Arabian Journal of Chemistry (2016) 9, S185–S190

King Saud University

Arabian Journal of Chemistry

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

A quantitative structure–activity relationship study on HIV-1 integrase inhibitors using genetic algorithm, artificial neural networks and different statistical methods

Ghasem Ghasemi [a,*], Mahyar Nirouei [b], Shahab Shariati [a], Parviz Abdolmaleki [c], Zinab Rastgoo [a]

[a] *Department of Chemistry, Rasht Branch, Islamic Azad University, Rasht, Iran*
[b] *Department of Electrical Engineering, Lahijan Branch, Islamic Azad University, Lahijan, Iran*
[c] *Department of Bio-Medical Physics, Faculty of Science, Tarbiat Modares University, P.O. Box 14115/175, Tehran, Iran*

Received 13 January 2011; accepted 3 March 2011
Available online 9 March 2011

KEYWORDS

Quantitative structure–activity relationship;
Tricyclic phthalimide;
Genetic algorithm;
Artificial neural network

Abstract In this work, quantitative structure–activity relationship (QSAR) study has been done on tricyclic phthalimide analogues acting as HIV-1 integrase inhibitors. Forty compounds were used in this study. Genetic algorithm (GA), artificial neural network (ANN) and multiple linear regressions (MLR) were utilized to construct the non-linear and linear QSAR models. It revealed that the GA–ANN model was much better than other models. For this purpose, ab initio geometry optimization performed at B3LYP level with a known basis set 6–31G (d). Hyperchem, ChemOffice and Gaussian 98W softwares were used for geometry optimization of the molecules and calculation of the quantum chemical descriptors. To include some of the correlation energy, the calculation was done with the density functional theory (DFT) with the same basis set and Becke's three parameter hybrid functional using the LYP correlation functional (B3LYP/6–31G (d)). For the calculations in solution phase, the polarized continuum model (PCM) was used and also included optimizations at gas-phase B3LYP/6–31G (d) level for comparison. In the aqueous phase, the root–mean–square errors of the training set and the test set for GA–ANN model using jack–knife method, were 0.1409, 0.1804, respectively. In the gas phase, the

* Corresponding author. Tel.: +98 1316613730; fax: +98 1314224949.
E-mail addresses: ghasemi@iaurasht.ac.ir, ghassemi47@gmail.com (G. Ghasemi).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.arabjc.2011.03.006

root–mean–square errors of the training set and the test set for GA–ANN model were 0.1408, 0.3103, respectively. Also, the R^2 values in the aqueous and the gas phase were obtained as 0.91, 0.82, respectively.

1. Introduction

HIV-1 integrase (IN) catalyzes two distinct reactions: the terminal cleavage at each 3′ end of the proviral DNA removing a pair of bases and the strand transfer which results in the joining of each 3′ end to 5′-phosphates in the target DNA. Such integration is essential for the production of progeny viruses and therefore therapeutic agents that can inhibit this process should be effective anti-HIV agents. HIV-IN has also been recognized as a safe target against HIV because there are no similar enzymes involved in human cellular function (Sakai et al., 1993; Taddeo et al., 1994; Engelman et al., 1995).

The purpose of QSAR study is to find a relation between the composition or structure of a compound with its bio or chemical activity, in order to design a new compound with expected properties or predict the properties of an unknown compound. Up to now, a lot of successful applications have been reported in many different types of cases, e.g., medicine design, environmental chemistry exploration, pesticide searching, etc (Topliss and Edwards, 1979; Hasegawa and Miyashita, 1992).

The artificial neural networks (ANNs) are known as a good method in expressing highly non-linear relationship between the input and output variables, hence, greater interests were attracted in applying them to the pattern classification of complex compounds (Huuskonen, 2000; Schneider et al., 1999; Jalali-Heravi and Parastar, 2000; Burden and Winkler, 1999; Burden et al., 2000).

Genetic algorithms (GAs) were introduced by Holland. They mimic nature's evolutionary method of adaptation to a changing environment. GAs are stochastic optimization methods that provide powerful means to perform directed random searches in a large problem space as encountered in chemometrics and drug design (Hasegawa, 1999; Handschuh and Gasteiger, 2000; Kimura, 1998).

In multiple linear regression (MLR), for a given data set consisting of a target variable and M descriptors for n compounds, a model is made with good fitting to define the combination of m descriptors ($m < M$) on target variable. Running through all combinations usually is too time-consuming. Therefore, several approximate methods have been proposed for this reason, but none of them guarantied to find very best combination in all cases. The best found model for a given data set may differ from one method to another method. So a real QSAR model should be compared to pseudo models based on random numbers preferably using the same descriptor selection procedure (Livingstone and Salt, 2005). In order to evaluate the effectiveness of different methods in obtaining QSAR models, cross-validation method is used.

2. Computational details

The 3D structures of the molecules were generated using the built optimum option of Hyperchem software (version 8.0), Then, the structures were fully optimized based on the ab initio method, using DFT level of theory. Hyperchem, ChemOffice and Dragon (version 3.0) programs were employed to calculate the molecular descriptors.

All calculations were performed using Gaussian 98W program series. Geometry optimization of forty compounds was carried out by B3LYP method employing 6–31G (d) basis set with no initial symmetry restrictions and assuming C1 point group which were drawn in Hyperchem. In order to show the effect of solvent environment on the structures, all structures were optimized in H_2O solvent.

In this study, the independent variables were molecular descriptors and the dependent variables were the actual half maximal inhibitory concentration (IC_{50}) values. Overall, more than 1039 theoretical descriptors were selected and calculated. These descriptors can be classified into several groups including: (i) topological, (ii) geometrical, (iii) MoRSE, (iv) RDF, (v) GETAWAY, (vi) autocorrelations and (vii) WHIM descriptors.

For each compound in the training sets, the correlation equation was derived with the same descriptors. Then, the obtained equation was used to predict log ($1/IC_{50}$) values for the compounds from the corresponding test sets. The efficiency of QSAR models for prediction of log ($1/IC_{50}$) values was estimated using the cross-validation method.

In the present work, stepwise multiple linear regression (stepwise-MLR) and GA variable subset selection methods were used for the selection of the most relevant descriptors from all of the descriptors. These descriptors would be used as inputs of the ANN.

Totally 1039 descriptors were generated that were too many to be fitted in our models. So, it was necessary to reduce the number of descriptors through an objective feature selection which was performed in three steps. First, descriptors that had the same value for at least 70% of compounds within the dataset were removed. In next step, descriptors with correlation coefficients less than 0.4 with the dependent variable were regarded redundant and removed. Finally, since highly correlated descriptors provide approximately identical information, a pair wise correlation was performed. When their correlation coefficient exceeded 0.90, one of two descriptors was randomly removed.

GA was utilized as the mean for non-linear feature selection. After calculation of the correlation between descriptors, 63 descriptors were used as input of the ANN in aqueous phase. In other words, the defined chromosome contains 58 genes, one gene for each feature, which can take two values. A value of 0 indicates that the corresponding feature is not selected, and a value of 1 means that the feature is selected. Therefore, there are 2^{63} possible feature subsets. GA selects the best features from these possible feature subsets during different generations. In each generation, the population is probabilistically modified, generating new chromosomes that may have a better chance of solving the

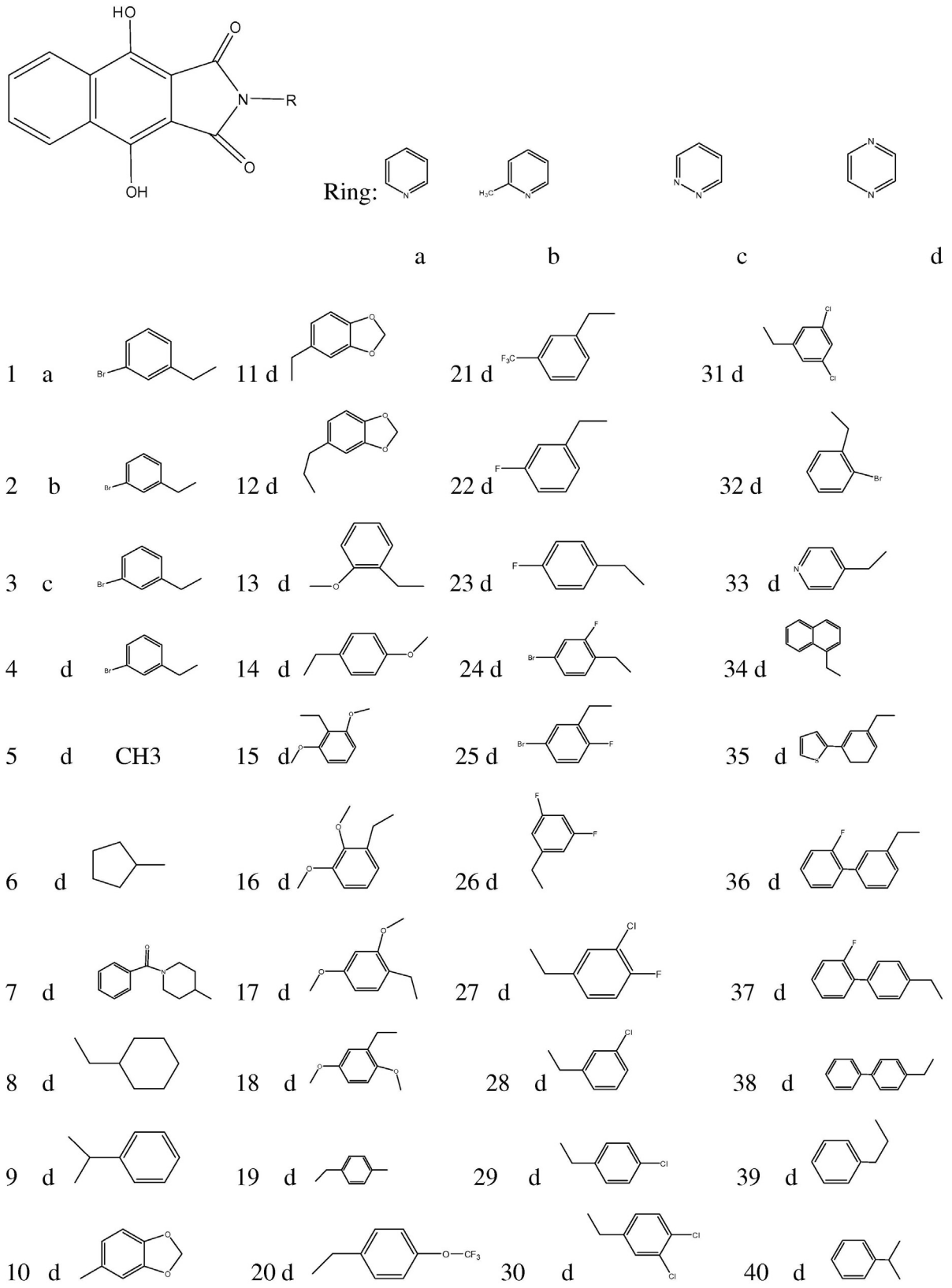

Figure 1 The molecular structure of phthalimide analogues.

problem. New characteristics are introduced into a chromosome by crossover and mutation. The probability of survival or reproduction of an individual depends more or less on its fitness to the environment. Each feature in a given feature space is treated as a gene and is encoded by a binary digit (bit) in a chromosome.

Table 1 Experimental and predicted values of log ($1/IC_{50}$) using Jack–Knife model.

Calculated (Jack–Knife) gas	Calculated (Jack–Knife) PCM	Observed log ($1/IC_{50}$)
6.5516	6.7009	6.420
6.6398	6.2908	6.590
6.4050	6.6533	5.440
6.4311	6.6774	6.680
3.1604	4.5125	4.310
3.9496	4.6055	4.980
6.1276	6.0073	5.620
5.3327	5.4687	5.660
5.5812	5.8385	5.980
5.4801	5.3665	5.000
6.0373	5.9212	6.250
6.0143	6.1461	5.850
5.8932	5.9627	6.090
6.2142	6.1039	6.250
6.3538	6.2872	5.700
6.2527	6.3440	6.370
6.4281	6.2605	6.370
6.2004	5.9841	6.110
6.4004	5.7789	5.800
6.3610	5.7886	5.690
5.7097	6.4423	6.370
6.2716	6.2876	6.250
6.0468	6.1329	6.390
6.8675	6.3648	6.360
6.5993	6.4879	6.450
6.3413	6.5234	6.430
6.5138	6.5931	6.750
6.0691	6.5876	6.660
6.3272	6.2880	6.380
6.8970	6.8502	6.980
6.8162	6.7013	6.730
5.9714	6.0203	6.200
5.5838	5.6964	5.170
6.0169	5.6681	5.850
5.7583	6.2011	6.170
6.1521	5.9052	6.300
6.1695	6.2395	6.000
5.3564	5.9874	5.840
5.7120	5.7381	5.690
5.8028	5.8792	5.690

Table 2 Descriptors values for GA–MLR model.

Molecule	GATS6v	RDFo35m	QZZv	P2e	HATS2e
1	0.811	10.576	7.716	0.405	0.079
2	0.856	10.578	7.880	0.398	0.077
3	0.786	10.292	7.630	0.403	0.079
4	0.838	11.119	8.709	0.402	0.079
5	1.168	6.965	4.004	0.218	0.057
6	1.125	7.471	7.419	0.274	0.065
7	1.150	11.585	11.123	0.318	0.050
8	1.019	12.088	9.345	0.346	0.053
9	0.980	11.032	12.137	0.335	0.061
10	0.934	9.957	7.155	0.295	0.076
11	1.006	12.744	10.049	0.366	0.052
12	1.030	11.872	9.731	0.298	0.049
13	0.825	9.796	10.421	0.351	0.051
14	0.918	11.040	8.597	0.366	0.049
15	0.754	10.365	7.931	0.340	0.047
16	0.997	14.062	8.803	0.364	0.052
17	0.818	10.426	11.300	0.333	0.048
18	0.888	11.776	8.828	0.344	0.047
19	1.015	9.364	10.388	0.324	0.042
20	0.949	12.544	10.610	0.336	0.049
21	0.798	11.406	11.093	0.341	0.062
22	0.920	10.232	9.605	0.330	0.097
23	0.935	10.463	9.994	0.348	0.100
24	1.003	9.154	10.843	0.388	0.087
15	0.832	11.409	9.902	0.393	0.080
26	0.957	12.476	9.595	0.327	0.102
27	0.918	11.013	11.794	0.361	0.134
28	0.900	10.320	11.694	0.369	0.086
29	1.052	10.535	11.319	0.377	0.085
30	1.005	10.393	13.928	0.404	0.113
31	0.849	10.672	14.405	0.404	0.085
32	0.935	10.297	11.026	0.302	0.072
33	1.050	10.173	9.423	0.343	0.051
34	0.875	10.970	7.625	0.268	0.061
35	1.020	10.570	12.187	0.369	0.050
36	1.041	11.604	11.780	0.366	0.049
37	1.020	10.570	12.187	0.369	0.050
38	1.040	10.664	10.546	0.368	0.049
39	0.970	10.712	10.643	0.360	0.062
40	1.015	11.593	10.457	0.321	0.063

In our study, two point binary crossover and binary mutation were performed. The roulette wheel selection strategy was also used in the algorithm for parent selection. The relevant parameter settings such as population size: 40; number of generation: 100; probability of crossover: 0.8; probability of mutation: 0.1 were used. A lot of fitness functions were tested and the optimal fitness function, as the object of minimization by GA was found to be as follows:

$$F = 100 \times \mathrm{RMSE}_{\mathrm{CVSET}} \times \mathrm{RMSE}_{\mathrm{TSET}} \quad (1)$$

where $\mathrm{RMSE}_{\mathrm{CVSET}}$ and $\mathrm{RMSE}_{\mathrm{TSET}}$ are the root–mean–square errors of the training set and the test set, respectively.

Each fitness value was obtained in a cross validation procedure by removing eight cross validation (CVSET) individuals from the data set, remaining other 32 train set (TSET) ones each time. This was done in a way that each compound was used four times as a TSET member and once as a CVSET one. In this way, the average result of five different simulations was reported as the fitness value.

Table 3 The statistical parameters of different constructed QSAR models.

Method	RMSE test	RMSE train	R^2
GA–ANN Jack–Knife (gas)	0.3103	0.1408	0.82
GA–ANN (gas) cross validation	0.3836	0.1532	–
GA–ANN Jack–Knife (PCM)	0.1804	0.1409	0.91
GA–ANN (PCM) cross validation	0.5440	0.14010	–

3. Results and discussion

The structures of the tricyclic phthalimide analogues used in this study are shown in Fig. 1.

Table 4 The descriptors selected using GA–ANN model.

Descriptors	
Aqueous	Gas
X2A	X2A
R1e+	IDDE
X3A	IC1
PW4	ATS7m
BAC	ATS8m
IC1	MATS6p
IC2	GATS6v
CIC3	GATS1p
SEigm	RDF035m
ATS8m	RDF095m
MATS8m	RDF055e
MATS3p	Mor27u
GATS3v	Mor05m
RDF035m	Mor29v
Mor27u	Mor32v
Mor27v	E3m
E3m	E3p
E1v	H6m
E2p	H3e
E3s	HATS6p
H6m	RTu+
HATS5v	R1e+
H1e	
H3e	
R6m	
R4v+	

Table 5 The results of genetic algorithm.

Descriptor symbol	Descriptor group	Meaning
X2A	Topological (1D)	Average connectivity index chi-2
MATS6p	Autocorrelation (2D)	Moran autocorrelation lag 6/weighted by atomic polarizability
GATS6v	Autocorrelation (2D)	Geary autocorrelation lag 6 weighted by atomic van der walls volumes
RDFo35m	RDF (3D)	Radial distribution function 3.5 weighted by atomic masses
RDFo55e	RDF (3D)	Radial distribution function 5.5 weighted by atomic sanderson electronegativities
Moro5m	3D MoRSE	3D MoRSE signal 05 weighted by atomic masses
Mor32v	3D MoRSE	3D MoRSE signal 32 weighted by atomic van der walls volumes
E3m	WHIM (3D)	3rd component accessibility directional WHIM index weighted by atomic masses
H6m	GETAWAY (3D)	H autocorrelation of lag 6 weighted by atomic masses
R1e+	GETAWAY (3D)	R maximal autocorrelation of lag 1 weighted by atomic sanderson electronegativity

The efficiency of the QSAR model to predict log (IC_{50}) value was also estimated using the internal cross-validation method. The resulted predictions of the log ($1/IC_{50}$) in gas and aqueous phases are given in Table 1.

Considering the experimental error, the overall prediction of the log ($1/IC_{50}$) values was quite satisfactory (without compound **3**). As shown in Table 1, the results of aqueous phase were much better than gas phase.

Two linear and non-linear variable selection methods were used to select the most significant descriptors (stepwise-MLR and GA) (Table 2). The selected descriptors through these methods were used to construct some linear and non-linear models by using MLR and ANN methods. Based on the types of variable selection method and also the types of the feature mapping technique, these models can be shown as MLR–ANN, GA–MLR and GA–ANN (de Weijer et al., 1992; Sheridan and Bush, 1993; Tominaga, 1999; Manallack and Livingstone, 1999). It revealed that the GA–ANN model was much better than other models (Table 3). Statistical parameters of different constructed QSAR models are shown in Table 3.

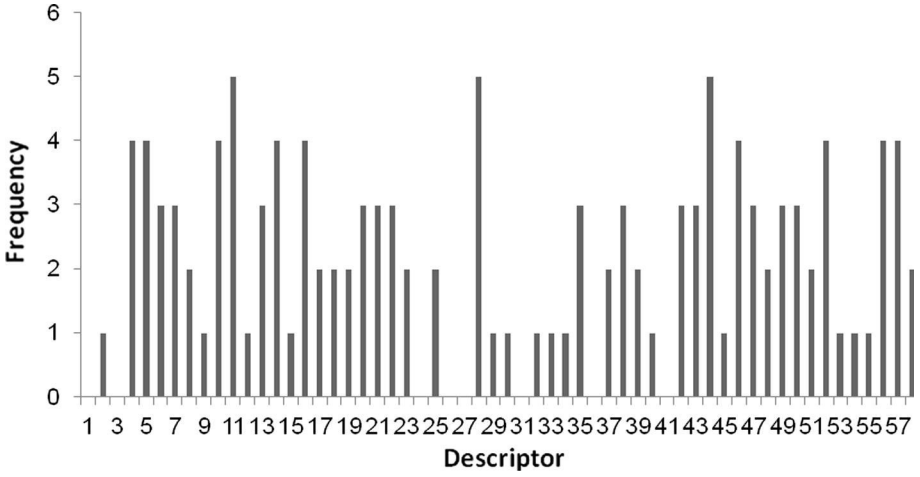

Figure 2 The results of Ga–ANN in gas phase.

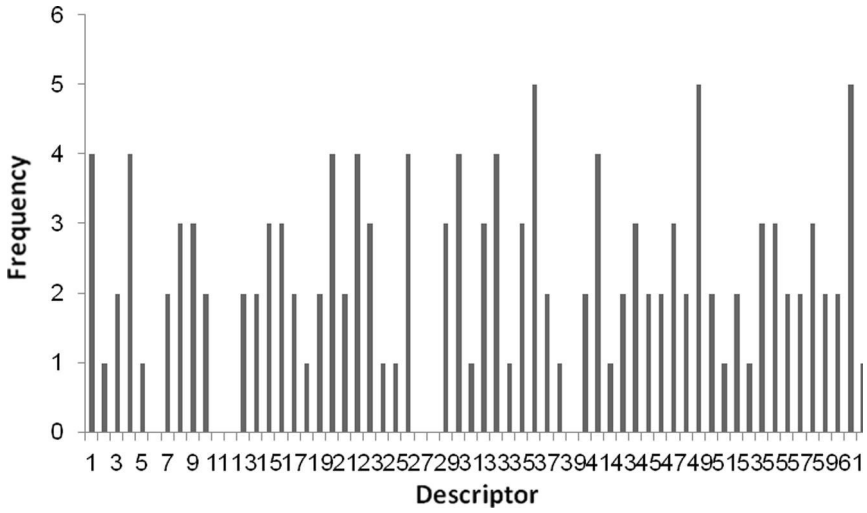

Figure 3 The results of Ga–ANN in aqueous phase.

As can be seen from this table, R^2 and RMSE values in aqueous phase are better than gas phase.

Since the chemical variation of the considered compounds is low, the selection of chemical descriptors, which can encode small variations between structures of molecules in data set, is very important. In this way, GETAWAY and WHIM descriptors are very informative 3D descriptors that can encode structural features of molecules and they are included in the GA–ANN model (Table 4). The ten most significant descriptors which were selected by GA are as follows: (Todeschini and Consonni, 2000; Consonni et al., 2002a,b) (with PCM)

X2A, MATS6p, GATS6v, RDFo35m, RDFo55e, Moro5m, Mor32v, E3m, H6*m* and R1e + .

These GA selected descriptors were used as inputs for the construction of ANN model (Table 5). As can be seen from this table, atomic mass, electronegativity and atomic polarizability were important descriptors in our study.

In the present study, two linear and non-linear variable selection methods were used to select the most significant descriptors. The MLR, ANN and GA were used to construct a quantitative relation between activities of tricyclic phthalimide analogues and their calculated descriptors (Figs. 2 and 3).

We have evaluated several layers ([3,1], [5,1], [7,1], [9,1], [11,1]) in GA and results are shown in Figs. 1 and 2.

4. Conclusion

In the present study, two linear and non-linear variable selection methods were used to select the most significant descriptors, and the MLR, ANN and GA were used to construct a quantitative relation between the activities of phthalimide analogues and their calculated descriptors. ANN has been successfully used for finding a QSAR model for tricyclic phthalimide analogues. It provides the best results among those we have tested. Our present attempt to correlate the log ($1/IC_{50}$) with theoretically calculated molecular descriptors has led to a relatively successful QSAR model that relates this complex molecular property to structural characteristics of the molecules.

The results obtained from this work indicate that the linear regression and ANN models exhibit reasonable prediction capabilities. Though the linear model was developed mainly for the purpose of structure–activity interpretation, the ANN model was primarily developed for predictive ability and classification.

Acknowledgment

The authors thank Dr. H. Fallah for various helpful contributions at all stages of the work.

References

Burden, F.R. et al., 2000. J. Chem. Inf. Comput. Sci. 40, 1423–1430.
Burden, F.R., Winkler, D.A., 1999. J. Med. Chem. 42, 3183–3187.
Consonni, V., Todeschini, R., Pavan, M., 2002a. J. Chem. Inf. Comput. Sci. 42, 682.
Consonni, V., Todeschini, R., Pavan, M., 2002b. J. Chem. Inf. Comput. Sci. 42, 693.
de Weijer, A.P., Buydens, L., Kateman, G., 1992. Chemom. Intell. Lab. Syst. 16, 77–86.
Engelman, A., Englund, G., Orenstein, J.M., Martin, M.A., Craigie, R., 1995. J. Virol. 69, 2729.
Handschuh, S., Gasteiger, J., 2000. J. Mol. Model. 6, 358–378.
Hasegawa, K., 1999. J. Chem. Inf. Comput. Sci. 39, 112–120.
Hasegawa, K., Miyashita, Y., 1992. Chemom. Intell. Lab. Syst. 16, 69–75.
Huuskonen, J., 2000. J. Chem. Inf. Comput. Sci. 40, 773–777.
Jalali-Heravi, M., Parastar, F., 2000. J. Chem. Inf. Comput. Sci. 40, 147–154.
Kimura, T., 1998. J. Chem. Inf. Comput. Sci. 38, 276–282.
Livingstone, D.J., Salt, D.W., 2005. Rev. Comput. Chem. 21, 287–348.
Manallack, D.T., Livingstone, D.J., 1999. Eur. J. Med. Chem. 34, 195–208.
Sakai, H., Kawamura, M., Sakuragi, J., Sakuragi, S., Shibata, R., Isimoto, A., Ono, N., Ueda, S., Adachi, A., 1993. J. Virol. 67, 1169.
Schneider, G. et al., 1999. J. Med. Chem. 42, 5072–5076.
Sheridan, R.P., Bush, B.L., 1993. J. Chem. Inf. Comput. Sci. 33, 756.
Taddeo, B., Haseltine, W.A., Farnet, C.M., 1994. J. Virol. 68, 8401.
Todeschini, R., Consonni, V., 2000. Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, Germany.
Tominaga, Y., 1999. Chemom. Intell. Lab. Syst. 49, 105–115.
Topliss, J.G., Edwards, R.P., 1979. J. Med. Chem. 10, 1238–1244.