5,7-trimethyl-1-adamantanol	1304	1327.72	22
2-adamantanon	1320	1298.76	22
2-chloro adamantane	1342	1333.38	22
1-propyl adamantane	1347	1311.47	22
2-methyl-2-adamantanol	1348	1397.78	22
2-isopropyl adamantane	1349	1337.58	22
2-propyl adamantane	1371	1361.32	22
1-bromo adamantane	1382	1405.31	22
1-hydroxy methyl adamantane	1402	1378.07	22
1-chloromethyladamantane	1404	1331.85	22
2-isobuthyl adamantane	1416	1393.56	22
3-ethyl-5,7-dimethyl -1-adamantanol	1421	1386.82	22
3-5 dimethyl 1 hydroxy methyl adamantane	1425	1399.76	22
5-7-dimethyl1-3 adamantandiol1.	1438	1434.22	22
1-buthyl adamantane	1443	1475.02	22
methyl-(1-adamanthyl) ketone	1443	1401.97	22
methyl-(2-adamanthyl)ketone	1445	1407.23	22
2-ethyl-2-adamantanol	1446	1424.14	22
2-buthyl adamantane	1465	1440.08	22

Adamantane-2-carboxylic acid methyl ester	1467	1475.90	22
methyl ester of 3,5 di methyl adamantane1-carboxilic acid	1467	1490.51	22
1-bromomethyl adamantane	1488	1497.74	22
2-methyl-1-hydroxy methyl adamantane	1490	1440.80	22
3-isopropyl-1-adamantanol	1506	1433.87	22
Adamantane -1-carboxylic acid ethyl ester	1508	1562.30	22
Methyl esters of 2-methyl adamantane -1-carboxylic acid	1512	1519.59	22
Adamantane-2-carboxylic acid ethyl ester	1529	1563.23	22
ethyl-(1-adamanthyl)ketone	1529	1489.78	22
Adamantane-1-carboxylic acid iso propyl ester	1532	1600.17	22
Adamantane-1-carboxylic acid tert-buthyl ester	1556	1654.40	22
2-isobuthyl-2-adamantanol	1570	1529.36	22
Methyl ester of -3-ethyl adamantane -1-carboxylic asid	1579	1573.66	22
3-buthyy-1-adamantanol	1595	1540.40	22
esters of adamantane 1-carboxylic acid propyl ester	1603	1619.45	22
2-buthyl-2-adamantanol	1620	1575.30	22
Adamantane -1-carboxylic acid sec-buthyl ester	1631	1638.87	22
Adamantane -1-carboxylic acid iso buthyl ester	1658	1615.18	22
Di methyl ester of 5,7-di methyl adamantane -1-3 di carboxylic acid	1769	1773.14	22

Table 2. the corresponding observed and Test RI values by the MLR method

Name	EX	Test	Ref
1-methyladamantane	1137	1170.27	22
2-fluoro adamantane	1182	1284.40	22
1-adamantanol	1268	1245.38	22
2-ethyl adamantane	1284	1287.60	22
2-adamantanol	1329	1341.85	22
1-isopropyl adamantane	1358	1310.92	22
3 5-dimethyl -1-bromo adamantane	1401	1420.78	22
2-bromoadamantane	1426	1479.65	22
esters of adamantane1-carboxylic acid methyl ester	1449	1393.90	22
3-propyl-1-adamantanol	1495	1458.60	22
2-propyl-2-adamantanol	1526	1494.15	22
3-(1-adamanthyl)pentane	1559	1423.37	22
propyl-(1-adamanthyl) ketone	1609	1536.92	22

As can be seen from Table 1, the calculated values for the RI are in good agreement with those of the experimental values. The predicted values for RI for the compounds in the training and test sets using equation ,were plotted against the experimental RI values in Figure 1 . The Graph of experimental verses the predicted values for the present RI model and the comparison between Retention Index using prediction and the experimental are shown in Figure 2.

Figure 1. The predicted versus the experimental RI by GA-MLR

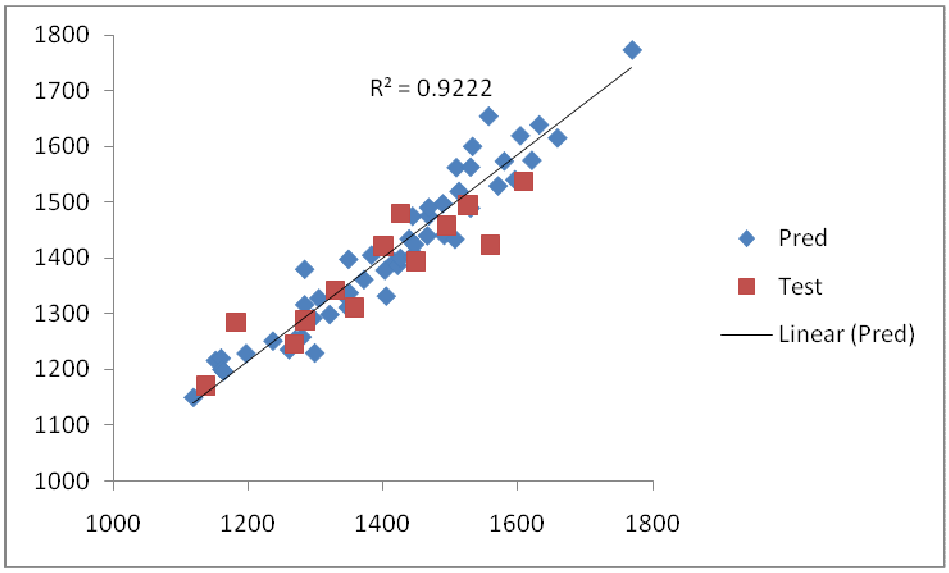
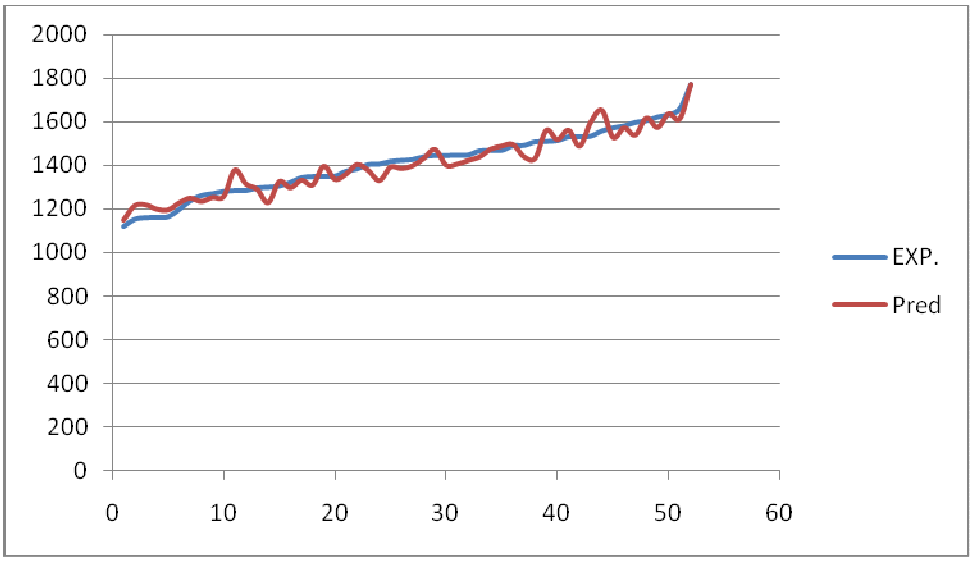


Figure 2. the comparison between RI using Predection and the Experimental.



As can be seen the model did not show any proportional and systematic error, because the propagation of the residuals on both sides of zero are random.The real usefulness of QSAR models is not just their ability to reproduce known data, verified by their fitting power (R^2), but is mainly their potential for predictive application. For this reason the model calculations were performed by maximising the explained variance in prediction, verified by the leave-one-out cross-validated correlation coefficientQ^2_{LOO}. To avoid the danger of overfitting and the possibility of overestimating the model predictivity by using Q^2_{LOO}, and Q^2_{ext} ,the internal predictive ability of the models was also verified using the bootstrap Q^2_{BOOT} procedure, as is strongly recommended for QSAR modeling. The robustness of the proposed models and their predictive ability was guaranteed by the high Q^2_{BOOT} based on the bootstrapping being repeated5000 times. The Q^2_{LOO}, Q^2_{ext} and $Q2_{BOOT}$ for the MLR model are shown in Equation. This indicates that the obtained regression model has a good internal and external predictive power. Also, in order to assess the robustness of the model, the Y-randomisation test was applied in this study. The dependent variable vector (RI) was randomly shuffled and a new QSPR model

developed using the original independent variable matrix. The new QSAR models (after several repetitions) would be expected to have low R^2 and Q^2_{LOO} values(Table 3). If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modelling method and data.

Table 3. The $R2_{train}$ and $Q2_{LOO}$ values after several Y-randomisation tests.

NO	Q^2	R^2
1	0.019532	0.073782
2	0.052444	0.216412
3	0.002706	0.130557
4	0.02119	0.162731
5	0.034956	0.058804
6	0.0661	0.043855
7	0.001646	0.099392
8	0.335116	0.011723
9	0.017008	0.139496
10	0.011897	0.068495

The $R2_{train}$ and $Q2_{LOO}$ values after several Y-randomisation tests.Interpretation of descriptors As well as demonstrating statistical significance, QSAR models should also provide useful chemical insights into the mechanism of inhibitory activity. For this reason, an acceptable interpretation of the QSAR results is provided below**.** The linear model based on the seven parameters selected by the GA-MLR method.The negative sign suggests that the RI value is inversely related to this descriptor. The linear model based on the seven parameters selected by the GA-MLR method. Commonly used descriptors in the QSAR analysis are presented in Table 4.

Table 4. The calculated descriptors used in this study.

	Symbol	Example
Descriptor	Molecular Polarizability	MP
	Electrostatic Potentialc	EP
	solvation Free Energy(in Octanol)	ΔGOCT
	Salvation Free Energy (in Cyclohexane)	ΔG_{Cyclo}
	Isotropic Parameterd	σ
	Mulliken Charge	MC
	Partition Coefficient	Log P
	Molecule surface area	SA
	Hydration Energy	HE
	Refractivity	REF
	Mass	M

CONCLUSION

Predictive QSPR model which is based on molecular descriptors is proposed in this study to correlate the Kovats retention indices of some of Adamantane derivatives. Application of the developed model to a testing set of 13 compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation. Since the QSPR was developed on the basis of theoretical molecular descriptors calculated exclusively from molecular structure, the proposed model could potentially provide useful information about the retention indices of

Adamantane derivatives. We have developed here a useful QSPR equation derived from theoretical descriptors associated with retention indices property.a MLR is successfully presented for prediction retention indices property (RI) of various compounds with diverse chemical structures using a linear quantitative structure– property relationship. A model with high statistical quality and low prediction errors was obtained. The model could predict the retention indices property of the compounds accurately. Development of quantitative structure property/activity relationships (QSPR/QSAR) on theoretical descriptors is a powerful tool not only for prediction of the chemical, physical and biological properties/activities of compounds. MLR analysis provided useful equation that can be used to predict the RI of chemicals based upon Electrostatic Potentialc, Isotropic Parameterd , Hydration Energy,Mass,Salvation Free Energy in Cyclohexane parameters. This model allowed us to achieve a precise and relatively fast method for determination of RI of different series of Adamantane Derivatives and to predict with sufficient accuracy the RI of new drug derivatives.

REFERENCES

[1]. Li X. IRAK4 in TLR/IL-1R *Eur J Immunol* **2008**;38:614–618.
[2]. Buckley GM, Ceska TA, Fraser JL, Gowers L, Groom CR, Higueruelo AP, Jenkins K, Mack SR, Morgan T, Parry DM, Pitt WR, Rausch O, Richard MD, Sabin V. *Bioorg Med Chem Lett* **2008**;18:3291–3295.
[3]. Buckley GM, Fosbeary R, Fraser JL, Gowers L, Higueruelo AP, James LA, Jenkins K, Mack SR, Morgan T, Parry DM, Pitt WR, Rausch O, Richard MD,Sabin V. *Bioorg Med Chem Lett* **2008**;18:3656–3660.
[4]. Buckley GM, Gowers L, Higueruelo AP, Jenkins K, Mack SR, Morgan T,Parry DM, Pitt WR, Rausch O, Richard MD, Sabin V, Fraser JL. *Bioorg Med Chem Lett* **2008**;18:3211–3214.
[5]. Sammes PG, Taylor JB. Comprehensive Medicinal Chemistry. Oxford: Pergamon Press, **1990**:766.
[6]. Riahi S, Pourbasheer E, Dinarvand R, Ganjali MR, Norouzi P. *Chem Biol Drug Des* **2008**;74:165–172.
[7]. Riahi S, Pourbasheer E, Ganjali MR, Norouzi P. *J Hazard Mater* **2009**;166:853–859.
[8]. Riahi S, Pourbasheer E, Ganjali MR, Norouzi P. *Chem Biol Drug Des* **2009**;73:558–571.
[9]. Depczynski U, Frost VJ, Molt K. *Anal Chim Acta* **2000**;420:217.
[10]. Alsberg BK, Marchand-Geneste N, King RD. Chemometr Intel Lab **2000**;54:75–91.
[11]. Jouanrimbaud D, Massart DL, Leardi R, Denoord OE. *Anal Chem* **1995**;67:4295–4301.
[12]. Riahi S, Ganjali MR, E Pourbasheer, Divsar F, Norouzi P, Chaloosi M. *Curr Pharm Anal* **2008**;4:231–237.
[13]. Riahi S, Ganjali MR, Pourbasheer E, Norouzi P. *Chromatographia* **2008**;67:917–922.
[14]. Riahi S, Pourbasheer E, Ganjali MR, Norouzi P, Zeraatkar Moghaddam A. *J Chin Chem Soc* **2008**;55:1086–1093.
[15]. Riahi S, Ganjali MR, Moghaddam AB, Pourbasheer E, Norouzi P. *Curr Anal Chem* **2009**; 5: 42–47.
[16]. Tropsha A, Gramatica P, Gombar VK. *QSAR Comb Sci* **2003**;22:69–77.
[17]. Riahi S, Ganjali MR, Norouzi P, Jafari F. *Sens. Actuators B* **2008**;132:13–19.
[18]. Eriksson L, Johansson E, Muller M, Wold S. *J Chemometr* **2000**;14:599–616.
[19]. Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropsha A. *J Comput Aided Mol Des* **2003**;17:241–253.*Journal of Enzyme Inhibition and Medicinal Chemistry* Downloaded from informahealthcare.com by Tehran University on 04/30/10For personal use only. 10 *Eslam Pourbasheer et al.*
[20]. Gramatica P, Pilutti P, Papa E. Validated QSAR prediction of OH tropospheric degradability: splitting into training-test set and consensus modeling. **2004**;44:1794–1802.

[21]. Todeschini R, Consonni V, Pavana M. http://www.disat.unimib.it/chm/.
[22] Burkhard, J., Vais, J., Vodicka L. and Landa S.: *Journal of Chromatography*. *Chrom*. 4057.