Available on line www.jocpr.com

Journal of Chemical and Pharmaceutical Research

ISSN No: 0975-7384
CODEN(USA): JCPRC5

J. Chem. Pharm. Res., 2011, 3(1):48-55

Computational Approaches to the Predication of the Kovats Retention Index(RI) for Adamantane Derivative (AD) as a drug

Z. Bayat*and M. Fakoor Yazdan Abad

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

ABSTRACT

*A quantitative structure–property relationship (QSPR) study was performed to develop models those relate the structures of 32 Kovats retention index (RI) of AD.Molecular descriptors derived solely from 3D structures of the molecular compounds. A genetic algorithm was also applied as a variable selection tool in QSPR analysis. The models were constructed using 25 molecules as training set, and predictive ability tested using 7 compounds. Modeling of RI of ADD as a function of the theoretically derived descriptors was established by multiple linear regression (MLR). The usefulness of the quantum chemical descriptors, calculated at the level of the HF theories using 6-31+G** basis set for QSAR study of AD was examined. The use of descriptors calculated only from molecular structure eliminates the need for experimental determination of properties for use in the correlation and allows for the estimation of RI for molecules not yet synthesized. Application of the developed model to testing set of 7 drug organic compounds demonstrates that the model is reliable with goo predictive accuracy and simple formulation. The prediction results are in good agreement with the experimental value. A multi-parametric equation containing maximum Four descriptors at B3LYP/6-31+G** method with good statistical qualities (R^2_{train}=0.922, F_{train}=109.04, R^2_{test}=0.848, F_{test}=4.35, Q^2_{LOO}=0.904, R^2_{adj}=0.914,Q^2_{LGO}=0.862) was obtained by Multiple Linear Regression using stepwise method.*

Keywords: Adamantane derivative, drug, Kovats retention indices(RI), genetic algorithm,MLR,QSPR,HF

INTRODUCTION

Diamondoids are classed with organic nanostructures; therefore, AD have become particularly popular with the development of nanotechnologies. The applications of AD are diverse: from antiviral drugs to nanorobots and molecular machines [1-3]. Particular attention is given to the

chromatographic behavior of AD, because various chromatographic methods allow not only separation of multi component mixtures of isomers and structurally related framework hydrocarbons and their derivatives, but also qualitative and quantitative analysis of these mixtures [4]. 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 [5], have been widely used in the field of chromatographic sciences [6–13]. 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 RI have been reported for different types of organic compounds. Acevedo-Mart´ınez et al. [14-18], developed linear and nonlinear models to study the Kovats retention indices of the immine family using topological, topographical and quantum chemical descriptors Correlations between the sorption and structural characteristics of AD were found based on the QSPR method. 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 [19]. In a QSAR study the model must be validated for its predictive value before it can be used to predict the response of additional chemicals. Validating QSPR with external data (i.e. data not used in the model development), although demanding, is the best method for validation [20–21]. 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.

EXPERIMENTAL SECTION

Methodology

Data set

The properties data used in this study are the n Kovats retention index (RI) of of the set of 32 AD [22]. The data set was randomly divided into two subsets: the training set containing 25 compounds (80%) and the test set containing 7 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. The properties data for the complete set of compounds are presented in Table 1and 2.To derive QSAR models, an appropriate representation of the chemical structure is necessary. For this purpose, descriptors of the structure are commonly used.

Table 1. Experimental values of RI for AD training set

Name	EXP.	Pred.	Ref.
Adamantane	1118	1150	22
1 3 dimethyl adamantine	1151	1216	22
1-fluoro adamantine	1159	1220	22
2-methylene adamantine	1160	1200	22
1,3,5 -trimethyl adamantine	1163	1196	22
2-methyl adamantine	1196	1228	22
1 2-dimethyl adamantine	1236	1251	22
1-ethyl adamantine	1260	1235	22
2 2-dimethyl adamantine	1269	1254	22
1-ethyl-3,5 di methyl adamantine	1279	1258	22

1-chloroadamantane	1298	1229	22
2-adamantanon	1320	1298	22
2-chloro adamantine	1342	1333	22
1-propyl adamantine	1347	1311	22
2-isopropyl adamantine	1349	1337	22
2-propyl adamantine	1371	1361	22
1-bromo adamantine	1382	1405	22
1-chloromethyladamantane	1404	1331	22
2-isobuthyl adamantine	1416	1393	22
1-buthyl adamantine	1443	1475	22
methyl-(1-adamanthyl) ketone	1443	1401	22
methyl-(2-adamanthyl)ketone	1445	1407	22
2-buthyl adamantine	1465	1440	22
1-bromomethyl adamantine	1488	1497	22
ethyl-(1-adamanthyl)ketone	1529	1489	22

Table 2. Experimental values of RI for AD test set

Name	EXP	Test	Ref.
1-methyladamantane	1137	1170	22
2-ethyl adamantine	1284	1287	22
1-isopropyl adamantine	1358	1310	22
3, 5-dimethyl -1-bromo adamantine	1401	1420	22
2-bromoadamantane	1426	1479	22
3-(1-adamanthyl)pentane	1559	1423	22
propyl-(1-adamanthyl) ketone	1609	1536	22

Molecular descriptor generation

To derive QSAR models, an appropriate representation of the chemical structure is necessary. For this purpose, descriptors of the structure are commonly used. These descriptors are generally understood as being any term, index or parameter conveying structure information. Commonly used descriptors in the QSAR analysis are presented in Table 3. Some of the descriptors are obtained directly from the chemical structure, e. g. constitutional, geometrical, and topological descriptors. Other chemical and physicochemical properties were determined by the chemical structure (lipophilicity, hydrophilicity descriptors, electronic descriptors, energies of interaction). In this work, we used Gaussian 03 for ab initio calculations. HF method at 6-31+G** were applied for optimization of Adamantane derivatives and calculation of many of the descriptors. At first AD were built by Hyperchem software and some o the descriptors such as surface area, hydration energy, and refractivity were calculated through it. The rest of the descriptors were obtained of Gaussian calculations. A large number of descriptors were calculated by Gaussian package and Hyperchem software. One way to avoid data redundancy is to exclude descriptors that are highly intercorrelated with each other before performing statistical analysis. Reduced multi collinearity and redundancy in the data will facilitate selection of relevant variables and models for the investigated endpoint. Variable-selection for the QSAR modeling was carried out by stepwise linear regression method. A stepwise technique was employed that only one parameter at a time was added to a model and always in the order of most significant to least significant in terms of F-test values. Statistical parameters were calculated subsequently for each step in the process, so the significance of the added parameter could be verified.

Table 3. The calculated descriptors used in this study

Descriptors	Symbol	Abbreviation	Descriptors	Symbol	Abbreviation
Quantum chemical descriptors	Molecular Dipole Moment	MDP	Quantum chemical descriptors	difference between LUMO and HOMO	E_{GAP}
	Molecular Polarizability	MP		Hardness [η=1/2 (HOMO+LUMO)]	H
	Natural Population Analysis	NPA		Softness (S=1/ η)	S
	Electrostatic Potentialc	EP		Electro negativity [χ= -1/2 (HOMO–LUMO)]	*X*
	Highest Occupied Molecular Orbital	HOMO		El Electro philicity ($\omega=\chi^2$ /2 η)	Ω
	Lowest Unoccupied Molecular Orbital	LUMO		Mullikenl Chargeg	MC
	Partition Coefficient	Log P		Molecule surface area	SA
Chemical properties	Mass	M	Chemical properties	Hydration Energy	HE
	Molecule volume	V		Refractivity	REF

Genetic algorithm

Genetic algorithms (GAs) are governed by biological evolution rules [23]. These are stochastic optimisation methods that have been inspired by evolutionary principles. The distinctive aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores a different region in the parameter of space [24]. To select the most relevant descriptors, the evolution of the population was simulated [25-27]. The first generation population was randomly selected; each individual member in the population was defined by a chromosome of binary values and represented a subset of descriptors. The number of the genes at each chromosome was equal to the number of the descriptors. A gene was given the value of 1, if its corresponding descriptor was included in the subset; otherwise, it was given the value of zero. The number of the genes with the value of 1 was kept relatively low to have a small subset of descriptors [28]. As a result, the probability of generating 0 for a gene was set greater (at least 60 %) than the value of 1. The operators used here were the crossover and mutation operators. The application probability of these operators was varied linearly with a generation renewal (0–0.1 % for mutation and 60–90 % for crossover). The population size was varied between 50 and 250 for the different GA runs. For a typical run, the evolution of the generation was stopped when 90% of the generations took the same fitness. The fitness function used here was the leave-one-out cross-validated correlation coefficient, Q^2_{LOO}. The GA program was written in Matlab 6.5 [29].

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 25 compounds, which were used to build the model and a test set of 7 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 (±1219.061) EP_5 - 5.39879 (± 1219.061) σ_9 - 73.0629 (±9.99241) ΔG_{CYCLO} + 5.362559 (±0.250731) M+ 0.048231 (± 0.013282) HF- 43237.4(±180) (HF/6-31+G**)

R^2_{train}=0.922 F_{train}=109.038 R^2_{test}=0.848 F_{test} = 4.35 R^2_{adj}=0.914
Q^2_{LOO}=0.904 Q^2_{LGO}= 0.862 N_{train}= 25, N_{test} = 7

In this equation, N is the number of compounds, R^2 is the squared correlation coefficient, Q^2_{LOO} and Q^2_{LGO} 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. The built model was used to predict the test set data. The prediction and the test results are given in Table 1and Table 2 respectively. The predicted values for RI for the compounds in the training and test sets using equation RI were plotted against the experimental RI values in Figure 1.and the comparison between Retention Index using prediction and the experimental .A plot of the residual for the predicted values of RI for both the training and test sets against the experimental RI values are shown in Figure 2. As can beseen 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 coefficient,Q^2_{LOO}.

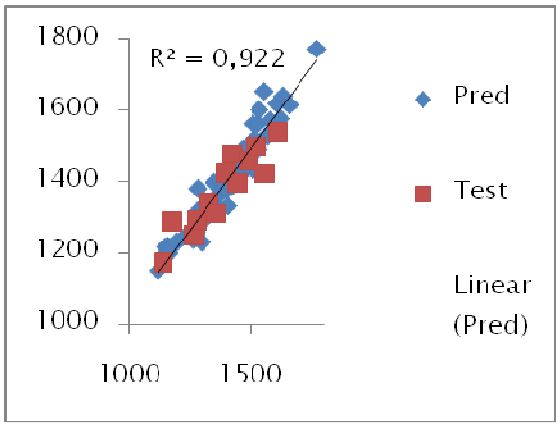


Figure1. The predicted versus the experimental RI by MLR

maximising the explained variance in prediction, verified by the leave-one-out cross-validated correlation coefficient, Q^2LOO. To avoid the danger of over fitting and the possibility of overestimating the model predictivity by using Q^2LOO procedure, as is strongly recommended for QSAR modeling. The Q^2 LOO and Q^2 GLO for the MLR model are shown in Equation RI. This indicates that the obtained regression model has a good internal and external predictive power.

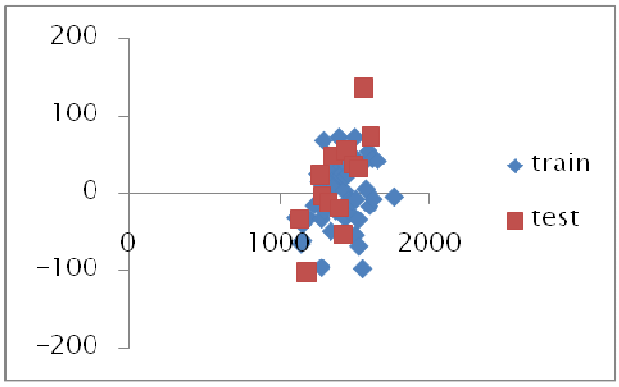


Figure 2. The residual versus the experimental RI by GA-MLR. (See colour version of this figure online at www.informahealthcare.com/enz)

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 The new QSAR models (after several repetitions) would be expected to have low R^2 and Q^2_{LOO} values (Table 4). If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

Table 4. The R^2_{train} and Q^2_{LOO} values after several Y-randomisation tests

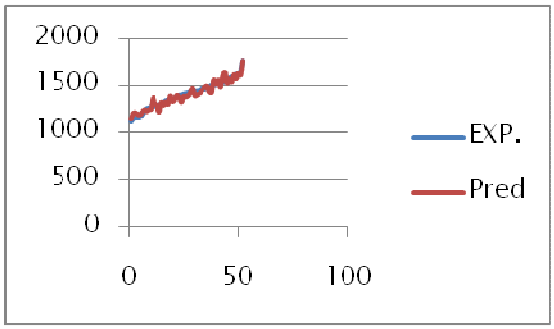
NO	Q^2	R^2
1	0.019532	0.073782
2	0.052444	0.216412
3	0.002706	0.130557
4	0.02119	0.162731
5	0.034956	0.058804
6	0.0661	0.043855
7	0.001646	0.099392
8	0.335116	0.011723
9	0.017008	0.139496
10	0.011897	0.068495

The MLR analysis was employed to derive the QSAR models for different AD. MLR and correlation analyses were carried out by the statistics software SPSS (Table 5).

Table 5.The correlation coefficient existing between the variables used in different MLR and equations with HF/6-31+G** method.

HF	**HF**	**M**	**ΔG_{CYCLO}**	**σ_9**	**EP5**
M	1	0	0	0	0
ΔG_{CYCLO}	0.36896	1	0	0	0
σ_9	0.125324	0.101574	1	0	0
EP5	0.12166	0.05804	0.11331	1	0
HF	0.31452	0.277738	0.36183	0.627972	1

Figure 3 has showed that results were obtained from equation B3LYP/6-31G* to the experimental values.



Series 1: the values of RI were obtained by using prediction.
Series 2: the values of RI were obtained by using Experimental methods
Figure 3. The comparison between properties (RI) using experimental and prediction

Interpretation of descriptors

The QSPR developed indicated that Nuclear magnetic Resonance (σ_9), free energy solvation (ΔG_{CYCLO}), electrostatic potential (EP_5) and Hartee-fuck energy (HF) compound Kovats retention index. Positive values in the regression coefficients indicate that the indicated descriptor contributes positively to the value of RI, whereas negative values indicate that the greater the value of the descriptor the lower the value of RI. In other words, increasing the σ_9, ΔG_{CYCLO} and EP_5 will decrease RI and increasing the HF and M increases extent of RI of the AD. The standardized regression coefficient reveals the significance of an individual descriptor presented in the regression model.

CONCLUSION

In this article, a QSAR study of 32 anti-cancer drugs was performed based on the theoretical molecular descriptors calculated by the GAUSSIAN software and selected. The built model was assessed comprehensively (internal and external validation) and all the validations indicated that the QSPR model built was robust and satisfactory, and that the selected descriptors could account for the structural features responsible for the AD properties of the compounds. The QSPR model developed in this study can provide a useful tool to predict the RI of new compounds and also to design new compounds with high RI.

REFERENCES

[1]. Janssens S, Beyaert R. *Mol Cell* **2003**;11:293–302.
[2]. Li S, Strelow A, Fontana EJ, Wesche H. *Proc Natl Acad Sci USA* **2002**;99:5567–5572.
[3]. Medvedev AE, Lentschat A, Kuhns DB, Blanco JC, Salkowski C, Zhang S, Arditi M, Gallin JI, Vogel SN. *J Exp Med* **2003**;198:521–531.
[4]. Picard C, Puel A, Bonnet M, Ku CL, Bustamante J, Yang K, Soudais C, Dupuis S, Feinberg J, Fieschi C, Elbim C, Hitchcock R, Lammas D, Davies G, Al- Ghonaium A, Al-Rayes H, Al-Jumaah S, Al-Hajjar S, Al-Mohsen IZ, Frayha HH, Rucker R, Hawn TR, Aderem A, Tufenkeji H, BHaraguchi S, Day NK, Good RA, Gougerot-Pocidalo MA, Cassanova JL. *Science* **2003**;299:2076–2079.
[5]. Li X. *Eur J Immunol* **2008**;38:614–618.

[6]. 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.
[7]. 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.
[8]. 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.
[9]. Sammes PG, Taylor JB. Comprehensive Medicinal Chemistry. Oxford: Pergamon Press, **1990**:766.
[10]. Riahi S, Pourbasheer E, Dinarvand R, Ganjali MR, Norouzi P. *Chem Biol Drug Des* **2008**;74:165–172.
[11]. Riahi S, Pourbasheer E, Ganjali MR, Norouzi P. *J Hazard Mater* **2009**;166:853–859.
[12]. Riahi S, Pourbasheer E, Ganjali MR, Norouzi P. *Chem Biol Drug Des* **2009**;73:558–571.
[13]. Depczynski U, Frost VJ, Molt K. *Anal Chim Acta* **2000**;420:217.
[14]. Alsberg BK, Marchand-Geneste N, King RD. *Chemometr Intel Lab* **2000**;54:75–91.
[15]. Jouanrimbaud D, Massart DL, Leardi R, Denoord OE. *Anal Chem* **1995**;67:4295–4301.
[16]. Riahi S, Ganjali MR, E Pourbasheer, Divsar F, Norouzi P, Chaloosi M. *Curr Pharm Anal* **2008**;4:231–237.
[17]. Riahi S, Ganjali MR, Pourbasheer E, Norouzi P. *Chromatographia* **2008**;67:917–922.
[18]. Riahi S, Pourbasheer E, Ganjali MR, Norouzi P, Zeraatkar Moghaddam A. *J Chin Chem Soc* **2008**;55:1086–1093.
[19]. Riahi S, Ganjali MR, Moghaddam AB, Pourbasheer E, Norouzi P. *Curr Anal Chem* **2009**;5:42–47.
[20]. Tropsha A, Gramatica P, Gombar VK. *QSAR Comb Sci* **2003**;22:69–77.
[21]. Riahi S, Ganjali MR, Norouzi P, Jafari F. *Sens. Actuators* B **2008**;132:13–19.
[22]. Jiri Burkhard,Jiri Vais,Ludek Vodicka And Stanislav Landa. *Jornal of Chromatography*. *Chrom*.4057.
[23]. Holland H. Adaption in Natural and Artificial Systems. Ann Arbor, MI: The University of Michigan,**1975**;342–375.
[24].Cartwright HM. Applications of Artificial Intelligence in Chemistry. Oxford: Oxford University, **1993**;760–765.
[25] Hunger J, Huttner G., *J Comput Chem* **1999**;20:455–471.
[26]. Ahmad S, Gromiha MM. *J Comput Chem* **2003**;24:1313–1320.
[27]. Waller CL, Bradley MP. *J Chem Inf Comput Sci* **1999**;39:345–355.
[28]. Aires-de-Sousa J, Hemmer MC, Casteiger J. *Anal Chem* **2002**;74:80–90.
[29]. The Mathworks. Genetic Algorithm and Direct Search Toolbox Users Guide. Massachusetts: MathWorks, **2002**;50–65.