

Application of different chemometric tools in QSAR study of azolo-adamantanes against influenza A virus

R. Karbakhsh^{1,*} and R. Sabet²

¹Department of Chemistry, Faculty of Advanced Sciences and Technology, Pharmaceutical Sciences Branch, Islamic Azad University, Tehran, I.R.Iran.

²Department of Medicinal Chemistry and Isfahan Pharmaceutical Science Research Center, School of Pharmacy and Pharmaceutical Science, Isfahan University of Medical Science, Isfahan, I.R.Iran.

Abstract

Quantitative relationships between molecular structure and azolo-adamantanes derivatives were discovered by different chemometric tools including factor analysis based multiple linear regressions (FA-MLR), principle component regression analysis (PCRA), and genetic algorithm-partial least squares GA-PLS. The FA-MLR describes the effect of geometrical and quantum indices on enzyme inhibition activity of the studied molecules. The quality of PCRA equation was found to be better than those derived from FA-MLR. GA-PLS analysis indicated that the topological (IC4 and MPC06), constitutional (nf) and geometrical (G (N..S)) parameters were the most significant ones on influenza A virus activity. Comparison of the different statistical methods employed revealed that GA-PLS represented superior results and it could explain and predict 85% and 77% of variances in the pIC₅₀ data, respectively.

Keywords: Influenza A, Azolo-adamantanes, QSAR, GA-PLS, PCRA, FA-MLR.

INTRODUCTION

Synthesis and evaluation of biological effects of new compounds usually consumes a lot of time and money. Nowadays, the application of computational methods for designing biologically active compounds has opened a new window to modern drug discovery research. Computational methods can accelerate the procedure of discovering new drugs by designing new compounds and predict their potency or activity. Quantitative structure activity relationships (QSAR) studies, as one of the most important areas in chemometrics, play a fundamental role in predicting the biological activity of new compounds and identifying ligand-receptor interactions (1-5). QSAR models are mathematical equations that provide a deeper knowledge into the mechanism of biological activity of compounds by constructing a relationship between chemical structures and biological activities. The most important step in building QSAR models is the appropriate representation of the structural and

physicochemical features of chemical entities (6-9). These features which are defined as molecular descriptors are the ones with higher impact on the biological activity of interest (10-13). Molecular descriptors have been classified into different categories according to different approaches including physicochemical, constitutional, geometrical, topological, and quantum chemical descriptors. Dragon and Gaussian are two well-known computational softwares which can provide more than 1000 of these descriptors (14,15). The first step in constructing the QSAR/QSPR models is the selection of molecular descriptors that represent variation in the interested property of the molecules by a number. The selected descriptors then will be used for constructing statistical models. There are two types of QSAR/QSPR models: regression models and classification models. Multiple linear regression (MLR), principle component regression (PCR), and partial least squares (PLS) are considered as regression models. Although MLR equations can describe the structure property relationships appropriately,

*Corresponding author: Reza Karbakhsh
Tel. 0098 912 2046978, Fax. 0098 21 44471682
Email: chemrk@gmail.com

some information will be disregarded in MLR analysis. Due to the co-linearity problem in MLR analysis, one may remove the collinear descriptors before the development of the MLR model. There are several variable selection methods including forward, backward, and stepwise selection. There are also some other methods inspired by the nature of which genetic algorithm is the most widely used. Factor analysis identifies the important predictor variables contributing to the response variable and avoids collinearities among them. PLS analysis as a factor analysis-based method omits the multicollinearity problem in the descriptors. In this method, the descriptors data matrix is decomposed to orthogonal matrices with an inner relationship between the dependent and independent variables. Because a minimal number of latent variables are used for modeling in PLS, this modeling method coincides with noisy data better than MLR (11-13).

Each winter, millions of people suffer from influenza, a highly contagious infection. The influenza virions are enveloped, mostly as spherical particles containing an outer lipid membrane. The genome of influenza virus is represented by eight separate segments of single-strand negative RNA associated with nucleoprotein and several molecules of the three subunits of its RNA polymerase. Unlike eukaryotic RNA polymerase, viral polymerase complex lacks error-prone activity. For this reason, similar to other RNA viruses, influenza virus has a very high rate of mutations in its genome leading to the fast selection of drug-resistant strains. Despite numerous steps in the viral life cycle that are potential targets for drug intervention, only two of them are now available for clinical usage. Currently, two main classes of chemical compounds are used for the treatment of influenza. They differ in their viral targets and mechanisms of action. The antiviral drugs amantadine and rimantadine block a viral ion channel and prevent the virus from infecting cells. Oseltamivir and zanamivir are designed to halt the spread of the virus in the body (16).

The structural invariants obtained from whole molecular structures and three different chemometric methods were used to make connections between structural parameters and

azolo-adamantanes. These methods included partial least squares combined with genetic algorithm for variable selection (GA-PLS), factor analysis-MLR (FA-MLR) and principle component regression analysis (PCRA).

MATERIALS AND METHODS

Software

A Pentium IV personal computer (CPU at 3.06 GHz) with windows XP operating system was used. Geometry optimization was performed by Hyperchem (version 7.0 Hypercube, Inc.) Dragon software was used for calculation of constitutional, topological, geometrical, and functional group descriptors. Gaussian software was used for calculation of quantum descriptors. SPSS software (version 11.50, IBM, Inc.) was used for PCR and FA-MLR analysis. GA-PLS regression and other calculations were performed in the MATLAB (version 7.1, MathWorks, Inc.) environment.

Activity data and descriptor generation

The biological data used in this study were anti influenza A activity, (in terms of $-\log IC_{50}$), of a set of forty six azolo-adamantanes derivatives (16). The structural features and biological activity of these compounds are listed in Table 1 and then used for subsequent QSAR analysis as dependent variable. The two-dimensional structures of molecules were drawn using Hyperchem 7.0 software. The final geometries were obtained with the semi-empirical AM1 method in Hyperchem program. The molecular structures were optimized using Polak-Ribiere algorithm until the root mean square gradient was $0.01 \text{ kcal mol}^{-1}$. Some chemical parameters including molecular volume (V), molecular surface area (SA), hydrophobicity (Log P), hydration energy (HE) and molecular polarizability (MP) were calculated using the Hyperchem Software. The resulted geometry by the Hyperchem software was transferred into Dragon program, which was developed by Milano Chemometrics and QSAR Group (14). Different functional groups, topological, geometrical and constitutional descriptors for each molecule were calculated by Dragon software. Z-matrices of the structures were provided by the Hyperchem software and

Table 1. Chemical structures of azolo-adamantanes analogues used in this study and their experimental activity against influenza A virus.

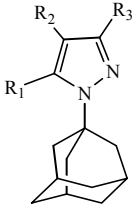
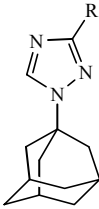
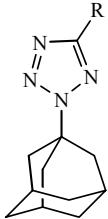
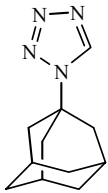
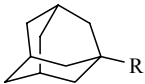
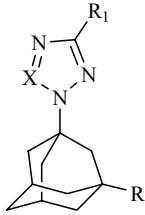
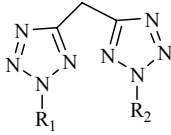
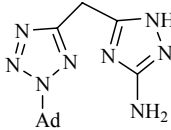
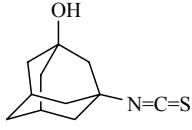
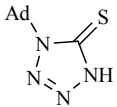
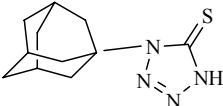
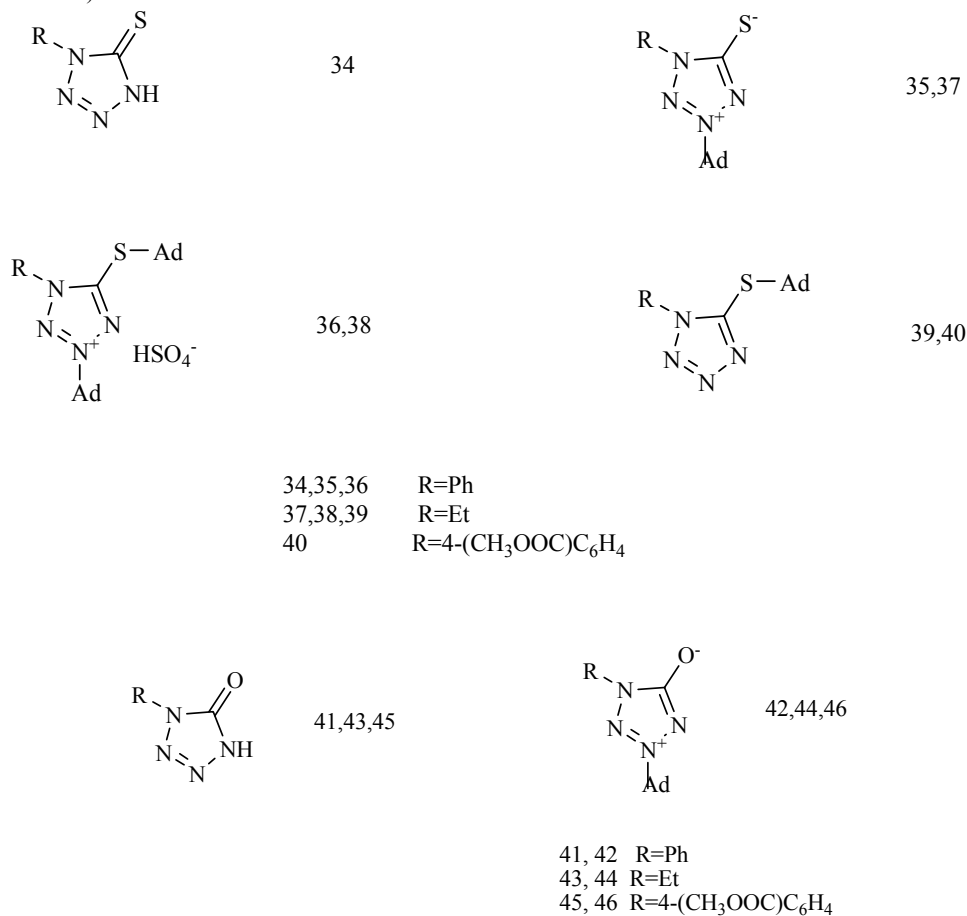
	1: R ₁ =R ₂ =CH ₃ R ₃ =NH ₂ 2: R ₁ =H R ₂ =Br R ₃ =COOH		3=H 4=Cl
	5: R=H 6: R=CH ₃ 7: R=CH ₂ COOH 8: R=CH ₂ COOC ₂ H ₅ 9: R=CH ₂ COO(CH ₂) ₃ CH ₃ 10: R=CH ₂ COONHNH ₂ 11: R=CH ₂ COONHCH ₂ CH ₂ OH		12
	13: R=NH ₂ 14: R=CH(CH ₃)NH ₂		
	X=C, R=NH ₂ X=C, R=CH(CH ₃)NH ₂ X=N, R=NH ₂ X=N, R=CH(CH ₃)NH ₂	15: R ₁ =H 16: R ₁ =Br 17: R ₁ =CH ₃ 18: R ₁ =NO ₂ 19: R ₁ =Cl 20: R ₁ =H 21: R ₁ =CH ₃ 22: R ₁ =C ₂ H ₅ 23: R ₁ =CH ₂ COOC ₂ H ₅ 24: R ₁ =H 25: R ₁ =CH ₃ 26: R ₁ =C ₂ H ₅ 27: R ₁ =CH ₂ COOC ₂ H ₅	
	28: R ₁ =H R ₂ =Ad 29: R ₁ =R ₂ =Ad		
	30		32
	31		33

Table 1. (Continued)

No.	Experimental pIC_{50}	Predicted pIC_{50}			No.	Experimental pIC_{50}	Predicted pIC_{50}		
		FA-MLR	PCR	GA-PLS			FA-MLR	PCR	GA-PLS
*1	7.42	7.57	7.54	7.37	24	7.36	7.21	7.14	7.29
2	7.23	7.58	7.43	7.38	25	7.17	7.19	7.16	7.23
*3	7.23	7.35	7.32	7.33	26	7.96	7.81	7.81	7.78
4	7.38	7.23	7.32	7.28	27	7.80	7.86	7.79	7.81
5	7.49	7.38	7.42	7.31	28	7.55	7.51	7.59	7.63
6	7.14	7.28	7.06	7.48	29	7.72	7.80	7.78	7.69
7	7.09	7.25	7.13	7.25	30	7.80	7.84	7.72	7.86
8	6.96	7.35	7.06	7.16	*31	7.82	7.83	7.70	7.88
9	7.26	7.29	7.16	7.18	32	7.74	7.66	7.89	7.70
10	7.38	7.35	7.27	7.37	33	7.77	7.76	7.88	7.81
*11	7.29	7.44	7.35	7.25	34	7.72	7.76	7.77	7.93
12	7.34	7.50	7.52	7.38	35	8.05	7.79	7.85	7.95
13	7.37	7.20	7.39	7.21	36	7.82	7.82	7.75	8.00
14	7.43	7.41	7.41	7.41	37	7.92	7.80	7.85	7.94
15	7.52	7.35	7.48	7.53	38	7.57	7.80	7.76	7.75
*16	7.28	7.28	7.47	7.34	*39	7.60	7.72	7.74	7.67
17	7.26	7.28	7.20	7.24	40	7.70	7.51	7.64	7.63
18	7.21	7.17	7.21	7.31	41	8.05	7.99	7.97	7.92
19	7.27	7.23	7.35	7.28	42	7.77	7.81	7.72	7.87
20	7.51	7.55	7.53	7.41	43	7.72	7.79	7.76	7.83
21	7.43	7.68	7.38	7.70	44	7.35	7.46	7.60	7.51
22	7.30	7.55	7.40	7.42	*45	8.00	7.77	7.93	7.90
*23	7.60	7.42	7.44	7.48	46	7.26	7.50	7.31	7.27

* Compounds used as prediction set

transferred to Gaussian 98 program. Complete geometry optimization was performed taking the most extended conformation as starting geometries. Semi-empirical molecular orbital calculation (AM1) of the structures was performed using the Gaussian 98 program (15). The Gaussian program calculated different quantum chemical descriptors including, dipole moment (DM), local charges, and HOMO and LOMO energies. Hardness (η), softness (S), electronegativity (χ) and electrophilicity (ω) were calculated according to the method proposed by Thanikaivelan and coworkers (17). The calculated descriptors from whole molecular structures are briefly described in Table 2.

Data Pretreatment and model building

Anti influenza A activity was used as dependent variable. The calculated descriptors (independent variables) were collected in a data matrix whose number of rows and columns were the number of molecules and descriptors, respectively. In order to test the final model performances, about 18% of the data (8 molecules out of 46) were selected as external test set molecules. These samples

were selected based on descriptors spaces. The data matrix containing the total descriptors was subjected to principle component analysis and the first two principle components were plotted against each other. GA-PLS, MLR with factor analysis as the data pre-processing step for variable selection and PCRA methods were used to derive the QSAR equations.

RESULTS

GA-PLS

In this study, GA-PLS was employed to model the structure azolo-adamantanes activity relationships more appropriately.(18-19). Application of PLS method thus allows the construction of larger QSAR equations while still avoiding over-fitting and eliminating most variables. This method is normally used in combination with cross-validation to obtain the optimum number of components (20-21). The PLS regression method used was the NIPALS-based algorithm existed in the chemometric toolbox of MATLAB software (version 7.1 MathWorks, Inc.). In order to obtain the optimum number of factors based

Table 2. Brief description of some descriptors used in this study.

Descriptor type	Molecular Description
Constitutional	Molecular weight, no. of atoms, no. of non-H atoms, no. of bonds, no. of heteroatoms, no. of multiple bonds (nBM), no. of aromatic bonds, no. of functional groups (hydroxyl, amine, aldehyde, carbonyl, nitro, nitroso, etc.), no. of rings, no. of circuits, no of H-bond donors, no of H-bond acceptors, no. of Nitrogen atoms (nN), chemical composition, sum of Kier-Hall electrotopological states (Ss), mean atomic polarizability (Mp), number of rotatable bonds (RBN), mean atomic Sanderson electronegativity (Me), etc.
Topological	Molecular size index, molecular connectivity indices (X1A, X4A, X2v, X1Av, X2Av, X3Av, X4Av), information content index (IC), Kier Shape indices, total walk count, path/walk-Randic shape indices (PW3, PW4, Zagreb indices, Schultz indices, Balaban J index (such as MSD) Wiener indices, topological charge indices, Sum of topological distances between F..F (T(F..F)), Ratio of multiple path count to path counts (PCR), Mean information content vertex degree magnitude (IVDM), Eigenvalue sum of Z weighted distance matrix (SEigZ), reciprocal hyper-detour index (Rww), Eigenvalue coefficient sum from adjacency matrix (VEA1), radial centric information index, 2D petijean shape index (PJI2), etc.
Geometrical	3D petijean shape index (PJI3), Gravitational index, Balaban index, Wiener index, etc.
Quantum	Highest occupied Molecular Orbital Energy (HOMO) , Lowest Unoccupied Molecular Orbital Energy (LUMO), Most positive charge (MPC), Least negative charge (LNC), Sum of squares of charges (SSC), Sum of square of positive charges (SSPC), Sum of square of negative charges (SSNC), Sum of positive charges (SUMPC), Sum of negative charges (SUMNC), Sum of absolute of charges (SAC), Total dipole moment (DM _t), Molecular dipole moment at X-direction (DM _x), Molecular dipole moment at Y-direction (DM _y), Molecular dipole moment at Z-direction (DM _z).
Functional group	Number of total tertiary carbons (nCt), Number of H-bond acceptor atoms (nHAcc), number of total hydroxyl groups (nOH), number of unsubstituted aromatic C(nCaH), number of ethers (aromatic) (nRORPh), etc.
Chemical	LogP (Octanol-water partition coefficient), Hydration Energy (HE), Polarizability (Pol), Molar refractivity (MR), Molecular volume (V), Molecular surface area (SA).

on the Haaland and Thomas F-ratio criterion, leave-one-out cross-validation procedure was used (22).

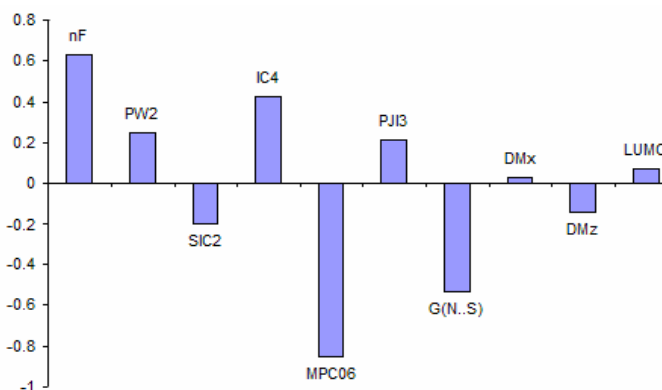
Genetic algorithm is a novel and simple optimization method based on the evolution process of living beings in which simplicity and effectiveness have been applied to the various types of optimization problems in many scientific fields. It uses genetic rules such as reproduction, crossover and mutation to build pseudo organisms that are then selected, on the basis of a fitness criterion to survive and pass information on to the next generation (23-25). Each individual of the population was defined by a chromosome of binary values representing a subset of descriptors. The population size varied between 50 and 250 for different GA runs. The population of the first generation was selected randomly. The number of genes at each chromosome was equal to the number of descriptors (26). A gene took a value of 1 if its corresponding descriptor was included in the subset; otherwise, it took a value of 0. The number of genes with a value of 1 was kept relatively low to have a small subset of descriptors, that is, the probability of generating 0 for a gene was set greater (at least 70%) than the value of 1 (25). The operators used here were crossover and mutation. The probability of the application of these operators varied linearly with generation renewal (0-10% for mutation and 60-90% for crossover). For a typical run, the evolution of the generation was stopped when 90% of the generation took the same fitness. A maximum generation number of 500 were used throughout. The fitness function (predictability of the model) was computed by cross-validation procedure based on the sum of squares of errors (SSECV) value. The inverse of SSECV was considered as fitness function (27). The chromosomes with the least numbers of selected descriptors and the highest fitness were marked as informative chromosomes (26).

In PLS analysis, the descriptors data matrix is decomposed to orthogonal matrices with an inner relationship between the dependent and independent variables. The multi-collinearity problem in the descriptors is omitted by PLS

analysis because a minimal number of latent variables are used for modeling in PLS (26). Since redundant variables degrade the performance of PLS analysis, similar to other regression methods, a variable selection method must be employed to find the more convenient set of descriptors. Here, GA was used as variable selection method. These samples were selected based on descriptors spaces. To do so, the data matrix containing the total descriptors was subjected to principle component analysis and the first two principle components were plotted against each other. The data set (n=46) was divided into two groups: calibration set (n=38) and prediction set (n=8). Given 38 calibration samples; cross-validation procedure was used to find the optimum number of latent variables for each PLS model. GA produces a population of acceptable models in each run. In this work, many different GA-PLS runs were conducted using different initial set of populations (50-250) and therefore a large number of acceptable models were created.

The most convenient GA-PLS model that resulted in the best fitness contained 10 descriptors including four topological indices (PW2, SIC2, IC4 and MPC06), one constitutional (nf), two geometrical (G (N..S) and PJI3) and three quantum parameters (LUMO, DMz, DMx). The PLS estimate of the regression coefficients are shown in Fig. 1. Since these constants were calculated based on the normalized descriptor values, they can be used as a measure of the importance of the corresponding descriptor. As it is observed, the topological (IC4 and MPC06), constitutional (nf) and geometrical (G (N..S)) parameters represent the most significant contribution in the obtained QSAR model followed by the functional geometrical and topological parameters (PJI and SIC2).

The statistical parameters of the resulted PLS-based QSAR model are given in Table 3. This GA-PLS model possessed high statistical quality $R^2=0.86$ and $Q^2=0.77$. It could explain and predict about 77% of variances in the anti influenza A activity of the studied molecules. The predictive ability of the model was measured by application to 8 external test set molecules. The correlation coefficient of

Fig. 1. PLS regression coefficients for the variables used in GA-PLS model**Table 3.** Statistical parameters for testing prediction ability of the FA-MLR, PCR and GA-PLS models

Model	q^2 ^a	$RMSE_{CV}$ ^b	r_p^2 ^c	$RMSE_p$ ^d
FA-MLR	0.64	0.17	0.78	0.19
PCR	0.82	0.11	0.82	0.12
GA-PLS	0.77	0.13	0.85	0.14

^a q^2 = Cross validation correlation coefficient. ^b $RMSE_{CV}$ = Root mean square error of cross validation.

^c r_p^2 = Regression coefficient for prediction set. ^d $RMSE_p$ = Root mean square error of prediction set.

prediction set is 0.85, which means that the resulted QSAR model could predict 85% of variances in the inhibitory activity data and standard error of prediction was 0.13.

The predicted activities are represented in Table 1 and are plotted against the corresponding experimental values in Fig. 2 Comparison between the results obtained by GA-PLS and the other employed regression methods indicates higher accuracy of this method in describing anti influenza A activity of the azolo-adamantanes derivative. Difference in accuracy of the different regression methods used in this study is visualized in Fig. 2 by plotting the predicted activity (by cross-validation) against the experimental values. As it is observed, the plot of data resulted by GA-PLS represents the lowest scattering of data around a straight line and that obtained by PCRA analysis is in the second order of accuracy.

Some criteria for the prediction of the model are suggested by Tropsha. If these criteria are satisfied, it can then be concluded that the model is predictive:

$$\begin{array}{l} R_{LOO}^2 > 0.5 \quad R^2 > 0.6 \\ \frac{R^2 - R_o^2}{R^2} < 0.1 \quad \frac{R^2 - R_o'^2}{R^2} < 0.1 \\ 0.85 < k < 1.15 \quad \text{or} \quad 0.85 < k' < 1.15 \end{array}$$

where, R^2 is the correlation coefficient of regression between the predicted and observed activities of compounds in training and test set. R_o^2 is the correlation coefficient for regressions between predicted versus observed activities through the origin, $R_o'^2$ is the correlation coefficient for the regressions between observed versus predicted activities through the origin, and the slopes of the regression lines through the origin are assigned by k and k' , respectively. Details of the definitions of parameters such as R_o^2 , $R_o'^2$, k and k' are presented in the literature. In addition, according to Roy and coworkers, it is necessary to study the differences between the values of R_o^2 and $R_o'^2$. They suggest the following modified R^2 form: if R_m^2 value for the given model is >0.5 , indicates good external predictability of the developed model (28).

$$R_m^2 = R^2 \left(1 - \sqrt{R^2 - R_o^2} \right)$$

Robustness and applicability domain of the models

Leverage is one of standard methods for this purpose. The numerical value of leverage has certain properties: (a) the value is always greater than zero, (b) the lower the value, the higher is the confidence in the prediction.

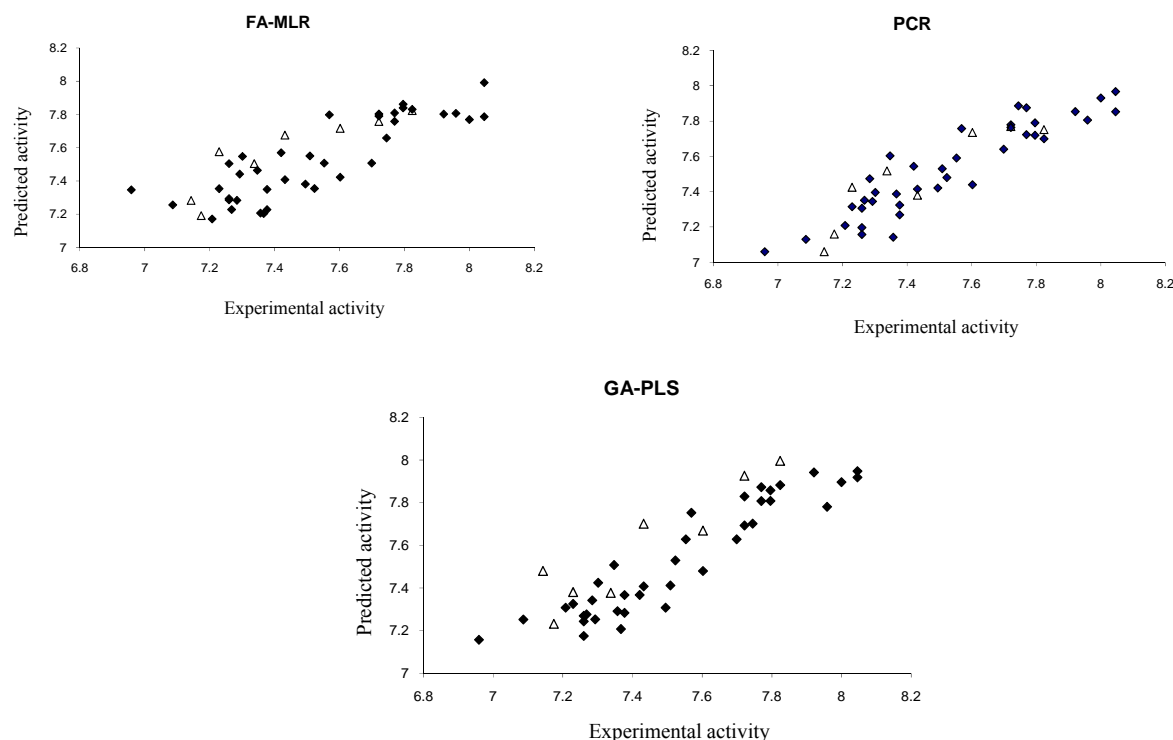


Fig. 2. Plots of the cross-validated predicts activity against the experimental activity for the different models obtained against Influenza A

Table 4. Statistical parameters obtained for the developed model of the investigated compounds

parameter	Training set	Test set
$R^2 - R_v^2 / R^2$	-0.012	-0.031
$R^2 - R'_v{}^2 / R^2$	-0.007	-0.030
K	1.032	0.876
K'	1.014	0.765
R_m^2	0.810	0.793

A value of 1 indicates very poor prediction. A value of 0 indicates perfect prediction and usually is not achievable, (c) If there are P coefficients in the model, the sum of values for leverage at each experimental point of calibration adds up to P . Warning leverage (h^*) is another criterion for interpretation of the results. The warning leverage is, generally, fixed at $3k/n$, where n is the number of training compounds and k is the number of model parameters. A leverage greater than warning leverage h^* means that the predicted response is the result of substantial extrapolation of the model and therefore may not be reliable (29). The calculated leverage values of the test set samples for different MLR and PCR models

are listed in Table 4. The warning leverage, as the threshold value for accepted prediction, is also given in Table 5. As seen, the leverages of all test samples are lower than h^* for all models. This means that all predicted values are acceptable.

FA-MLR and PCRA

FA-MLR was performed on the dataset. Factor analysis (FA) was used to reduce the number of variables and to detect structure in the relationships among them. This data-processing step is applied to identify the important predictor variables and to avoid collinearities (30). PCRA, was tried for the data set along with FA-MLR. With PCRA,

Table 5. Leverage (*h*) of the external test set molecules for different models. The last row (*h*^{*}) is the warning leverage.

Molecule No.	FA-MLR	PCRA	GA-PLS
4	0.101	0.098	0.0322
8	0.203	0.030	0.232
10	0.240	0.045	0.123
25	0.032	0.234	0.543
29	0.313	0.123	0.233
32	0.022	0.021	0.098
38	0.032	0.322	0.032
40	0.043	0.123	0.126
<i>h</i> [*]	0.650	0.609	0.534

Table 6. Numerical values of factor loading numbers 1–4 for descriptors after VARIMAX rotation

	1	2	3	4	Commonality
nF	0.133	0.882	0.133	0.054	0.800
PW2	0.026	0.800	0.234	0.070	0.700
SIC2	0.596	0.341	-0.365	0.184	0.639
IC4	0.074	-0.479	0.605	0.284	0.681
MPC06	0.298	0.477	0.727	0.136	0.864
PJI3	-0.032	-0.179	-0.758	0.371	0.745
G(N..S)	0.882	-0.076	0.129	-0.232	0.853
DMx	-0.055	-0.298	0.040	-0.678	0.553
DMz	-0.031	-0.133	-0.025	0.850	0.741
LUMO	-0.671	-0.270	-0.016	-0.323	0.628
PIC₅₀	-0.790	0.055	-0.470	0.056	0.850
%variance	21.139	20.097	17.249	14.742	73.226

collinearities among X variables are not a disturbing factor and the number of variables included in the analysis may exceed the number of observations (31). In this method, factor scores, as obtained from FA, are used as the predictor variables (30). In PCRA, all descriptors are assumed to be important while the aim of factor analysis is to identify relevant descriptors. Table 6 shows the 4 factor loadings of the variables (after VARIMAX rotation) for the compounds tested against influenza A. As it is observed, about 73% of variances in the original data matrix could be explained by the selected 4 factors.

Based on the procedure explained in the experimental section, the following three-parametric equation was derived:

$$pIC_{50} = 6.590 (\pm 0.353) - 0.054 (\pm 0.007) G(N..S) + 1.742 (\pm 0.391) PJI3 - 0.050 (\pm 0.021) DMz$$

$$r^2 = 0.72 \quad S.E = 0.21 \quad F = 29.80 \quad q^2 = 0.64$$

$$RMS_{cv} = 0.17 \quad N = 38 \quad (E_1)$$

Equation 1 could explain about 72% of the variance and predict 64% of the variance in pIC₅₀ data. This equation describes the effect of geometrical (G (N..S) and PJI3) and Quantum (DMz) indices on enzyme inhibitory activity of the studied molecules.

When factor scores were used as the predictor parameters in a multiple regression equation using forward selection method (PCRA), the following equation was obtained:

$$pIC_{50} = 7.520 (\pm 0.018) - 0.215 (\pm 0.018) f_1 - 0.144 (\pm 0.019) f_3$$

$$r^2 = 0.85 \quad S.E. = 0.13 \quad F = 97.82 \quad q^2 = 0.82$$

$$RMS_{cv} = 0.11 \quad N = 38 \quad (E_2)$$

Equation 2 could explain and predict 85% and 82% of the variances in pIC_{50} data, respectively. Since factor scores are used instead of selected descriptors, and any factor-score contains information from different descriptors, loss of information is thus avoided and the quality of PCRA equation is better than those derived from FA-MLR (32).

As seen in Table 6, in the case of each factor, the loading values for some descriptors are much higher than those of the others. These high values for each factor indicate that this factor contains more information about descriptors. It should be noted that all factors have information from all descriptors but the contribution of descriptor in different factors are not equal. For example, factors 1 and 2 have higher loadings for the geometrical, topological and constitutional indices, whereas information about the topological, geometrical and quantum descriptors are highly incorporated in factor 3 and 4. Therefore, from the factor scores used by equation E_2 , significance of the original variables for modeling the activity can be obtained. Factor score 1 indicates importance of G (N..S) (Geometrical indices). Factor score 2 indicates importance of nf and PW2 (the constitutional and topological descriptors) and factor scores 3 and 4 signify the importance of MPC06, PJI3 and DMz (the topological, geometrical and Quantum descriptors).

The predicted values of the activity for calibration set (by cross-validation) and prediction set for FA-MLR and PCRA are listed in Table 1 and are plotted against the corresponding experimental values in Fig. 2. The statistical parameters of prediction set are listed in Table 3. The correlation coefficient of prediction for FA-MLR analysis is 0.78, which means that the obtained QSAR model could predict 78% of variances in the anti influenza A activity data. It has a root mean square error of 0.19. The correlation coefficient of prediction for PCRA analysis is 0.82. This means that the derived QSAR model could predict 82% of variances in the inhibitory activity data. The root mean square error of PCRA analysis was 0.12. Whilst the data of this analysis shows acceptable prediction, we see that the predicted values of some molecules are near to each other.

DISCUSSION

Quantitative relationships between molecular structure and anti influenza activity were discovered by GA-PLS, FA-MLR and PCRA. As it was shown in Fig. 1 the topological (IC4 and MPC06), constitutional (nf) and geometrical (G (N..S)) parameters represent the most significant contribution in the obtained QSAR model followed by the functional geometrical and topological parameters (PJI and SIC2). FA-MLR was performed on the dataset. Equation 1 describes the effect of geometrical (G (N..S) and PJI3) and Quantum (DMz) indices on enzyme inhibitory activity of the examined molecules. PCRA was performed on the dataset and equation 2 could explain and predict 85% and 82% of the variances in pIC_{50} data.

CONCLUSION

Quantitative relationships between molecular structure and anti influenza A activity of a series of azolo-adamantanes derivatives were discovered by different chemometric tools including FA-MLR, PCRA and GA-PLS. The FA-MLR describes the effect of geometrical and quantum indices on inhibitory activity of the examined molecules. The quality of PCRA equation is better than those derived from FA-MLR. Factor scores 1 and 2 indicate importance of geometrical, constitutional and topological indices. Factor scores 3 and 4 show the importance of geometrical, topological and quantum descriptors. GA-PLS analysis indicated that the topological (IC4 and MPC06), constitutional (nf) and geometrical (G (N..S)) parameters were the most significant parameters on inhibitory activity. A comparison between the different statistical methods employed revealed that GA-PLS represented superior results and it could explain and predict 85% and 77% of variances in the pIC_{50} data, respectively.

REFERENCES

1. Schmidi H. Multivariate prediction for QSAR. *Chemom Intell Lab Syst.* 1997;37:125-134.
2. C. Hansch, A. Kurup, R. Garg, H. Gao. *Chemoinformatics and QSAR: A review of QSAR*

- lacking positive hydrophobic terms. *Chem Rev.* 2001;101:619-672.
3. Wold S, Trygg J, Berglund A, Antti H, Some recent developments in PLS modeling. *Chemom Intell Lab Syst.* 2001;58:131-150.
 4. Hemmateenejad B, Miri R, Akhond M, Shamsipur M. QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. An application of genetic algorithm for variable selection in MLR and PLS methods. *Chemom Intell Lab Syst.* 2002;64:91-99.
 5. Hemmateenejad B, Miri R, Akhond M, Shamsipur M. Quantitative structure-activity relationship study of recently synthesized 1,4-dihydropyridine calcium channel antagonists. Application of the Hansch analysis method. *Arch Pharm.* 2002;335:472-480.
 6. Hansch C, Fujita T. ρ - σ - π Analysis. A method for the correlation of biological activity and chemical structure. *J Am Chem Soc.* 1964;86:1616-1626.
 7. Wang J, Zhang L, Yang G, Zhan CG. Quantitative structure-activity relationship for cyclic imide derivatives of protoporphyrinogen oxidase inhibitors: A study of quantum chemical descriptors from density functional theory. *J Chem Inf Comput Sci.* 2004;44:2099-2105.
 8. Hansch C, Hoekman D, Gao H. Comparative QSAR: Toward a deeper understanding of chemobiological interactions. *Chem Rev.* 1996;96:1045-1075.
 9. Todeschini R, Consonni V. *Handbook of Molecular Descriptors.* Wiley-VCH, Weinheim, 2000. p. 178-180.
 10. Horvath D, Mao B. Neighborhood behavior. Fuzzy molecular descriptors and their influence on the relationship between structural similarity and property similarity. *QSAR Comb Sci.* 2003;22:498-509.
 11. Putta S, Eksterowicz J, Lemmen C, Stanton R. A novel subshape molecular descriptor. *J Chem. Inf Comput Sci.* 2003;43:1623-1635.
 12. Gupta S, Singh M, Madan AK. Superpendent index: A novel topological descriptor for predicting biological activity. *J Chem Inf Comput Sci.* 1999;39:272-277.
 13. Consonni V, Todeschini R, Pavan M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J Chem Inf Comput Sci.* 2002;42:693-705.
 14. Dragon software, version 2.1, Milano Chemometrics and QSPR Group., Milano, Italy, 2002.
 15. Frisch MJ, Trucks MJ, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. *Gaussian 98, Revision A.7,* Gaussian, Inc., Pittsburgh, PA, 1998.
 16. Zarubaev VV, Golod EL, Anfimov PM, Shtro AA, Saraev VV, Gavrilov AS, et al. Synthesis and antiviral activity of azolo-adamantanes against influenza A virus. *Bioorg & Med Chem.* 2010;18:839-848.
 17. Thanikaivelan P, Subramanian V, Rao JR, Nair BU. Application of quantum chemical descriptor in quantitative structure activity and structure property relationship. *Chem Phys Lett.* 2000;323:59-70.
 18. Leardi R. Application of genetic algorithm-PLS for feature selection in spectral data sets. *J Chemomtr.* 2000;14:643-655.
 19. Leardi R, Gonzalez AL. Genetic algorithm applied to feature selection in PLS regression: how and when to use them. *Chemom Intell Lab Syst.* 1998;41:195-207.
 20. Fassihi A, Sabet R. QSAR Study of p56^{lck} Protein Tyrosine Kinase Inhibitory Activity of Flavonoid Derivatives Using MLR and GA-PLS. *Int J Mol Sci.* 2008;9:1876-1892.
 21. Leardi R. Genetic Algorithms in Chemometrics and Chemistry: A Review. *J Chemometrics.* 2001;15:559-569.
 22. Hemmateenejad B. Optimal QSAR analysis of the carcinogenic activity of drugs by correlation ranking and genetic algorithm-based. *J Chemometrics.* 2004;18:475-485.
 23. Cho SJ, Hermsmeier MA. Genetic algorithm guided selection: Variable selection and subset selection. *J Chem Inf Comput Sci.* 2002;42:927-936.
 24. Ahamed M, Gromiha M. Design and training of a neural network for predicting the solvent accessibility of proteins. *J Comp Chem.* 2003;24:1313-1320.
 25. Hemmateenejad B, Safarpour M.A, Miri R, Taghavi F. Application of ab initio theory for the QSAR study of 1,4-dihydropyridine-based calcium channel antagonist. *J Comput Chem.* 2004;25:1495-1503.
 26. Deeb O, Hemmateenejad B, Jaber A, Garduno-Juarez R, Miri R. Effects of the Electronic and Physicochemical Parameters on the Carcinogenesis Activity of Some Sulfa Drug Using QSAR Analysis Based on Genetic-MLR & Genetic-PLS. *Chemosphere.* 2007;67:2122-2130.
 27. Absalan G, Hemmateenejad B, Soleimani M, Akhond M, Miri R. Quantitative structure-micelization relationship study of Gemini surfactants using genetic-MLR and genetic-PLS. *QSAR Comb Sci.* 2004;23:416-425.
 28. Roy P, Roy K. On some aspects of variable selection for partial least squares regression models. *QSAR Comb Sci.* 2008;27:302-313.
 29. Brereton R, *Chemometrics Data Analysis for the Laboratory and Chemical Plant.* Wiley. 2004. p. 47-54.
 30. Franke R, Gruska A. *Chemometrics Methods in molecular design,* in: H. van Waterbeemd, , *Methods and Principles in Medicinal Chemistry,* 3rd ed. Vol. 2. VCH, Weinheim. 1995. p. 113-119.
 31. Kubinyi H. The quantitative analysis of structure-activity relationships, in: M.E. Wolff, *Burger's Medicinal Chemistry and Drug Discovery.* 5th Ed. Vol. 1. New York: Wiley. 1995. p. 506-509.
 32. Sabet R, Fassihi A. QSAR Study of Antimicrobial 3-Hydroxypyridine-4-one and 3-Hydroxypyran-4-one Derivatives Using Different Chemometric Tools. *Int J Mol Sci.* 2008;9:2407-2423.