

Quantitative Structure-activity Relationships of Imidazole-containing Farnesyltransferase Inhibitors Using Different Chemometric Methods

Ali Shayanfar^{1,2*}, Saeed Ghasemi², Somaieh Soltani², Karim Asadpour-Zeynali³, Robert J. Doerksen⁴ and Abolghasem Jouyban⁵

¹Liver and Gastrointestinal Diseases Research Center, Students' Research Committee, Tabriz University of Medical Sciences, Tabriz, Iran

²Department of Medicinal Chemistry, Faculty of Pharmacy, Tabriz University of Medical Sciences, Tabriz, Iran

³Department of Analytical Chemistry, Faculty of Chemistry, University of Tabriz, Tabriz, Iran

⁴Department of Medicinal Chemistry and Research Institute of Pharmaceutical Sciences, School of Pharmacy, University of Mississippi, University, MS 38677-1848, USA

⁵Drug Applied Research Center and Faculty of Pharmacy, Tabriz University of Medical Sciences, Tabriz, Iran

Abstract: Farnesyltransferase inhibitors (FTIs) are one of the most promising classes of anticancer agents, but though some compounds in this category are in clinical trials there are no marketed drugs in this class yet. Quantitative structure-activity relationship (QSAR) models can be used for predicting the activity of FTI candidates in early stages of drug discovery. In this study 192 imidazole-containing FTIs were obtained from the literature, structures of the molecules were optimized using Hyperchem software, and molecular descriptors were calculated using Dragon software. The most suitable descriptors were selected using genetic algorithms-partial least squares (GA-PLS) and stepwise regression, and indicated that the volume, shape and polarity of the FTIs are important for their activities. 2D-QSAR models were prepared using both linear methods, i.e., multiple linear regression (MLR), and non-linear methods, i.e., artificial neural networks (ANN) and support vector machines (SVM). The proposed QSAR models were validated using internal and external validation methods. The results show that the proposed 2D-QSAR models are valid and that they can be applied to predict the activities of imidazole-containing FTIs. The prediction capability of the 2D-QSAR (linear and non-linear) models is comparable to and somewhat better than that of previous 3D-QSAR models and the non-linear models are more accurate than the linear models.

Keywords: Imidazole-containing farnesyltransferase inhibitors, cancer, QSAR, multiple linear regression, artificial neural network, support vector machine.

INTRODUCTION

Ras proteins play an essential role in regulating and stimulating proteins involved in cell growth. Mutations in the *ras* gene can cause permanent activation of proteins leading to uncontrolled cell growth and division. Mutation of the *ras* gene is found in 30% of cancers. To be activated, Ras proteins should be coupled with a 15-carbon isoprenyl group in a reaction that is catalyzed by protein farnesyltransferase (FT). Therefore farnesyltransferase inhibitors (FTIs) have been studied extensively as candidates for interfering with Ras operation and hence for cancer chemotherapy. Recently, other mechanisms for FTIs to modulate tumor growth have been reported [1-5]. In addition, FTIs can cause lysis of *Plasmodium falciparum* and have been studied as novel antimalarial agents [6,7] and as therapeutic interventions for

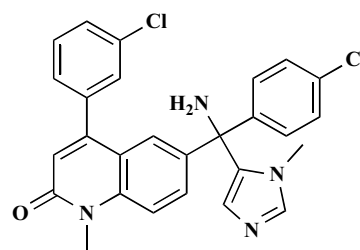


Fig. (1). Structure of Tipifarnib.

other parasitic diseases such as Chagas disease [8]. By inhibition of FT, the signal transduction pathway was stopped and cell proliferation was arrested. Many classes of FTIs have been reported such as non-thiol, non-peptidic, imidazole- or non-imidazole-containing compounds, but among them the imidazole-containing candidates are the most active class of FTIs and some of them have been studied in clinical trials, such as tipifarnib (Fig. 1) which is currently being tested for treatment of acute myeloid leukemia (AML) [4, 5]. Because of its high potency as an FTI, many compounds

*Address correspondence to this author at the Department of Medicinal Chemistry, Faculty of Pharmacy Tabriz University of Medical Sciences, Tabriz, Iran; Tel: 0098 411 3341315; Fax: 0098 411 3344798; E-mails: alishayanf@yahoo.co.uk; shayanfara@tbzmed.ac.ir

were synthesized based on the tipifarnib pharmacophore, but with modifications to its structure, such as those of Abbott Laboratories, based on elimination of ring B and transfer of ring D, which have considerable FT inhibitory activity (see references cited in [9]).

Modeling and prediction of the activities of drugs and drug-like molecules are critical for decreasing the cost and time of new drug discovery as well for understanding the mechanism of drug action. Quantitative structure-activity relationships (QSAR) are mathematical or computational models constructed to find a correlation between the structures and the activities of drugs [10]. The most common kinds of QSAR models are 2D and 3D. 2D-QSAR models correlate the activities of drug-like molecules to structural patterns without consideration of the 3-dimensional (3D) conformations of the molecules, but 3D-QSAR models correlate the activities to properties that manifestly depend on the 3D conformations of the molecules, such as non-covalent interaction fields calculated at points surrounding the molecules [11].

There are some studies predicting the activities of different classes of FTIs in the literature using 2D-QSAR models. González and coworkers investigated the activities of thiol and non-thiol peptidomimetic FTIs using genetic neural network methods [12]. Fernandez *et al.* correlated FTI activities to trihalobenzocycloheptapyridine structures with 2D autocorrelation descriptors using multiple linear regression (MLR) and artificial neural network (ANN) methods [13]. The inhibitory activities of a benzo[*f*]perhydroisoindole (BPHI) series on FT were analyzed using partial least squares and response surface modeling [14]. Bayesian regularized neural networks were applied to predict the activities of FTIs with diverse structures [15]. In addition, Chaurasia and coworkers proposed QSAR models to predict tetrahydroquinoline-based FTI activities using physicochemical properties [16]. Recently, FTI activities of 2,5-diaminobenzophenone derivatives were predicted using 2D chemical drawings [17] and also QSAR of imidazole containing tetrahydrobenzodiazepines as FTIs were studied [18]. Xie *et al.* applied 3D-QSAR models to predict the activities of imidazole-containing FTIs in a recent study [19,9]. They used comparative molecular field analysis (COMFA) and comparative molecular similarity indices analysis (COMSIA) (the most popular approaches of 3D-QSAR) to predict the activities of imidazole-containing FTIs [9]. However, no 2D-QSAR study has been reported for those compounds. 3D-QSAR models have some advantages, but also major disadvantages [11,20]. From a practical aspect, 2D-QSAR models are easy to use and can be easier and faster to develop than 3D-QSAR models [21]. 2D-QSAR models avoid the obstacle of how best to align the 3D-structures of the molecules, which is required in 3D-QSAR. They can also prove to be practical and helpful since the QSARs can be expressed in terms of important and interpretable descriptors.

In this study 192 imidazole-containing FTIs were employed to construct 2D-QSAR models using different chemometric methods. The models were developed using MLR, ANN and support vector machine (SVM) methods to predict the IC_{50} s of the imidazole-containing FTIs. Molecular descriptors were selected by genetic algorithms-partial

least squares (GA-PLS) and stepwise-regression methods and the validities of the developed models were checked by internal and external validation methods. The accuracy of the models was compared with that of the previously reported 3D-QSAR models [9].

MATERIALS AND METHODS

Data Set

The pIC_{50} (negative logarithm of the 50% enzyme inhibitory concentration) values of 192 imidazole-containing FTIs were collected from the literature [9]. This data set is composed of eight different groups of imidazole-containing FTIs. Fig. (2) shows the structures of these compounds. In order to compare the results of this study (2D-QSAR) with the previous study (3D-QSAR), the same carefully-selected training and test sets were used in model development [9].

Molecular Descriptors

In order to calculate molecular descriptors, the structures of the compounds were drawn using Hyperchem 8.0 software and pre-optimized with the molecular mechanics force field (MM+) method and then AM1 semiempirical calculations were performed to optimize the 3D geometries of the molecules with the Polak-Ribière (conjugate gradient) algorithm. The optimized structures from Hyperchem 8.0 software were fed into the Dragon 3.0 software and the molecular descriptors of these compounds were calculated.

Selection of Descriptors

In order to reduce the number of descriptors, the descriptors of 156 compounds in the training set with more than 50% repeated values or collinear descriptors ($R > 0.9$) were excluded and then further reduction of the number of descriptors was performed with GA-PLS. GA simulates the process of natural evolution and has been shown to be an acceptable method to reduce the number of descriptors [22]. Also a combination of PLS, a valuable tool for data reduction, and GA was applied to reduce the number of descriptors [23,24]. GA-PLS was run in MATLAB 7.8 software using the program written by Leardi [25]. The population size of genetic algorithms was considered as 100. 10% of the descriptors with top scores were selected and then the descriptor selection was done using stepwise regression. High correlations with response and low inter-correlation between descriptors (using Pearson correlation) were considered as selection criteria before stepwise regression.

Model Building

MLR Model

The selected descriptors were used to develop a MLR equation using SPSS 11.5 software. Statistical properties of the proposed equation such as the correlation coefficient (R), the adjusted correlation coefficient (R_{adj}), the standard error of estimate (SEE), the probability values (p -value) of each descriptor, and the Fischer statistic or variance ratio (F) recommended by Dearden *et al.* [26] were considered. The proposed model was validated using the leave one-out (LOO) method to evaluate prediction capability of the model.

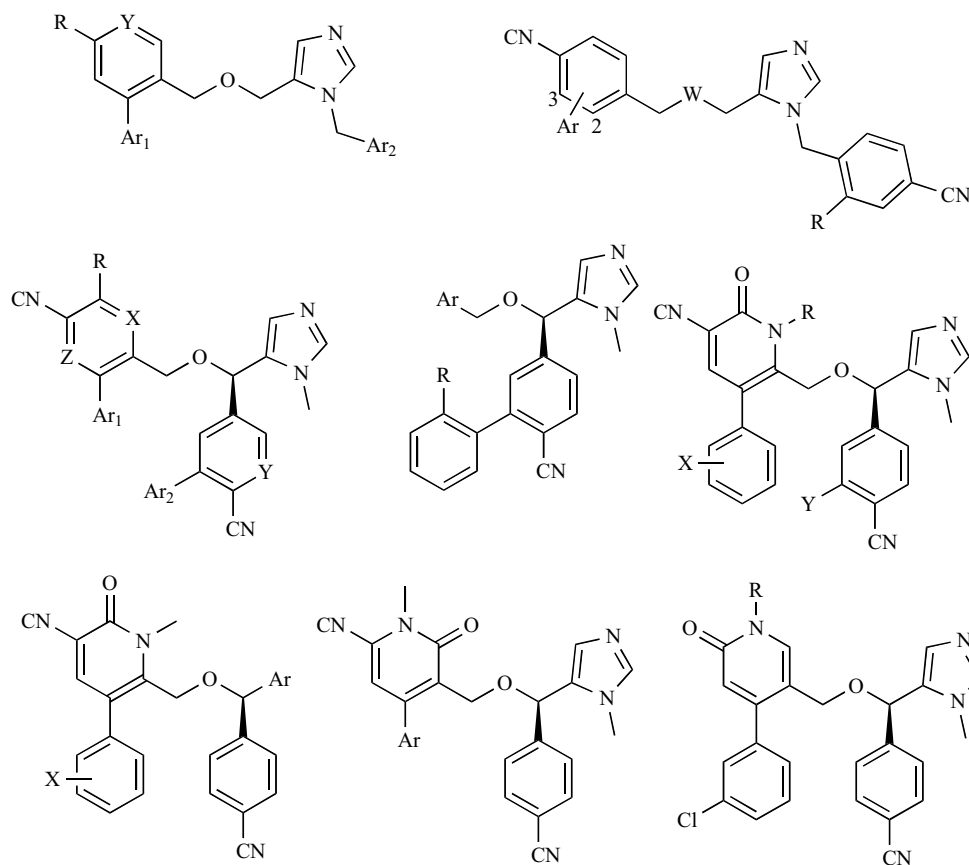


Fig. (2). Structures of the studied FTIs.

ANN with the Levenberg-Marquardt Algorithm

ANNs mimic human brain process information. An ANN has a multilayer structure (input layer, hidden layer and output layer) that consists of neurons and connections between neurons by synapses. Selected descriptors are neurons of the input layer and pIC_{50} values of compounds are the output neurons. Neurons of the hidden layer connect neurons of the input and output layers. This network can form a non-linear relationship between independent (input) and dependent (output) parameters. The strength of the synapse between two neurons is calculated by its weight [27]. There are different algorithms for weight update functions in the literature. Recently, the Levenberg–Marquardt algorithm was characterized as one of the most effective algorithm in QSAR study [27–29]. In this study, we used the nftool (network-fitting tool) toolbox of MATLAB 7.8 software to train the network. This toolbox is user-friendly and uses Levenberg–Marquardt back propagation algorithms (Trainlim) for ANN training. For training a valid network to avoid over-fitting, the 156 data points of the training set described above for MLR were randomly classified into training (70%), validation (15%) and test (15%) sets using the software. A three-layer network with three neurons in the hidden layer was designed.

SVM

SVM is a statistical learning method proposed by Vapnik [30]. This is a relatively new non-linear method in QSAR study. Some QSAR models were proposed using SVM in

recent years [31–33]. This method constructs a hyperplane in a multidimensional space which provides the minimum error by employing a non-linear kernel function for classification or regression tasks. There are some parameters which should be optimized in SVM analysis. One of the parameters is the capacity parameter (C) which is a regularization parameter that adjusts maximizing the distance from the hyperplane to any training set data points and minimizing the error. Another parameter is ϵ which is related to noise in the data. A common type of kernel function is a radial basis function (RBF) [34–37]. This function has a parameter (γ) which should be optimized and controls the generalization ability of the SVM. The C and ϵ parameters were optimized using the leave-many-out cross-validation method. SVM was performed using STATISTICA 7 software.

External Validation of the Proposed Models and Comparison of the Models

In order to check the validity of the proposed models and to compare the prediction capabilities of the models, an external data set (test set) composed of 36 data points is used in this study. A set of statistical criteria proposed in the literature [38] for use with an external test set was considered, including:

$$1. R^2 > 0.6$$

where R^2 is the coefficient of determination (correlation coefficient) between the predicted and observed values of pIC_{50} .

Table 1. Selected Descriptors by GA-PLS and Stepwise Regression from DRAGON Software

Number	Symbol	Definition	Class
1	De	Total accessibility index / weighted by atomic Sanderson electronegativities	WHIM descriptors
2	MATS8e	Moran autocorrelation - lag 8 / weighted by atomic Sanderson electronegativities	2D autocorrelations
3	Mor32m	3D-MorSE - signal 32 / weighted by atomic masses	3D-MorSE descriptors
4	SPH	Spherosity index	Geometrical descriptors
5	BELv1	Lowest eigenvalue n. 1 of Burden matrix / weighted by atomic van der Waals volumes	BCUT descriptors
6	MATS7v	Moran autocorrelation - lag 7 / weighted by atomic van der Waals volumes	2D autocorrelations
7	R1v+	Maximal autocorrelation of lag 1 / weighted by atomic van der Waals volumes	GETAWAY descriptors
8	TIE	E-state topological parameter	Topological descriptors
9	BEHe8	Highest eigenvalue n. 8 of Burden matrix / weighted by atomic Sanderson electronegativities	BCUT descriptors
10	GGI10	Topological charge index of order 10	Galvez topological charge indices
11	GATS6e	Geary autocorrelation - lag 6 / weighted by atomic Sanderson electronegativities	2D autocorrelations

$$2. (R^2 - R_0^2) / R^2 < 0.1 \text{ or } (R^2 - R_0'^2) / R^2 < 0.1$$

where R_0^2 is the correlation coefficient obtained using predicted values relative to a regression line fit to experimental values and required to pass through the origin and $R_0'^2$ is the corresponding correlation coefficient obtained using experimental values relative to a regression line fit to predicted values and required to pass through the origin.

$$3. 0.85 \leq K \leq 1.15 \text{ or } 0.85 \leq K' \leq 1.15.$$

K and K' are the slopes of regression lines through the origin for fits to experimental and predicted data, respectively [38].

Another criteria also was proposed by Roy and Roy [39] to evaluate the external predictability of QSAR models:

$$4. R_m^2 = R^2 \left(1 - \sqrt{R^2 - R_0^2} \right)$$

in which $R_m^2 > 0.5$ indicates the good external predictability of QSAR models.

The accuracy of the proposed models was compared using the average absolute error (AAE), which is defined as:

$$AAE = \frac{\sum |Calculated\ pIC_{50} - Experimental\ pIC_{50}|}{N} = \frac{\sum AE}{N}$$

RESULTS AND DISCUSSION

Selection of Descriptors

Table 1 shows the details of the selected descriptors obtained using GA-PLS and stepwise regression in which less than 1 descriptor per 14 compounds was selected. These results show that a combination of 2D and 3D descriptors was best for predicting the pIC_{50} of the studied structures. Four of

the selected descriptors are 2D-autocorrelation descriptors. These descriptors are used in QSAR studies to calculate the spatial distribution of molecular properties. Fernandez and coworkers used 2D-autocorrelation descriptors to construct QSAR models of trihalobenzocycloheptapyridine containing FTIs [13]. In addition, other 2D descriptors such as topological, BCUT, and Galvez topological charge indices are selected. Also, there are four 3D descriptors including geometrical, 3D-MorSE, WHIM and GETAWAY among the selected descriptors. Based on the selected descriptors, both the volume and polarity of the molecules are important for the activity of the studied compounds. In addition, the shape of the molecule is important.

A correlation matrix shows that there is no inter-correlation ($R < 0.6$) between the selected descriptors (Table 2) and reveals that the selected descriptors are linearly independent and hence can be used together in the development of QSAR models.

Model Building Using Different Methods

The selected descriptors were used to develop QSAR models using MLR, ANN and SVM. A linear model was proposed as the simplest and most straightforward model. Table 3 shows the coefficients, their SEE and the p-value of the selected descriptors. Statistical information that is necessary for validation of QSAR models is listed in Table 4 for the proposed models in this study. The results show that the correlation coefficient is acceptable and there is no significant difference between R and R_{adj} . Internal cross-validation using leave-one-out (LOO) analysis ($q_{loo}=0.727$) shows that the proposed MLR is predictive. Fig. (3) shows the influence of the number of descriptors on R and R_{adj} for the developed model. The increase in R_{adj} after the addition of each descriptor confirms the influence of all the selected descriptors [26].

Table 2. Correlation Matrix Between Selected Descriptors

	De	MATS8e	Mor32m	SPH	BELv1	MATS7v	R1v+	TIE	BEHe8	GGI10	GATS6e
De	1.00										
MATS8e	0.24	1.00									
Mor32m	0.25	0.00	1.00								
SPH	0.23	0.15	0.07	1.00							
BELv1	0.29	0.40	0.08	0.08	1.00						
MATS7v	0.43	0.41	0.14	0.10	0.39	1.00					
R1v+	0.47	0.21	0.04	0.16	0.53	0.22	1.00				
TIE	0.03	0.41	0.13	0.02	0.05	0.20	0.26	1.00			
BEHe8	0.09	0.04	0.28	0.19	0.41	0.19	0.56	0.49	1.00		
GGI10	0.19	0.32	0.33	0.07	0.04	0.27	0.28	0.62	0.56	1.00	
GATS6e	0.21	0.08	0.00	0.09	0.10	0.20	0.34	0.05	0.04	0.07	1.00

Table 3. Coefficients and Statistical Properties of Selected Descriptors of the Most Accurate MLR Model

Descriptors	Coefficient	SEE	p-value
Constant	-32.1519	7.6633	<0.001
De	-3.6908	1.3815	0.008
MATS8e	1.8430	0.8032	0.023
Mor32m	-0.8094	0.2544	0.002
SPH	-1.0655	0.3915	0.007
BELv1	23.6167	3.8210	<0.001
R1v+	17.3016	5.3155	0.001
MATS7v	2.9490	0.9424	0.002
TIE	0.0022	0.0006	0.001
BEHe8	-2.5360	0.7510	0.001
GGI10	1.2870	0.5666	0.025
GATS6e	0.6081	0.2748	0.028

Table 4. Statistical Information for the Proposed Models for the Training Set

MLR					
	N	R	R _{adj}	SEE ^a	F ^a
Training set	156	0.775	0.756	0.478 (0.471)	232.0 (19.7)
ANN					
Training set	110	0.823	0.822	0.434	227.4
Validation set	23	0.801	0.791	0.457	37.6
Test set	23	0.768	0.756	0.440	30.2
Overall	156	0.807	0.806	0.440	288.3
SVM					
Training set	156	0.800	0.798	0.447	273.4

^aThe values are calculated according to correlation between experimental and prediction. The values in parentheses were computed using a MLR model with 11 descriptors. (F and SEE are dependent on the number of independent descriptors and the number of degrees of freedom [26].)

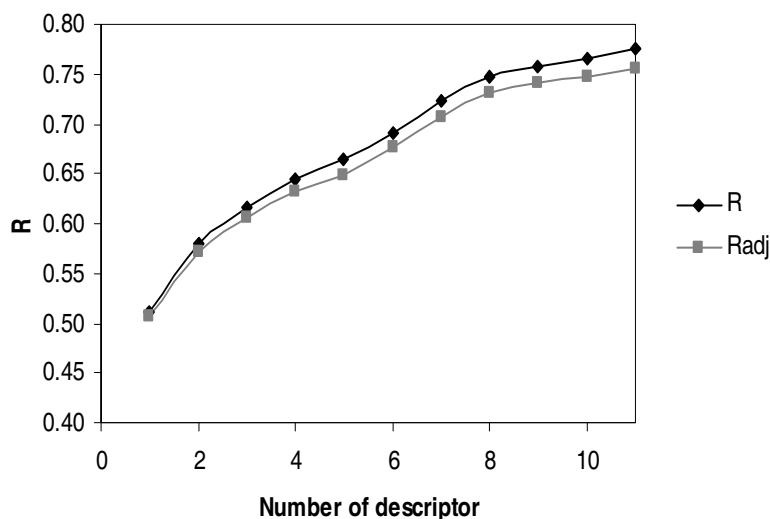


Fig. (3). Effects of the number of descriptors (according to Table 1) on R and R_{adj} to evaluate the influence of the selected descriptors.

In the next stage, data was used to develop an ANN model in which the number of optimal hidden neurons is three. The statistical parameters of the developed ANN model for the data set which was divided into training, validation and test sets are shown in Table 4. There are no significant changes between statistical properties of these sets. In addition, external validation was performed (see Materials and Methods section) and the results show that the trained network is valid and no over-fitting occurred.

Finally, SVM models were developed using the selected descriptors. The optimization of the SVM parameters (C , ϵ and γ) was done with 10-fold cross-validation using the STATISTICA 7 software. A robust model is developed by selecting parameters that give the lowest error. The optimized values of C , ϵ and γ were 7, 0.001 and 0.110 and the statistical properties of the proposed SVM model for the training set are listed in Table 4. Predicted and absolute error (AE) values using MLR, ANN and SVM models are listed in Table 5.

Table 5. Experimental, Predicted and Absolute Error (AE) Values of 156 Training and 36 Test Set Compounds

No.	pIC ₅₀ exp	MLR	AE	ANN	AE	SVM	AE
		pIC ₅₀ pred		pIC ₅₀ pred		pIC ₅₀ pred	
			Training set				
1	9.21	8.61	0.597	8.75	0.463	8.69	0.520
2	9.43	8.96	0.467	9.13	0.304	8.98	0.445
3	7.02	8.48	1.463	8.34	1.321	8.59	1.570
4	8.66	8.66	0.002	8.64	0.023	8.64	0.019
5	8.92	8.89	0.031	8.92	0.003	8.92	0.003
6	9.31	8.59	0.716	8.82	0.491	8.79	0.518
7	9.21	8.72	0.486	8.58	0.631	8.62	0.586
8	7.89	8.48	0.593	8.65	0.756	8.45	0.555
9	8.89	9.06	0.168	9.07	0.180	8.92	0.032
10	9.36	9.14	0.216	9.28	0.077	9.07	0.295
11	8.12	8.37	0.254	8.39	0.270	8.38	0.264
12	9.43	8.80	0.633	9.13	0.304	8.92	0.512
13	9.22	8.95	0.272	8.65	0.569	8.88	0.336
14	8.08	8.44	0.355	8.56	0.482	8.33	0.249
15	9.09	8.21	0.877	8.21	0.880	8.42	0.665
16	9.04	9.35	0.311	9.16	0.120	9.21	0.165

Table 5. contd....

No.	pIC ₅₀ exp	MLR	AE	ANN	AE	SVM	AE
		pIC ₅₀ pred		pIC ₅₀ pred		pIC ₅₀ pred	
17	8.89	9.29	0.402	9.09	0.202	9.04	0.154
18	9.34	8.38	0.956	8.72	0.619	8.47	0.874
19	9.57	8.80	0.770	8.92	0.645	8.97	0.600
20	10.00	8.71	1.293	8.87	1.131	8.92	1.081
21	8.37	8.44	0.066	8.68	0.314	8.37	0.004
22	8.77	7.91	0.865	8.03	0.738	7.96	0.806
23	8.52	8.65	0.134	8.85	0.328	8.84	0.317
24	9.24	8.87	0.368	8.95	0.293	9.08	0.156
25	8.52	8.79	0.267	8.91	0.387	9.01	0.494
26	8.00	7.93	0.067	7.81	0.193	8.00	0.004
27	6.47	7.42	0.947	7.40	0.932	7.67	1.200
28	8.17	8.87	0.702	8.79	0.624	8.76	0.592
29	9.02	9.14	0.118	9.27	0.254	9.02	0.003
30	8.19	8.00	0.190	7.87	0.319	8.04	0.150
31	8.70	8.58	0.121	8.28	0.415	8.54	0.161
32	9.21	8.95	0.258	8.89	0.322	8.98	0.229
33	9.37	8.80	0.571	8.65	0.721	8.67	0.702
34	9.12	8.84	0.278	8.97	0.154	8.81	0.307
35	9.06	9.01	0.053	9.01	0.045	8.96	0.103
36	9.40	8.76	0.639	8.76	0.639	8.63	0.773
37	9.08	8.82	0.256	8.91	0.167	8.82	0.264
38	9.00	9.31	0.315	9.00	0.005	9.07	0.074
39	8.70	9.16	0.456	9.32	0.615	9.08	0.375
40	8.77	9.27	0.500	8.73	0.038	9.04	0.271
41	7.92	8.20	0.279	8.22	0.297	8.10	0.181
42	7.24	7.70	0.463	7.39	0.154	7.82	0.576
43	8.09	8.87	0.783	8.56	0.472	8.70	0.605
44	8.07	8.39	0.322	8.69	0.623	8.26	0.192
45	7.60	8.13	0.531	7.98	0.377	7.96	0.360
46	9.04	8.68	0.362	9.10	0.055	8.39	0.655
47	9.05	9.40	0.348	9.06	0.007	9.05	0.003
48	9.00	8.77	0.228	8.97	0.031	8.96	0.036
49	9.01	9.05	0.036	9.06	0.046	9.01	0.001
50	9.24	8.94	0.299	9.17	0.067	8.97	0.267
51	9.14	8.56	0.580	8.80	0.342	8.65	0.493
52	9.60	9.71	0.109	9.50	0.099	9.64	0.045

Table 5. contd....

No.	pIC ₅₀ exp	MLR	AE	ANN	AE	SVM	AE
		pIC ₅₀ pred		pIC ₅₀ pred		pIC ₅₀ pred	
53	8.41	9.10	0.687	9.04	0.627	9.24	0.834
54	9.74	9.00	0.741	9.11	0.632	9.00	0.735
55	9.74	8.78	0.956	8.88	0.855	8.81	0.930
56	9.29	9.05	0.239	9.16	0.129	9.13	0.157
57	9.15	9.29	0.145	9.38	0.227	9.27	0.119
58	9.35	8.84	0.507	9.09	0.256	9.07	0.281
59	7.36	7.95	0.594	7.70	0.341	8.22	0.858
60	8.72	8.15	0.575	8.51	0.213	8.12	0.597
61	8.32	8.03	0.285	7.88	0.438	8.05	0.270
62	8.72	8.34	0.377	8.42	0.296	8.45	0.274
63	8.26	8.21	0.045	8.19	0.073	8.30	0.035
64	7.33	8.20	0.870	8.23	0.897	8.25	0.921
65	8.82	8.28	0.542	8.12	0.701	8.32	0.499
66	8.68	7.70	0.981	7.62	1.059	7.79	0.886
67	6.80	7.63	0.828	7.36	0.559	7.85	1.046
68	9.70	8.90	0.796	8.92	0.784	8.81	0.894
69	8.46	8.79	0.332	8.86	0.398	8.84	0.380
70	8.89	8.92	0.029	8.83	0.063	8.89	0.001
71	8.70	8.93	0.227	8.92	0.216	8.70	0.000
72	8.89	8.97	0.082	8.91	0.016	8.71	0.178
73	8.44	8.61	0.165	8.70	0.264	8.46	0.020
74	8.74	8.40	0.343	8.57	0.166	8.48	0.256
75	8.77	8.07	0.697	8.12	0.653	8.37	0.398
76	8.59	8.12	0.472	8.26	0.329	8.45	0.140
77	8.89	8.42	0.468	8.48	0.413	8.67	0.218
78	8.11	8.06	0.052	8.32	0.206	8.11	0.002
79	8.12	8.49	0.372	8.81	0.690	8.43	0.308
80	8.04	7.70	0.344	8.00	0.044	8.04	0.003
81	8.00	8.19	0.190	8.48	0.484	8.26	0.262
82	7.08	7.72	0.640	7.74	0.661	7.70	0.616
83	7.27	7.51	0.237	7.49	0.218	7.53	0.262
84	7.21	7.56	0.354	7.36	0.147	7.47	0.263
85	7.27	8.05	0.779	7.83	0.562	7.94	0.674
86	7.29	7.30	0.008	7.39	0.097	7.29	0.001
87	7.06	7.14	0.077	6.85	0.206	7.12	0.062
88	7.77	8.13	0.357	7.80	0.027	7.98	0.210

Table 5. contd....

No.	pIC ₅₀ exp	MLR	AE	ANN	AE	SVM	AE
		pIC ₅₀ pred		pIC ₅₀ pred		pIC ₅₀ pred	
89	7.23	7.81	0.579	7.48	0.248	7.76	0.534
90	8.08	7.94	0.144	7.98	0.096	8.08	0.005
91	7.80	7.68	0.119	7.73	0.067	7.80	0.003
92	8.09	7.48	0.606	7.48	0.611	7.56	0.527
93	7.92	8.05	0.128	8.20	0.281	8.12	0.202
94	7.21	7.72	0.510	7.80	0.589	7.76	0.553
95	7.74	7.74	0.002	7.75	0.010	7.73	0.013
96	8.40	8.09	0.309	8.36	0.043	8.08	0.325
97	7.42	7.68	0.257	7.69	0.274	7.79	0.371
98	7.49	7.49	0.003	7.44	0.049	7.57	0.080
99	7.32	7.60	0.283	7.46	0.139	7.57	0.248
100	7.96	7.65	0.313	7.58	0.381	7.61	0.353
101	7.92	7.36	0.565	7.28	0.640	7.35	0.566
102	7.48	8.00	0.518	8.09	0.610	8.08	0.604
103	7.21	7.41	0.203	7.33	0.119	7.50	0.294
104	8.21	7.76	0.446	7.87	0.338	7.89	0.320
105	7.57	7.53	0.043	7.54	0.028	7.58	0.007
106	7.72	7.63	0.089	7.48	0.242	7.72	0.003
107	7.82	7.74	0.084	7.75	0.066	7.79	0.027
108	7.89	7.72	0.168	7.52	0.371	7.83	0.061
109	7.54	7.45	0.090	7.35	0.186	7.54	0.002
110	8.77	8.37	0.399	8.46	0.311	8.32	0.451
111	8.21	8.22	0.007	8.13	0.084	8.24	0.029
112	8.39	7.98	0.413	8.12	0.265	7.93	0.459
113	8.57	8.68	0.106	8.73	0.159	8.57	0.004
114	8.05	8.38	0.333	8.53	0.476	8.38	0.333
115	8.80	9.00	0.197	8.81	0.010	8.88	0.078
116	8.68	9.30	0.624	9.21	0.528	9.26	0.584
117	9.08	9.24	0.160	9.26	0.178	9.18	0.104
118	8.82	8.98	0.157	9.04	0.225	8.97	0.145
119	8.57	8.92	0.352	9.01	0.443	8.84	0.272
120	8.57	8.89	0.322	8.90	0.333	8.86	0.288
121	8.20	8.91	0.710	8.92	0.723	8.89	0.687
122	9.15	9.13	0.018	9.21	0.058	9.12	0.028
123	8.96	9.12	0.162	9.15	0.192	9.07	0.106
124	8.74	8.85	0.106	8.87	0.127	8.81	0.068

Table 5. contd....

No.	pIC ₅₀ exp	MLR	AE	ANN	AE	SVM	AE
		pIC ₅₀ pred		pIC ₅₀ pred		pIC ₅₀ pred	
125	9.29	8.88	0.412	8.88	0.411	8.88	0.408
126	9.39	9.13	0.260	9.11	0.280	9.25	0.142
127	9.14	8.93	0.215	8.91	0.228	8.97	0.168
128	9.38	8.87	0.513	8.97	0.408	9.02	0.357
129	9.16	8.67	0.490	8.70	0.457	8.82	0.340
130	9.01	9.14	0.129	9.24	0.230	9.01	0.005
131	9.12	9.17	0.049	9.22	0.096	9.13	0.005
132	9.00	9.07	0.070	8.83	0.166	9.08	0.075
133	9.34	9.64	0.303	9.07	0.271	9.34	0.003
134	9.11	9.17	0.065	9.28	0.168	9.08	0.030
135	9.38	9.13	0.246	9.07	0.314	9.02	0.363
136	9.09	9.68	0.592	9.08	0.013	9.33	0.244
137	9.44	9.03	0.410	9.12	0.323	9.00	0.437
138	9.05	8.92	0.135	9.00	0.053	8.93	0.124
139	8.41	8.98	0.566	8.66	0.247	8.90	0.495
140	8.89	8.68	0.213	8.65	0.241	8.74	0.153
141	8.19	8.58	0.392	8.55	0.362	8.74	0.553
142	9.07	9.32	0.248	9.08	0.011	9.07	0.001
143	10.44	9.38	1.062	9.35	1.086	9.35	1.088
144	9.30	8.89	0.406	9.08	0.225	9.03	0.269
145	7.15	7.86	0.707	7.54	0.385	7.75	0.604
146	8.28	8.50	0.218	8.57	0.290	8.43	0.148
147	8.00	8.34	0.337	8.17	0.166	8.26	0.257
148	8.92	8.88	0.035	8.91	0.014	8.92	0.004
149	8.12	8.73	0.609	8.65	0.535	8.70	0.575
150	7.96	8.27	0.313	8.50	0.537	8.37	0.409
151	7.92	8.60	0.678	8.51	0.592	8.71	0.792
152	7.72	8.67	0.950	8.94	1.221	8.71	0.994
153	8.52	8.42	0.101	8.72	0.199	8.52	0.001
154	8.64	8.70	0.057	8.70	0.061	8.64	0.005
155	8.30	8.69	0.386	8.96	0.664	8.67	0.365
156	8.68	8.57	0.108	8.80	0.122	8.58	0.101
			Test set				
157	8.85	8.73	0.117	8.71	0.139	8.82	0.031
158	8.37	8.43	0.064	8.63	0.257	8.36	0.006
159	8.36	8.77	0.411	8.88	0.519	9.00	0.636

Table 5. contd....

No.	pIC ₅₀ exp	MLR	AE	ANN	AE	SVM	AE
		pIC ₅₀ pred		pIC ₅₀ pred		pIC ₅₀ pred	
160	8.82	8.59	0.226	8.80	0.020	8.88	0.061
161	8.32	8.78	0.457	8.82	0.504	8.97	0.655
162	8.32	8.38	0.065	8.23	0.094	8.16	0.163
163	9.19	8.78	0.405	9.03	0.163	8.83	0.359
164	7.17	8.49	1.316	8.42	1.249	8.40	1.225
165	9.29	8.64	0.652	8.75	0.539	8.66	0.627
166	9.09	8.93	0.158	9.05	0.038	9.14	0.051
167	9.80	9.75	0.051	9.43	0.375	9.56	0.236
168	9.10	9.11	0.009	8.82	0.279	9.11	0.013
169	7.85	8.85	0.997	8.52	0.670	8.85	0.998
170	8.92	8.64	0.284	8.88	0.042	8.73	0.186
171	7.89	7.56	0.326	7.70	0.192	7.58	0.312
172	7.23	7.51	0.284	7.42	0.190	7.63	0.401
173	7.15	7.74	0.595	7.80	0.647	7.74	0.595
174	7.72	7.68	0.037	7.54	0.176	7.72	0.003
175	7.89	7.68	0.212	7.64	0.248	7.69	0.203
176	7.70	7.68	0.018	7.67	0.033	7.65	0.054
177	7.89	8.44	0.549	9.02	1.131	8.22	0.327
178	8.12	7.84	0.280	7.98	0.135	7.93	0.187
179	7.42	7.70	0.285	7.66	0.244	7.80	0.383
180	7.09	7.98	0.888	7.86	0.770	7.85	0.764
181	9.36	9.00	0.364	9.08	0.280	8.98	0.379
182	9.31	9.59	0.277	9.18	0.126	9.34	0.029
183	8.21	9.05	0.841	8.83	0.625	8.90	0.692
184	9.17	9.05	0.117	9.05	0.120	9.22	0.051
185	9.38	8.92	0.458	8.89	0.493	9.06	0.316
186	8.89	9.23	0.335	9.25	0.355	9.14	0.250
187	8.70	9.05	0.349	8.91	0.207	9.10	0.401
188	9.14	9.31	0.171	9.30	0.161	9.34	0.204
189	8.72	8.71	0.006	8.72	0.001	8.55	0.168
190	8.96	8.88	0.076	9.23	0.270	8.83	0.132
191	8.89	8.68	0.209	8.82	0.071	8.68	0.211
192	8.68	8.50	0.181	8.66	0.017	8.54	0.144

External Validation of the Proposed Models

To confirm that the models would be useful for application to compounds other than those in the training set and for

comparison of the three proposed models, external validation was performed using the 36 data point test set that was not used in descriptor selection and model building. The predicted values for the test set for different models are given in

Table 6. AAE's of the Proposed Models Using Different Chemometrics Methods

	AAE of Training Set (N=156)	AAE of Test Set (N=36)
MLR	0.375±0.279	0.335±0.300
ANN	0.342±0.273	0.316±0.299
SVM	0.331±0.303	0.318±0.291

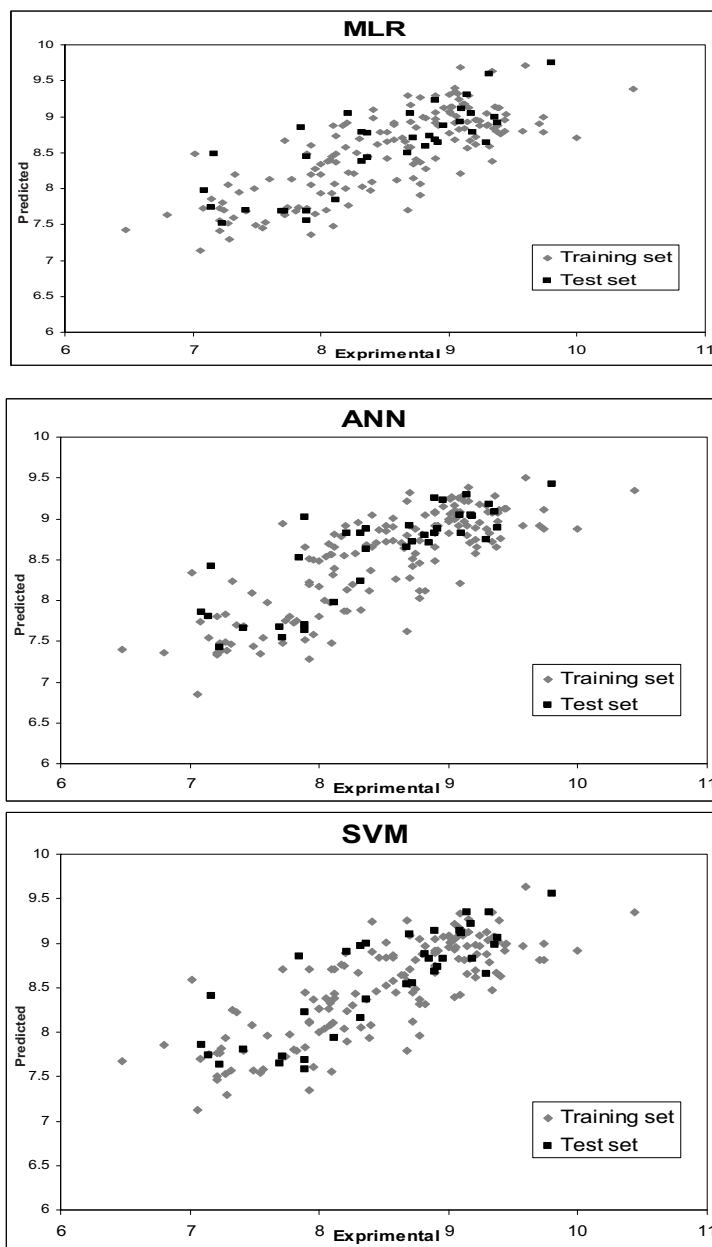
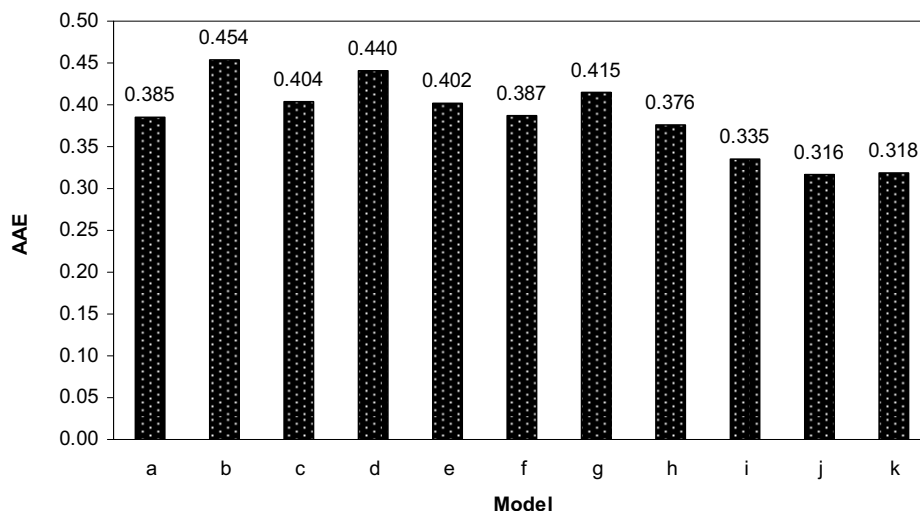
**Fig. (4).** Experimental versus predicted pIC₅₀ values using MLR, ANN and SVM models.

Table 5. Fig. (4) shows the experimental versus predicted values for training (156 data points) and test sets (36 data points). The AAE values of training and test compounds are summarized in Table 6. Careful review of these data reveals that the developed models possess good prediction capability. To further confirm this observation, some statistical cri-

teria for evaluating the prediction capability and robustness of the model were calculated for the external test set (see Table 7) and revealed that the proposed models built using MLR, ANN and SVM are robust and valid for external prediction.

Table 7. Statistical Parameters for External Validation of Three Proposed Models

Statistical Criteria	MLR	ANN	SVM
$R^2 > 0.6$	0.642	0.674	0.677
$(R^2 - R_0^2) / R^2 < 0.1$	0.000	0.003	0.003
$0.85 \leq K \leq 1.15$	0.988	0.988	0.987
$R_m^2 > 0.5$	0.633	0.644	0.644

**Fig. (5).** Comparison of the accuracy of different 3D-QSAR models (a to h) and the proposed 2D-QSAR models (i: MLR, j: ANN, k: SVM) in this study for 36 members of the test data set.

Comparison of the MLR, ANN and SVM 2D-models and 3D-models

Table 5 shows the predicted pIC_{50} values along with AE values for the MLR, ANN and SVM models for 192 data points (156 in the training set and 36 in the test set). The AAE's for the ANN and SVM models are better than those for the MLR model (Table 6) and R for these non-linear models is also greater than that for the MLR model (Tables 4 and 7), revealing that the SVM and ANN models are more accurate than the MLR model.

Also we compared AAE values of the new models for the test set (36 data points) with those of the 3D-QSAR models using COMFA and COMSIA [9] in Fig. (5). These results show that the 2D-QSAR models predict the IC_{50} 's of imidazole-based FTIs in the test set more accurately than the 3D-QSAR models. Considering that 2D-QSAR models are simpler than 3D-QSAR models, the new models represent a significant advance over the previous work.

CONCLUSION

In this study different chemometric methods were used to build models to predict the activities of imidazole-containing farnesyltransferase inhibitors. A large collection of descriptors was used to represent the FTI structures. The results

indicate that the volume, shape and polarity are important for the activity of the studied compounds and the proposed 2D-QSAR models constructed using linear (MLR) and non-linear (ANN and SVM) methods can be used to predict accurately the activities of FTIs. The non-linear models are superior to the linear model in this work. In addition, the prediction accuracy of the 2D-QSAR models is comparable and slightly better than that of previously published 3D-QSAR models. The proposed models could be used in drug design for evaluation of novel imidazole-containing FTIs.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

The authors would like to thank the Student Research Committee, Tabriz University of Medical Sciences for partial financial support under grant (under grant No. 90/2/3). Also we thank Dr. Mohammad Amin Abolghasemi Fakhree for his help in this study. RJD thanks US NIH National Center for Research Resources Research Facilities Improvements Program C06 RR-14503-01.

REFERENCES

- [1] Appels, N.M.G.M.; Beijnen, J.H.; Schellens, J.H.M. Development of farnesyl transferase inhibitors: A review. *Oncologist*, **2005**, *10*, 565-578.
- [2] Chatterjee, M.; van Golen, K.L. Farnesyl transferase inhibitor treatment of breast cancer cells leads to altered RhoA and RhoC GTPase activity and induces a dormant phenotype. *Int. J. Cancer*, **2011**, *129*, 61-69.
- [3] Wlodarczyk, N.; Le Broc-Ryckewaert, D.; Gilleron, P.; Lemoine, A.; Farce, A.; Chavatte, P.; Dubois, J.; Pommery, N.; Henichart, J.P.; Furman, C.; Millet, R. Potent Farnesyltransferase Inhibitors with 1,4-Diazepane Scaffolds as Novel Destabilizing Microtubule Agents in Hormone-Resistant Prostate Cancer. *J. Med. Chem.*, **2011**, *54*, 1178-1190.
- [4] Yanamandra, N.; Buzzeo, R.W.; Gabriel, M.; Hazlehurst, L.A.; Mari, Y.; Beaupre, D.M.; Cuevas, J. Tipifarnib-Induced Apoptosis in Acute Myeloid Leukemia and Multiple Myeloma Cells Depends on Ca(2+) Influx through Plasma Membrane Ca(2+) Channels. *J. Pharmacol. Exp. Ther.*, **2011**, *337*, 636-643.
- [5] Jabbour, E.; Kantarjian, H.; Ravandi, F.; Garcia-Manero, G.; Estrov, Z.; Verstovsek, S.; O'Brien, S.; Faderl, S.; Thomas, D.A.; Wright, J.J.; Cortes, J. A Phase 1-2 Study of a Farnesyltransferase Inhibitor, Tipifarnib, Combined With Idarubicin and Cytarabine for Patients With Newly Diagnosed Acute Myeloid Leukemia and High-Risk Myelodysplastic Syndrome. *Cancer*, **2011**, *117*, 1236-1244.
- [6] Bendale, P.; Olepu, S.; Suryadevara, P.K.; Bulbule, V.; Rivas, K.; Nallan, L.; Smart, B.; Yokoyama, K.; Ankala, S.; Pendyala, P.R.; Floyd, D.; Lombardo, L.J.; Williams, D.K.; Buckner, F.S.; Chakrabarti, D.; Verlinde, C.L.M.J.; Van Voorhis, W.C.; Gelb, M.H. Second generation tetrahydroquinoline-based protein farnesyltransferase inhibitors as antimalarials. *J. Med. Chem.*, **2007**, *50*, 4585-4605.
- [7] Roy, K.; Ojha, P.K. Advances in quantitative structure-activity relationship models of antimalarials. *Expert Opin. Drug Discov.*, **2010**, *5*, 751-778.
- [8] Kraus, J.M.; Tatipaka, H.B.; McGuffin, S.A.; Chennamaneni, N.K.; Karimi, M.; Arif, J.; Verlinde, C.; Buckner, F.S.; Gelb, M.H. Second Generation Analogues of the Cancer Drug Clinical Candidate Tipifarnib for Anti-Chagas Disease Drug Discovery. *J. Med. Chem.*, **2010**, *53*, 3887-3898.
- [9] Xie, A.; Odde, S.; Prasanna, S.; Doerksen, R.J. Imidazole-containing farnesyltransferase inhibitors: 3D quantitative structure-activity relationships and molecular docking. *J. Comput. Aided Mol. Des.*, **2009**, *23*, 431-448.
- [10] Puzyn, T.; Leszczynski, J.; Cronin, M.T.D. *Recent Advances in QSAR Studies*. Springer, Dordrecht **2010**.
- [11] Verma, J.; Khedkar, V.M.; Coutinho, E.C. 3D-QSAR in drug design - A review. *Curr. Top. Med. Chem.*, **2010**, *10*, 95-115.
- [12] González, M.P.; Caballero, J.; Tundidor-Camba, A.; Helguera, A.M.; Fernández, M. Modeling of farnesyltransferase inhibition by some thiol and non-thiol peptidomimetic inhibitors using genetic neural networks and RDF approaches. *Bioorg. Med. Chem.*, **2006**, *14*, 200-213.
- [13] Fernández, M.; Tundidor-Camba, A.; Caballero, J.M. 2D Autocorrelation modeling of the activity of trihalobenzocycloheptapyridine analogues as farnesyl protein transferase inhibitors. *Mol. Simul.*, **2005**, *31*, 575-584.
- [14] Giraud, E.; Luttmann, C.; Lavelle, F.; Riou, J.F.; Mailliet, P.; Laoui, A. Multivariate data analysis using D-optimal designs, partial least squares, and response surface modeling: A directional approach for the analysis of farnesyltransferase inhibitors. *J. Med. Chem.*, **2000**, *43*, 1807-1816.
- [15] Polley, M.J.; Winkler, D.A.; Burden, F.R. Broad-based quantitative structure-activity relationship modeling of potency and selectivity of farnesyltransferase inhibitors using a Bayesian regularized neural network. *J. Med. Chem.*, **2004**, *47*, 6230-6238.
- [16] Chaurasia, S.; Srivastava, A.K.; Nath, A.; Srivastava, M.K.; Pandey, A. Quantitative structure activity relationship studies on a series of tetrahydroquinoline-based farnesyltransferase inhibitors. *Oxid. Commun.*, **2007**, *30*, 778-787.
- [17] Cormanich, R.A.; Freitas, M.P.; Rittner, R. 2D Chemical Drawings Correlate to Bioactivities: MIA-QSAR Modelling of Antimalarial Activities of 2,5-Diaminobenzophenone Derivatives. *J. Braz. Chem. Soc.*, **2011**, *22*, 637-642.
- [18] Gaurav, A.; Gautam, V.; Singh, R. Exploring the structure activity relationships of imidazole containing tetrahydrobenzodiazepines as farnesyltransferase inhibitors: A QSAR study. *Lett Drug Des. Discov.*, **2011**, *8*, 506-515.
- [19] Xie, A.; Clark, S.R.; Prasanna, S.; Doerksen, R.J. Three-dimensional quantitative structure-farnesyltransferase inhibition analysis for some diaminobenzophenones. *J. Enzyme Inhib. Med. Chem.*, **2009**, *24*, 1220-1228.
- [20] Thomas, G. *Medicinal chemistry*, 2nd ed., John Wiley, Chichester Hoboken, NJ, **2007**.
- [21] Kharkar, P.S. Two-dimensional (2D) in silico models for Absorption, Distribution, Metabolism, Excretion and Toxicity (ADME/T) in drug discovery. *Curr. Top. Med. Chem.*, **2010**, *10*, 116-126.
- [22] Habibi-Yangjeh, A. QSAR study of the 5-HT1A receptor affinities of arylpiperazines using a genetic algorithm-artificial neural network model. *Monatsh. Chem.*, **2009**, *140*, 523-530.
- [23] Soltani, S.; Abolhasani, H.; Zarghi, A.; Jouyban, A. QSAR analysis of diaryl COX-2 inhibitors: Comparison of feature selection and train-test data selection methods. *Eur. J. Med. Chem.*, **2010**, *45*, 2753-2760.
- [24] Dastmalchi, S.; Hamzeh-Mivehroud, M.; Ghafourian, T.; Hamzei, H. Molecular modeling of histamine H3 receptor and QSAR studies on arylbenzofuran derived H3 antagonists. *J. Mol. Graph. Model.*, **2008**, *26*, 834-844.
- [25] Leardi, R.; Seasholtz, M.B.; Pell, R.J. Variable selection for multivariate calibration using a genetic algorithm: Prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Anal. Chim. Acta*, **2002**, *461*, 189-200.
- [26] Dearden, J.C.; Cronin, M.T.D.; Kaiser, K.L.E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.*, **2009**, *20*, 241-266.
- [27] Jalali-Heravi, M.; Asadollahi-Baboli, M.; Shabbazikhah, P. QSAR study of heparanase inhibitors activity using artificial neural networks and Levenberg-Marquardt algorithm. *Eur. J. Med. Chem.*, **2008**, *43*, 548-556.
- [28] Arab Chamjangali, M.; Beglari, M.; Bagherian, G. Prediction of cytotoxicity data (CC50) of anti-HIV 5-phenyl-1-phenylamino-1H-imidazole derivatives by artificial neural network trained with Levenberg-Marquardt algorithm. *J. Mol. Graph. Model.*, **2007**, *26*, 360-367.
- [29] Arab Chamjangali, M. Modelling of cytotoxicity data (CC50) of anti-HIV 1-[5-chlorophenyl] sulfonyl-1H-pyrrole derivatives using calculated molecular descriptors and levenberg-marquardt artificial neural network. *Chem. Biol. Drug Des.*, **2009**, *73*, 456-465.
- [30] Vapnik, V. *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, **1995**.
- [31] Shahlai, M.; Fassihi, A.; Saghale, L. Application of PC-ANN and PH-LS-SVM in QSAR of CCR1 antagonist compounds: A comparative study. *Eur. J. Med. Chem.* **2010**, *45*, 1572-1582.
- [32] Cheng, Z.; Zhang, Y.; Fu, W. QSAR study of carboxylic acid derivatives as HIV-1 Integrase inhibitors. *Eur. J. Med. Chem.*, **2010**, *45*, 3970-3980.
- [33] Darnag, R.; Mostapha Mazouz, E.L.; Schmitzer, A.; Villemain, D.; Jarid, A.; Cherqaoui, D. Support vector machines: Development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives. *Eur. J. Med. Chem.*, **2010**, *45*, 1590-1597.
- [34] Asadpour-Zeynali, K.; Soheili-Azad, P. Simultaneous polarographic determination of isoniazid and rifampicin by differential pulse polarography method and support vector regression. *Electrochim. Acta*, **2010**, *55*, 6570-6576.
- [35] Katritzky, A.R.; Kuanar, M.; Slavov, S.; Hall, C.D.; Karelson, M.; Kahn, I.; Dobchev, D.A. Quantitative correlation of physical and chemical properties with chemical structure: Utility for prediction. *Chem. Rev.*, **2010**, *110*, 5714-5789.
- [36] Louis, B.; Agrawal, V.K.; Khadikar, P.V. Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses. *Eur. J. Med. Chem.*, **2010**, *45*, 4018-4025.
- [37] Van De Waterbeemd, H.; Lennemäs, H.; Artursson, P. *Drug Bioavailability: Estimation of Solubility, Permeability,*

- Absorption, and Bioavailability*, 2nd ed., Wiley-VCH, Weinheim **2009**.
- [38] Golbraikh, A.; Tropsha, A. Beware of q²! *J. Mol. Graph. Model.* **2002**, *20*, 269-276.
- [39] Roy, P.P.; Roy, K. On some aspects of variable selection for partial least squares regression models. *QSAR Comb. Sci.*, **2008**, *27*, 302-313.

Received: January 10, 2012

Revised: July 05, 2012

Accepted: July 08, 2012