Monatsh Chem (2009) 140:15–27
DOI 10.1007/s00706-008-0049-7

ORIGINAL PAPER

Application of principal component-genetic algorithm-artificial neural network for prediction acidity constant of various nitrogen-containing compounds in water

Aziz Habibi-Yangjeh · Eslam Pourbasheer · Mohammad Danandeh-Jenagharad

Received: 28 June 2008 / Accepted: 29 July 2008 / Published online: 30 September 2008

Abstract Principal component-genetic algorithm-multiparameter linear regression (PC-GA-MLR) and principal component-genetic algorithm-artificial neural network (PC-GA-ANN) models were applied for prediction acidity constant (pK_a) for various nitrogen-containing compounds. A data set that consisted of 282 various compounds, including 55 anilines, 77 amines, 82 pyridines, 14 pyrimidines, 26 imidazoles and benzimidazoles, and 28 quinolines, is used in this work. A large number of theoretical descriptors were calculated for each compound. The first 179 principal components (PCs) were found to explain more than 99.9% of variances in the original data matrix. From the pool of these PCs, the genetic algorithm was employed for selection of the best set of extracted PCs for PC-MLR and PC-ANN models. The models were generated using 15 PCs as variables. For evaluation of the predictive power of the models, pK_a values of 56 compounds in the prediction set were calculated. Root mean square errors (RMSE) for PC-GA-MLR and PC-GA-ANN models are 1.4863 and 0.0750. Comparison of the results obtained by the models reveals superiority of the PC-GA-ANN model relative to the PC-GA-MLR model. Mean percent deviation for the PC-GA-ANN model in the prediction set is 2.123. The improvements are due to the fact that pK_a of the compounds demonstrates non-linear correlations with the PCs.

A. Habibi-Yangjeh · E. Pourbasheer (✉) · M. Danandeh-Jenagharad
Department of Chemistry, Faculty of Science, University of Mohaghegh Ardabili, Ardabil, Iran
e-mail: ehsan@khayam.ut.ac.ir

Keywords Quantitative structure-activity relationship · Acidity constant · Genetic-algorithm · Principal components · Artificial neural network

Introduction

The acid–base processes are one of the most important types of reactions in chemistry and biochemistry. It has been shown that the acid–base properties affect the toxicity, chromatographic retention behavior, and pharmaceutical properties of organic acids and bases. On the other hand, it is well known that the pharmacokinetic properties, such as bioavailability, capacity to diffuse across many membranes, and other physical barriers of a compound can be strongly affected by its acid–base properties [1, 2]. Experimentally determined pK_a values are not always available from literature sources, and often estimated values are employed instead. Therefore, it is of interest to develop methods for estimating the acidity and basicity of various compounds [3].

Application of computational techniques to biology, physics, and chemistry is growing rapidly. The prediction of physicochemical and biological properties/activities of organic molecules is the main objective of quantitative structure-property/activity relationships (QSPRs/QSARs). QSPR/QSAR models are obtained on the basis of the correlation between the experimental values of the property/activity and descriptors reflecting the molecular structure of the compounds [4–16]. The QSPR/QSAR models now correlate chemical structure to a wide variety of physical, chemical, biological (including biomedical, toxicological, ecotoxicological), and technological properties [8–12]. To obtain a significant correlation, it is crucial that appropriate descriptors should be employed.

A wide variety of molecular descriptors has been reported for using in QSPR/QSAR models [17]. However, as the number of descriptors (variables) increases, the model becomes complicated, and its interpretation is difficult if many variables are used in modeling. Therefore, the application of these techniques usually requires variable selection for building well-fitted models. A better predictive model can be obtained by ortogonalization of the variables by means of principal component analysis (PCA) [18–21]. The PCA was used to compress the descriptor groups into principal components (PCs). In order to reduce the dimensionality of the independent variable space, a limited number of PCs are used [22–25]. Hence, selecting the significant and informative PCs is the main problem in all of the PCA-based calibration methods. Different methods have been addressed to select the significant PCs for calibration purposes [22–28]. The simplest and most common one is a top-down variable selection where the PCs are ranked in the order of decreasing eigenvalues, and the PC with the highest eigenvalue is considered as the most significant one; subsequently, the PCs are introduced into the calibration model. However, the magnitude of an eigenvalue is not necessarily a measure of its significance for the calibration [25]. In the other method, which is called correlation ranking, the PCs are ranked by their correlation coefficient with the property and selected by the procedure discussed for eigenvalue ranking [22, 23]. Better results are often achieved by this method. Recently, the genetic algorithm (GA) has been applied for the selection of the most relevant PCs instead of the older methods [26, 27]. Comparison of the results obtained using GA principal component selection with the two above-mentioned methods shows that GA gives a better result and close to the correlation ranking [26–28]. GA is a stochastic method to solve optimization problems applying the evolution hypothesis of Darwin and different genetic functions, i.e., cross-over and mutation [29, 30]. GA is robust, global, and generally more straightforward to apply in situations where there is little or no a priori knowledge about the process to be controlled [29].

Artificial neural networks (ANNs) have become popular in QSPR/QSAR models due to their success where complex non-linear relationships exist amongst data [31, 32]. An ANN is formed from an artificial neuron connected with coefficients (weights), which constitute the neural structure and are organized in layers. The layers of neurons between the input and output layers are called hidden layers. Neural networks do not need explicit formulation of the mathematical or physical relationships of the handled problem. These give ANNs an advantage over traditional fitting methods for some chemical applications. For these reasons, in recent years ANNs have been applied to a wide variety of chemical problems [33–42].

In the last years many theoretical studies have been performed for the prediction of pK_a values of various compounds using theoretical descriptors, but in most cases linear equations have been used in these studies [43–46]. Recently, separate multi-parameter linear QSAR models have been proposed between the acidity constant and molecular descriptors of 282 various nitrogen-containing compounds [47]. Average of root-mean square error for the separate models is 0.9183. In the present work, a single model has been proposed for the prediction of the pK_a values for the same set of compounds. In order to increase the ability of the model to predict the pK_a values for the compounds using single-linear and non-linear models, principal component-genetic algorithm-multiparameter linear regression (PC-GA-MLR) and principal component-genetic algorithm-artificial neural network (PC-GA-ANN) models were employed to generate QSAR models between the PCs and pK_a of 282 various nitrogen-containing compounds with diverse chemical structures, and results for the models were compared with each other, the previous work, and the experimental values.

Results and discussion

Principal component analysis

After the elimination of the constant descriptors and one of the collinear ones, 406 descriptors remained from 1,481 theoretical descriptors. The results of application of PCA on the descriptors data matrix show that 99.9% of the variances in the descriptors data matrix are explained by 179 first PCs. Therefore, we focused our analysis on these PCs, and the reminders, which are noisy factors, were not considered.

Principal component-genetic algorithm-multiparameter linear regression

Obtaining the number of significant PCs is the main problem in the PCA-based methods. The first 179 PCs were found to explain more than 99.9% of variances in the original data matrix. As noted previously, not all of the PCs are informative for QSAR/QSPR modeling [25–27]. Then, we used GA for the selection of the most relevant PCs instead of the older methods. The selected PCs are PC2, PC4, PC5, PC6, PC8, PC9, PC10, PC12, PC14, PC16, PC19, PC22, PC49, PC80, and PC115. As can be seen, the selected PCs are not based on their eigenvalue. For example, the PC2 and PC4 are selected, and the PC1 and PC3 are not considered in the model. This is due to the fact that the information contents of some extracted PCs may not be in the same direction of the activity data.

Table 1 The descriptors that have main correlations with selected PCs

Descriptor	Descriptor type	Brief description
Mor12p	Three-dimensional descriptors	3D-MoRSE—signal 12/weighted by atomic polarizabilities
Mor31v	Three-dimensional descriptors	3D-MoRSE—signal 31/weighted by atomic van der Waals volumes
IC0	Topological descriptors	Information content index (neighborhood symmetry of 0-order)
BIC1	Topological descriptors	Bond information content (neighborhood symmetry of 1-order)
Ms	Constitutional descriptors	Mean electrotopological state
GATS1p	Autocorrelation descriptors	Geary autocorrelation—lag 1/weighted by atomic polarizabilities
H-050	Atom-centered group	H attached to heteroatom
nCp	Functional group	Number of total primary C(sp3)

The eigenvectors, resulting from the PCA procedure, are mathematical factors and do not have any physical or chemical meanings. A transformation procedure is needed to transform these imaginary factors into real chemical factors [20]. In our research, we did not aim to transform the extracted PCs, but attempts were made to find the descriptors that mainly contribute to these PCs [26]. To do this, the correlation of the selected PCs with each one of the descriptors used in this study was examined. The descriptors that had the highest correlations ($R > 0.9$) with PCs are listed in Table 1.

Obviously, the selected PCs are relating to eight different descriptors. First and second descriptors are Mor12p and Mor31v. These descriptors belong to the 3D-MoRSE descriptors. 3D MoRSE descriptors (3D Molecule Representation of Structures based on Electron diffraction) are derived from Infrared spectra simulation using a generalized scattering function [17]. Mor12p and Mor31v were proposed as signal 12/weighted by atomic polarizabilities and signal 31/weighted by atomic van der Waals volumes, which relate to polarizabilities and van der Waals volumes of the atoms, respectively.

IC0 and BIC1 are the other descriptors that related to information content. The information content of a system having n elements is a measure of the degree of diversity of the elements in the set [17].

Ms is the mean electrotopological state. The electrotopological state gives information related to the electronic and topological state of the atom in the molecule [17].

The sixth descriptor is GATS1p (Geary autocorrelation—lag 1/weighted by atomic polarizabilities), which is one of the 2D autocorrelation descriptors. In this descriptor the Geary coefficient is a distance-type function, that is, function is any physico-chemical property calculated for each atom of the molecule, such as atomic mass, polarizability, etc. Therefore, the molecule atoms represent the set of discrete points in space and the atomic property the function evaluated at those points. The physico-chemical property in this case is atomic polarizability.

The next descriptor is H-050 (H attached to heteroatom). It is one of the atom-centered fragment descriptors that describes each atom by its own atom type and the bond types and atom types of its first neighbors. This descriptor represents the first neighbor (hydrogen) of heteroatom.

The final descriptor is nCp, which is one of the functional group count descriptors. The nCp represents the number of total primary C (sp3).

In summary, it is concluded that polarizabilities and van der Waals volumes of the atoms, diversity of the elements in compounds, electronic state of the atoms in the molecule, the number of first neighbor (hydrogen) of heteroatom, and the number of total primary C (sp3) play main roles in the acidity constant of the compounds.

Multiparameter linear correlation of pK_a values for 170 various compounds in the training set was obtained using 15 PCs selected by GA. The calculated values of pK_a for the compounds in training, validation, and prediction sets using the PC-GA-MLR model have been plotted versus the experimental values of it (Fig. 1).

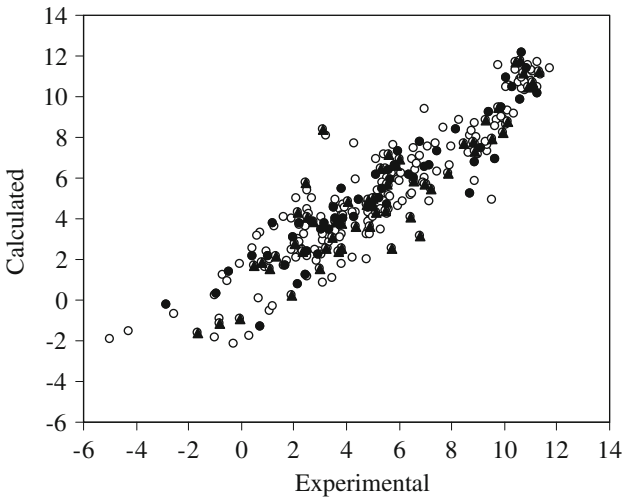

Fig. 1 Plot of the calculated values of pK_a from the PC-GA-MLR model versus the experimental values of it for training (*open circle*), validation (*filled circle*), and prediction (*filled triangle*) sets

Principal component-genetic algorithm-artificial neural network

To process the non-linear relationships existing between the acidity and the PCs, the ANN modeling method combined with PCA for dimension reduction and GA for feature selection was employed. A PC-GA-ANN model, which combines the PCs with ANN, is another PC-based calibration technique for non-linear modeling between the PCs and dependent variables [26–28]. The input vectors were the set of PCs that were selected by GA, and therefore, the number of nodes in the input layer was equal to the number of selected PCs. There are no rigorous theoretical principles for choosing the proper network topology, so different structures were tested in order to obtain the optimal hidden neurons and training cycles [34–42]. Before training the network, the number of nodes in the hidden layer was optimized. In order to optimize the number of nodes in the hidden layer, several training sessions were conducted with different numbers of hidden nodes (from 1 to 20). The root mean square error of training (RMSET) and validation (RMSEV) sets were obtained at various iterations for different numbers of neurons at the hidden layer, and the minimum value of RMSEV was recorded as the optimum value. The plot of RMSET and RMSEV versus the number of nodes in the hidden layer has been shown in Fig. 2. It is clear that the 15 nodes in the hidden layer is the optimum value.

This network consists of 15 inputs (PC2, PC4, PC5, PC6, PC8, PC9, PC10, PC12, PC14, PC16, PC19, PC22, PC49, PC80, and PC115), the same PCs in the PC-GA-MLR model, and one output for the pK_a. Then an ANN with architecture 15-15-1 was generated. It is noteworthy that training of the network was stopped when the RMSEV started to increase, i.e., when overtraining begins. The overtraining causes the ANN to lose its prediction power [32]. Therefore, during training of the network, it is desirable that iterations are stopped when overtraining begins. To control the overtraining of the network during the training procedure, the values of RMSET and RMSEV were calculated and recorded to monitor the extent of the learning in various iterations. Results showed that overfitting did not occur in the optimum architecture (Fig. 3).

The generated ANN was then trained using the training and validation sets for the optimization of the weights and biases. For the evaluation of the predictive power of the generated ANN, an optimized network was applied for prediction of the pK_a values in the prediction set, which were not used in the modeling procedure (Table 2).

The correlation equation for the all of calculated values of pK_a from the ANN model and the experimental values is as follows:

$$\mathrm{p}K_a(\mathrm{cal}) = 0.9925\mathrm{p}K_a(\mathrm{exp}) + 0.0397 \quad (1)$$

$$(R^2 = 0.9924; \quad \mathrm{MPD} = 2.259; \quad \mathrm{RMSE} = 0.3000; \quad F = 36560.57)$$

The calculated values of pK_a for the compounds in training, validation, and prediction sets using the ANN model have been plotted versus the experimental values of it in Fig. 4. As can be seen, there are four compounds in the data sets as outliers (compounds with numbers 111, 167, 121, and 170). The test that we used to detect the presence of outliers was a standard residual higher than $2 \times \delta$, where δ is equivalent to the standard deviation [48].

After removing of the outliers, the following equation was obtained:

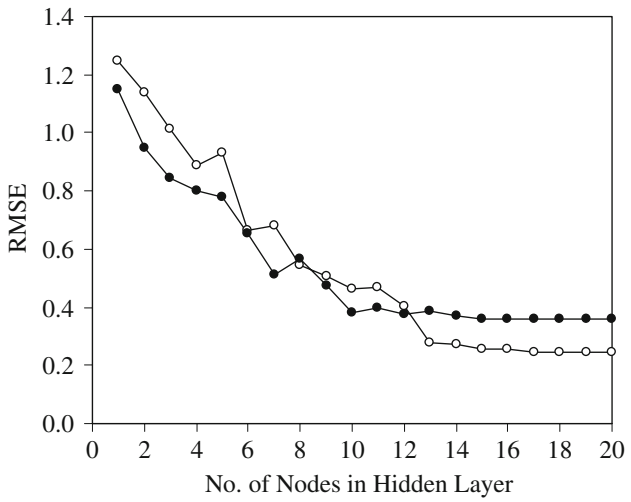

Fig. 2 Plot of RMSE for training (*open circle*) and validation (*filled circle*) sets versus the number of nodes in the hidden layer

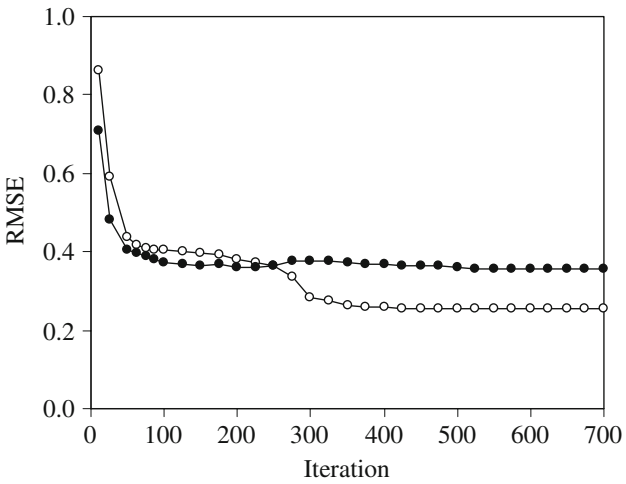

Fig. 3 Plot of RMSE for training (*open circle*) and validation (*filled circle*) sets versus the number of iterations

Table 2 Experimental and calculated values of pK_a for various nitrogen-containing compounds in water for training, validation, and prediction sets by PC-GA-MLR and PC-GA-ANN models along with individual percent deviation (IPD)

No.	Compounds	Exp.	Calculated 1	IPD 1	Calculated 2	IPD 2
	Training set					
2	4-Aminobenzoic acid	2.38	3.20	34.45	2.36	−1.03
3	*p*-Aminosalicylic acid	2.05	2.45	19.51	1.98	−3.61
4	4-Amino-phenol	5.48	5.25	−4.20	5.55	1.34
5	3-Amino-phenol	4.37	4.30	−1.60	4.31	−1.38
6	3,5-Dinitroaniline	0.30	−1.76	−686.67	0.30	−0.63
8	4-Aminobiphenyl	4.35	5.89	35.40	4.34	−0.20
9	3-Methyl-4-nitroaniline	1.64	1.69	3.05	1.64	−0.18
11	3,5-Dimethyl-4-nitrobenzenamine	2.54	2.13	−16.14	2.53	−0.43
16	Methyl *p*-aminobenzoate	2.47	3.59	45.34	2.43	−1.69
18	Propyl-*p*-aminobenzoate	2.49	5.36	115.26	2.50	0.37
19	*p*-Aminobenzoic acid, ethyl ester	2.51	4.44	76.89	2.51	0.02
20	3,4-Dichloroaniline	2.97	2.35	−20.88	2.96	−0.40
21	3-Trifluoromethylaniline	3.49	1.06	−69.63	3.50	0.35
23	3-Iodo-benzenamine	3.61	3.84	6.37	3.61	0.01
24	4-Iodo-benzenamine	3.78	4.45	17.72	3.78	0.00
25	*p*-Bromoaniline	3.86	4.71	22.02	3.88	0.55
27	2-Aminobenzoic acid	2.14	3.03	41.59	2.20	2.91
28	o-Aminophenol	4.84	4.28	−11.57	4.80	−0.91
29	2-Amino-4-nitrophenol	3.10	0.85	−72.58	3.07	−1.09
30	2,4-Dibromoaniline	2.30	2.54	10.43	2.31	0.52
31	3-Nitro-4-toluidine	3.03	2.25	−25.74	3.06	0.83
33	2,6-Dichloro-4-nitroaniline	−2.55	−0.72	−71.76	−2.54	−0.46
34	2,6-Dimethyl-4-nitrobenzenamine	0.98	2.41	145.92	0.96	−1.95
36	2,4-Dinitroaniline	−4.25	−1.55	−63.53	−4.25	0.00
37	4-Chloro-2-nitroaniline	−1.02	0.21	−120.59	−1.04	1.89
39	2,3,4,5,6-Pentafluoroaniline	−0.28	−2.17	675.00	−0.28	0.75
40	2,6-Dichloroaniline	0.42	2.56	509.52	0.46	9.45
43	2,3-Dichloroaniline	1.76	1.96	11.36	1.74	−0.93
46	Methyl anthranilate	2.23	3.69	65.47	2.23	0.03
48	2,3,5,6-Tetramethyl-4-nitrobenzenamine	2.36	2.46	4.24	2.36	0.00
49	2-Methoxy-5-nitroaniline	2.49	1.13	−54.62	2.55	2.55
51	2-Iodoaniline	2.60	4.26	63.85	2.61	0.35
52	2,6-Dinitroaniline	−5.00	−1.90	−62.00	−5.01	0.13
53	2,4,6-Trichloroaniline	−0.03	1.80	−6100.0	−0.01	−57.33
55	2,5-Dimethoxyaniline	3.93	4.51	14.76	3.93	0.06
56	2-Naphthalenamine, 1,2,3,4-tetrahydro-	9.93	8.92	−10.17	9.93	−0.04
58	Benzeneethanamine, a-methyl-, (R)-	10.13	9.29	−8.29	10.19	0.54
59	Propylamine	10.71	10.81	0.93	10.68	−0.32
60	*n*-Butylamine	10.78	10.33	−4.17	10.84	0.58
62	Cyclohexanamine	10.63	10.81	1.69	10.66	0.26
63	Benzylamine	9.33	7.31	−21.65	9.38	0.59
64	Benzenemethanamine, 4-methyl-	9.36	7.62	−18.59	9.31	−0.52
65	*p*-Methoxyamphetamine	9.53	7.86	−17.52	9.78	2.66
69	4-Phenylbutylamine	10.36	9.13	−11.87	10.34	−0.19
70	Isobutylamine	10.68	10.42	−2.43	10.64	−0.39
71	2-Aminomethylfuran	8.89	5.84	−34.31	8.89	−0.04
73	Mescaline	9.56	4.91	−48.64	9.56	0.00

Table 2 continued

No.	Compounds	Exp.	Calculated 1	IPD 1	Calculated 2	IPD 2
74	Allylamine	9.70	8.84	−8.87	9.67	−0.30
75	Sec-butylamine	10.56	10.66	0.95	10.58	0.21
79	Benzeneethanamine, *n*-methyl-	10.08	8.64	−14.29	10.45	3.64
80	Diethylamine	11.09	10.59	−4.51	11.14	0.45
82	Dibutylamine	11.39	11.05	−2.99	11.37	−0.13
86	Benzenemethanamine, *n*-ethyl-	9.64	8.52	−11.62	9.46	−1.90
87	Benzeneethanamine, *n*, b-dimethyl-	9.87	8.98	−9.02	9.87	0.03
90	Dipropylamine	11.00	10.79	−1.91	11.00	0.04
92	Diisopropylamine	11.07	11.22	1.36	10.99	−0.75
93	Piperidine, 2,2,6,6-tetramethyl-	11.72	11.36	−3.07	11.73	0.07
95	Azetidine	11.29	10.49	−7.09	11.22	−0.62
97	Anabasine	8.70	8.06	−7.36	8.67	−0.36
100	2-Propylpiperidine	11.00	11.29	2.64	10.95	−0.49
102	4-Ethylmorpholine	7.67	8.45	10.17	7.60	−0.86
103	Fenpropimorph	6.98	9.39	34.53	6.98	0.00
104	Ethyl dimethylamine	10.16	10.55	3.84	10.18	0.24
105	Dimethylbutylamine	10.19	10.54	3.43	10.21	0.15
106	*n*-Methylmorpholine	7.38	7.67	3.93	7.41	0.47
107	*n*, *n*-Dimethyl-3-pyridylmethylamine	8.00	7.51	−6.13	7.97	−0.39
108	2-Propen-1-amine, *n*, *n*-di-2-propenyl-	8.31	8.84	6.38	8.31	−0.04
109	2-Pyridineethanamine, *n*, *n*-dimethyl-	8.75	8.27	−5.49	8.77	0.27
111	*n*, *n*-Dimethyl-2-(3-pyridyl)ethylamine	8.86	7.66	−13.54	6.60	−25.54
112	*n*, *n*-Dimethylbenzylamine	8.91	8.70	−2.36	8.92	0.08
113	Methadone	8.94	8.02	−10.29	8.93	−0.08
115	3-Pyridylethyl-2-(*n*-pyrrolidine)	9.28	7.78	−16.16	9.28	−0.03
116	*n*-Methylpiperidine	10.08	10.49	4.07	10.07	−0.06
118	Triethylamine	10.78	10.66	−1.11	10.74	−0.37
119	1-Propylpiperidine	10.41	11.33	8.84	10.40	−0.07
120	Piperidine, 1,2,2,6,6-pentamethyl-	11.25	11.69	3.91	11.23	−0.19
122	bis(2-chloroethyl)ethylamine	6.57	7.49	14.00	6.58	0.13
124	Ethanamine, *n*, *n*-dimethyl-2-[5-methyl-2-(1-methylethyl)phenoxy]	8.66	7.24	−16.40	8.66	−0.06
128	Diphenhydramine	8.98	7.18	−20.04	8.97	−0.13
129	Trimethylamine	9.80	11.51	17.45	9.80	0.01
131	Tripropylamine	10.65	10.90	2.35	10.65	−0.03
132	tri *n*-butylamine	10.89	11.57	6.24	10.96	0.61
133	4-Acetylpyridine	3.59	3.50	−2.51	3.58	−0.37
136	Isonicotinic acid, ethyl ester	3.45	3.33	−3.48	3.37	−2.26
138	3-Formylpyridine	3.80	2.41	−36.58	3.70	−2.63
139	4-Chloropyridine	3.84	3.21	−16.41	3.86	0.49
140	4-Formylpyridine	4.77	2.02	−57.65	4.84	1.56
141	4,4′-Dipyridyl	4.82	3.39	−29.67	4.82	0.07
142	Nicotinic acid, methyl ester	3.13	2.91	−7.03	3.31	5.74
143	*i*-Nicotinic acid, methyl ester	3.26	2.55	−21.78	3.15	−3.25
144	3,5-Dichloropyridine	0.67	0.09	−86.57	0.67	0.34
145	Dinicotinic acid	1.10	−0.51	−146.36	1.10	0.32
146	3-Chloropyridine	2.84	1.94	−31.69	2.79	−1.85
150	Nicotinic acid	2.07	2.06	−0.48	2.11	2.14
152	4-Pyridinemethanol	5.33	4.43	−16.89	5.32	−0.12

Table 2 continued

No.	Compounds	Exp.	Calculated 1	IPD 1	Calculated 2	IPD 2
153	3-Pyridinepropanol	5.47	6.49	18.65	5.50	0.51
154	3-Ethylpyridine	5.56	5.99	7.73	5.67	1.92
156	4-Pyridineethanol	5.60	5.65	0.89	5.87	4.79
157	4-Vinylpyridine	5.62	4.46	−20.64	5.50	−2.06
158	3-Methylpyridine	5.63	5.14	−8.70	5.69	1.09
160	4-Ethylpyridine	5.87	6.03	2.73	5.85	−0.31
161	4-Methylpyridine	5.98	4.65	−22.24	5.94	−0.62
164	3,5-Dimethylpyridine	6.15	4.81	−21.79	6.16	0.12
165	3,4-Dimethylpyridine	6.46	5.17	−19.97	6.40	−0.90
167	*n*, *n*-Dimethyl-2-(3-pyridyl)ethylamine	4.30	7.66	78.14	6.60	53.43
168	Anabasine	3.21	8.07	151.40	3.22	0.39
169	4-Cyanopyridine	1.90	2.45	28.95	1.96	2.96
171	Nikethamide	3.50	4.65	32.86	3.51	0.37
172	(E)-3-Nicotinoylacrylic acid	3.82	1.75	−54.19	3.80	−0.48
176	4-Phenylpyridine	5.55	4.56	−17.84	5.56	0.10
177	3-Pyridinemethaneamine	5.96	5.90	−1.01	6.09	2.11
178	3-Nitropyridine	1.18	−0.29	−124.58	1.22	2.99
179	3-Hydroxypyridine	4.80	3.23	−32.71	4.72	−1.72
180	3-Methoxypyridine	4.91	3.93	−19.96	4.86	−1.07
181	5-Ethyl-2-methylpyridine	6.51	6.96	6.91	6.49	−0.26
182	2,3,5,6-Tetramethylpyridine	7.90	6.58	−16.71	7.81	−1.12
185	2-Bromopyridine	0.90	1.59	76.67	0.94	4.63
186	2-Vinylpyridine	4.98	4.77	−4.22	5.15	3.34
187	2-Benzylpyridine	5.13	6.95	35.48	5.13	0.00
191	2-t-Butylpyridine	5.76	7.65	32.81	5.77	0.22
195	2,6-Lutidine	6.60	6.40	−3.03	6.59	−0.19
197	2,4,6-Collidine	7.43	6.29	−15.34	7.57	1.82
199	2,3,4,5,6-Pentachlorpyridine	−1.00	−1.87	87.00	−1.00	0.24
200	2,3-Dichloropyridine	−0.85	−0.91	7.06	−0.84	−0.89
203	2-Methylpyridine	6.00	5.85	−2.50	6.00	0.07
207	2-Methylthiopyridine	3.59	4.92	37.05	3.48	−3.09
208	2-Hydroxypyridine	0.75	3.33	344.00	0.71	−5.48
209	2-Methoxypyridine	3.06	4.06	32.68	3.25	6.36
210	Pyridine, 2,6-dimethoxy-	1.60	4.08	155.00	1.60	−0.23
213	Picolinic acid, methyl ester	2.21	2.81	27.15	2.18	−1.50
214	Picolinic acid	1.06	2.03	91.51	1.16	9.16
216	2-Pyrimidinecarboxylicacid,methylester	−0.68	1.24	−282.35	−0.67	−1.53
217	Pyrimidine,2-(methylthio)	0.59	3.18	438.98	0.62	5.68
220	Pyrimidine, 2-ethoxy-	1.27	3.62	185.04	1.31	3.00
221	4-Methylpyrimidine	1.91	4.03	110.99	1.94	1.57
222	2-SMe-4,6-dimethylpyrimidine	2.12	5.01	136.32	2.15	1.49
223	4,6-Dimethylpyrimidine	2.70	4.98	84.44	2.63	−2.67
226	Fenclorim	4.23	2.06	−51.30	4.30	1.69
229	1-Methyl-4-nitro-1H-imidazole	−0.53	0.92	−273.58	−0.51	−3.04
234	1H-imidazole, 2-phenyl-	6.48	5.23	−19.29	6.48	0.05
235	1H-imidazole, 1-methyl-	6.95	5.98	−13.96	6.97	0.32
238	Anserine	7.04	5.72	−18.75	7.04	−0.07
239	Prochloraz	3.80	4.16	9.47	3.78	−0.51

Table 2 continued

No.	Compounds	Exp.	Calculated 1	IPD 1	Calculated 2	IPD 2
240	Pentostatin	5.20	4.99	−4.04	5.21	0.13
241	1H-imidazole, 1-(phenylmethyl)-	6.70	6.71	0.15	6.64	−0.83
243	Cimetidine	6.80	7.09	4.26	6.80	0.06
246	2-(4-methoxyphenylmethyl)-5-nitrobenzimidazole	4.26	4.27	0.23	4.24	−0.53
247	2-(4-chlorphenylmethyl)-5-chlorobenzimidazole	4.86	3.86	−20.58	4.84	−0.50
248	2-(2-methylphenyl)-5-nitrobenzimidazole	4.87	3.26	−33.06	4.83	−0.92
249	2-(2,4-dimethylphenyl)-5-nitrobenzimidazole	5.29	3.49	−34.03	5.30	0.17
250	2-(4-bromophenylmethyl)-5-chlorobenzimidazole	5.42	7.19	32.66	5.42	0.09
251	2-(4-methylphenyl)-benzimidazole	6.90	5.78	−16.23	6.93	0.44
252	2-(4-methylphenylmethyl)-5-chlorobenzimidazole	7.09	7.55	6.49	7.06	−0.47
253	2-(2-methoxyphenyl)-benzimidazole	7.17	4.88	−31.94	7.15	−0.30
255	Quinoline	4.90	4.79	−2.24	4.89	−0.12
256	5-Quinolinol	5.02	4.09	−18.53	5.15	2.69
258	6-Methoxyquinoline	5.03	4.81	−4.37	5.10	1.38
259	8-Methylquinoline	5.05	5.02	−0.59	5.15	2.05
260	3-Methylquinoline	5.17	5.39	4.26	5.08	−1.67
261	6-Methylquinoline	5.34	5.75	7.68	5.44	1.87
265	4-Methylquinoline	5.67	5.09	−10.23	5.53	−2.55
266	2-Methylquinoline	5.71	6.20	8.58	5.68	−0.54
268	4,7-Dichloro-quinoline	2.80	2.50	−10.71	2.79	−0.27
269	8-Chloroquinoline	3.12	3.22	3.21	3.14	0.59
270	8-Fluoroquinoline	3.34	2.83	−15.27	3.33	−0.35
272	7-Bromoquinoline	3.87	4.15	7.24	3.86	−0.32
274	2,6-Dimethylquinoline	6.10	6.24	2.30	6.05	−0.87
278	8-Quinolinol,5-chloro-	3.56	3.35	−5.90	3.44	−3.47
281	4-Methyl-8-quinolinol	5.56	4.25	−23.56	5.62	1.00
282	Cloquintocet-mexyl	3.75	3.04	−18.93	3.77	0.65
	Validation set					
1	3-Aminobenzoic acid	3.07	3.44	12.05	3.15	2.63
7	3,5-Dichloroaniline	2.51	2.42	−3.59	2.51	−0.13
10	4-Benzoylaniline	2.24	3.74	66.96	2.22	−1.11
12	2-Nitro-*p*-toluidine	0.40	2.16	440.00	0.40	0.82
15	*p*-Trifluoromethylaniline	2.45	1.22	−50.20	2.45	−0.07
32	2-Aminobiphenyl	3.83	5.44	42.04	3.90	1.72
38	2-Chloro-4-nitroaniline	−0.94	0.33	−135.11	−0.98	4.55
42	4-Nitro-2-toluidine	1.04	2.16	107.69	1.02	−1.52
44	2,4-Dichloroaniline	2.00	3.07	53.50	1.99	−0.54
61	Ethylamine	10.87	11.37	4.60	10.87	0.04
66	3,4-Methylenedioxyamphetamine	9.67	6.91	−28.54	9.69	0.17
76	Methylamine	10.62	9.83	−7.44	10.62	0.00
78	t-Butylamine	10.68	12.12	13.48	10.65	−0.30
81	Piperidine	11.28	10.15	−10.02	11.26	−0.14
88	Methamphetamine	9.87	9.48	−3.95	9.63	−2.40
91	Hexamethyleneimine	11.07	10.46	−5.51	11.05	−0.14
98	Pyrrolidine, 2-phenyl-	9.40	9.22	−1.91	9.44	0.47
114	Chlorpheniramine	9.13	7.49	−17.96	9.11	−0.20
117	1-Methyl-pyrrolidine	10.32	10.48	1.55	10.37	0.45
121	Nicotine	8.18	8.40	2.69	5.70	−30.33

Table 2 continued

No.	Compounds	Exp.	Calculated 1	IPD 1	Calculated 2	IPD 2
123	Arecoline	7.16	6.60	−7.82	7.07	−1.31
125	Moxisylyte	8.72	5.21	−40.25	8.72	0.00
126	Piperalin	8.90	6.74	−24.27	8.91	0.09
130	Fenpropidin	10.10	10.95	8.42	10.10	0.01
134	3-Bromopyridine	2.91	2.20	−24.40	2.91	−0.04
135	Nicotinic acid, ethyl ester	3.35	3.44	2.69	3.34	−0.32
137	4-Bromopyridine	3.78	3.68	−2.65	3.76	−0.64
148	3-Acetylpyridine	3.18	3.78	18.87	3.14	−1.29
151	3-Pyridinemethanol	4.90	4.66	−4.90	4.87	−0.68
155	4-Benzylpyridine	5.59	6.42	14.85	5.59	−0.08
159	4-Pyridinepropanol	5.84	6.62	13.36	5.72	−2.06
162	4-(tert-butyl)pyridine	5.99	7.33	22.37	6.03	0.70
174	Isonicotinic acid	1.70	1.69	−0.59	1.88	10.46
175	Pyridine	5.23	5.00	−4.40	5.20	−0.56
183	2-Fluoropyridine	−0.44	1.39	−415.91	−0.48	8.25
190	2-Methyl-5-vinylpyridine	5.67	5.89	3.88	5.67	−0.03
193	2,5-Dimethylpyridine	6.40	6.15	−3.91	6.36	−0.55
198	2,6-Dichloropyridine	−2.86	−0.24	−91.61	−2.89	0.98
202	2-Phenylpyridine	4.48	4.91	9.60	4.54	1.37
218	Pyrimidine,5-nitro	0.72	−1.32	−283.33	0.72	0.33
219	Pyrimidine	1.23	3.77	206.50	1.23	0.37
225	Pyrimethanil	3.52	4.57	29.83	4.16	18.08
227	2-Amino-4,6-dimethylpyrimidine	4.82	4.57	−5.19	4.16	−13.77
231	1-Methyl-5-nitroimidazole	2.13	0.74	−65.26	2.15	1.12
232	1H-imidazole, 1-acetyl-	3.60	3.99	10.83	3.63	0.95
236	Imidazole	6.95	6.57	−5.47	7.00	0.74
242	Pilocarpol	6.78	7.77	14.60	6.77	−0.21
244	Imazalil base	6.53	6.01	−7.96	6.53	0.00
245	Benzimidazole	5.53	4.66	−15.73	5.53	0.01
254	2-(4-aminophenylmethyl)-5-chlorobenzimidazole	7.47	7.28	−2.54	7.43	−0.57
262	7-Methylquinoline	5.34	5.50	3.00	5.38	0.76
267	3-Bromoquinoline	2.69	3.87	43.87	2.69	0.05
273	6-Bromoquinoline	3.87	3.99	3.10	3.87	−0.03
275	2,4-Dimethylquinoline	5.12	6.18	20.70	5.13	0.27
276	3-Quinolinol	4.28	4.07	−4.91	4.31	0.67
279	8-Methoxyquinoline	5.01	4.57	−8.78	4.99	−0.47
	Prediction set					
13	4-Methylsulfonylaniline	1.35	2.17	60.74	1.35	0.24
14	4-Chloro-3-nitro-benzenamine	1.90	0.26	−86.32	1.89	−0.67
17	Butyl-*p*-aminobenzoate	2.47	5.77	133.60	2.51	1.70
22	*m*-Bromoaniline	3.58	3.87	8.10	3.56	−0.54
26	3-Methyl-4-bromoaniline	4.05	4.85	19.75	4.08	0.65
35	2,5-Dichloroaniline	2.05	2.78	35.61	2.03	−0.76
41	4-Methoxy-2-nitro-benzenamine	0.77	1.81	135.06	0.69	−10.56
45	*o*-Aminobenzoic acid, ethyl ester	2.18	4.31	97.71	2.28	4.39
47	5-Nitro-2-toluidine	2.35	2.37	0.85	2.40	2.11
50	*o*-Bromoaniline	2.53	4.06	60.47	2.54	0.26
54	2,4,5-Trichloroaniline	1.09	1.57	44.04	1.11	1.83

Table 2 continued

No.	Compounds	Exp.	Calculated 1	IPD 1	Calculated 2	IPD 2
57	2-Phenylethylamine	9.96	8.24	−17.27	9.65	−3.12
67	Benzenepropanamine, a-methyl-	9.79	9.43	−3.68	9.80	0.06
68	3-Phenyl propylamine	10.16	8.74	−13.98	10.12	−0.43
72	2,2,2-Trifluoroethylamine	5.70	2.52	−55.79	5.70	0.00
77	Isopropylamine	10.63	11.75	10.54	10.61	−0.22
83	Morpholine	8.49	7.72	−9.07	8.53	0.47
84	Diallylamine	9.29	8.87	−4.52	9.34	0.52
85	*n*-Methylbenzylamine	9.54	7.92	−16.98	9.54	0.03
89	Methylbutylamine	10.90	10.43	−4.31	10.92	0.18
94	2-Methylpiperidine	11.08	10.67	−3.70	11.09	0.05
96	Pyrrolidine	11.31	11.23	−0.71	11.41	0.86
99	Dimethylamine	10.73	11.15	3.91	10.73	0.00
101	Dicyclohexylamine	10.40	11.68	12.31	10.44	0.40
110	3-Pyridylethyl-2-*n*-piperidine	8.81	7.79	−11.58	8.81	−0.05
127	Orphenadrine	8.91	7.33	−17.73	8.90	−0.10
147	3-Fluoropyridine	2.97	1.53	−48.48	2.96	−0.37
149	3-Iodopyridine	3.25	2.56	−21.23	3.25	−0.04
163	4-Propylpyridine	6.05	6.96	15.04	6.02	−0.44
166	4-Methoxypyridine	6.47	4.06	−37.25	6.53	0.92
170	Nicotine	3.10	8.40	170.97	5.70	83.84
173	3-Phenylpyridine	4.80	4.73	−1.46	4.72	−1.59
184	2-Chloropyridine	0.49	1.72	251.02	0.44	−9.57
188	2-Pyridineethanol	5.31	6.46	21.66	5.33	0.46
189	2-Pyridinepropanol	5.61	7.13	27.09	5.61	−0.06
192	2-Ethylpyridine	5.89	6.65	12.90	5.79	−1.73
194	2,3-Dimethylpyridine	6.57	5.81	−11.57	6.71	2.08
196	2,4-Dimethylpyridine	6.99	5.69	−18.60	6.88	−1.63
201	2,3,5,6-Tetrachlorpyridine	−0.80	−1.17	46.25	−0.79	−0.93
204	2,2-Bipyridine	4.33	3.61	−16.63	4.32	−0.35
205	2-Pyridinemethanol	4.86	4.93	1.44	4.79	−1.36
206	2-Amino-5-methylpyridine	7.22	5.45	−24.52	7.16	−0.77
211	2-Acetylpyridine	2.73	3.81	39.56	2.57	−5.77
212	2-Pyridinecarboxyaldehyde	3.80	2.54	−33.16	3.57	−6.11
215	Pyrimidine, 2-bromo-	−1.63	−1.60	−1.84	−1.63	0.01
224	2-Aminopyrimidine	3.45	3.09	−10.43	3.45	−0.03
228	2,4,6-Pyrimidinetriamine	6.81	3.12	−54.19	6.81	0.04
230	4-Nitroimidazole	−0.05	−0.93	1760.00	−0.07	47.00
233	Triflumizole	3.70	2.35	−36.49	3.70	0.11
237	1H-imidazole, 2-methyl-	7.85	6.24	−20.51	7.82	−0.44
257	7-Methoxyquinoline	5.03	4.99	−0.80	5.03	0.05
263	6-Hydroxyquinoline	5.15	4.30	−16.50	5.25	2.01
264	7-Quinolinol	5.46	4.36	−20.15	5.46	0.04
271	6-Chloroquinoline	3.85	3.79	−1.56	3.77	−1.97
277	8-Quinolinol	4.90	3.60	−26.53	4.76	−2.81
280	2-Methyl 8-quinolonol	5.55	5.70	2.70	5.55	0.01

1 and 2 denote to the values obtained by PC-GA-MLR and PC-GA-ANN models

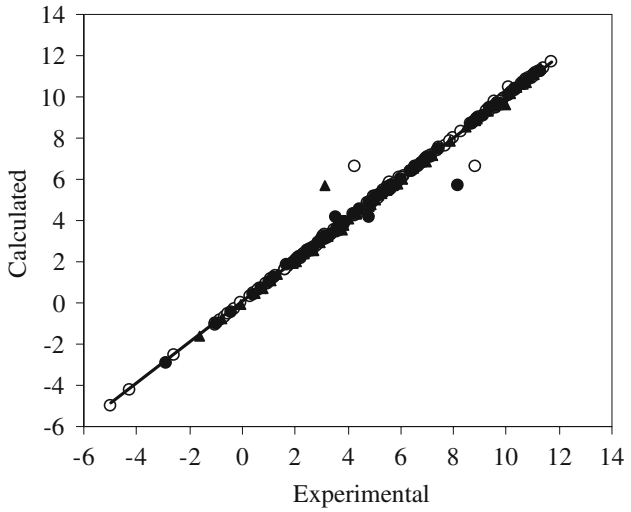

Fig. 4 Plot of the calculated values of pK_a from the PC-GA-ANN model versus the experimental values of it for training (*open circle*), validation (*filled circle*), and prediction (*filled triangle*) sets

$$\mathrm{p}K_a(\mathrm{cal}) = 0.9994\mathrm{p}K_a(\mathrm{exp}) + 0.0025 \quad (2)$$

$$\left(R^2 = 0.9994;\quad \mathrm{MPD} = 1.591;\quad \mathrm{RMSE} = 0.0858;\quad F = 440008.7\right)$$

Similarly, the correlation of pK_a (cal) versus pK_a (exp) values in the prediction set gives Eq. 3:

$$\mathrm{p}K_a(\mathrm{cal}) = 1.0007\ \mathrm{p}K_a(\mathrm{exp}) - 0.0192 \quad (3)$$

$$\left(R^2 = 0.9995;\quad \mathrm{MPD} = 2.123;\quad \mathrm{RMSE} = 0.0750;\quad F = 108858.9\right)$$

Table 3 compares the results obtained using the PC-GA-MLR and PC-GA-ANN models. The MPD and RMSE of the models for total, training, validation, and prediction sets show the potential of the ANN model for prediction of pK_a values of the various nitrogen-containing compounds in water.

As a result, it was found that a properly selected and trained neural network could fairly represent dependence of the acidity constant of the various nitrogen-containing compounds in water on the PCs. Then the optimized neural network could simulate the complicated non-linear relationship between pK_a values and the PCs. The MPD and RMSE of 64.62 and 1.4863 for the prediction set by the PC-GA-MLR model should be compared with the values of 2.123 and 0.0750, respectively, for the PC-GA-ANN model. It can be seen from Table 3 that although parameters appearing in the PC-GA-MLR model are used as inputs for the generated PC-GA-ANN model, the statistics show a large improvement. These improvements are due to the fact that pK_a values of various nitrogen-containing compounds show non-linear correlations with the PCs. It is clear that statistical parameters of the proposed single model for the prediction of the acidity constant is very much better than the separate models [47] that have been recently proposed (RMSE of the models are 0.0858 and 0.9183).

Conclusions

QSAR modelings have been applied on the acidity constant of various nitrogen-containing compounds in water using the PC-GA-MLR and PC-GA-ANN methods. Comparison of the values of MPD (and other statistical parameters in Table 3) for training, validation, and prediction sets for the PC-GA-MLR and PC-GA-ANN models demonstrates the superiority of the PC-GA-ANN model over the PC-GA-MLR model. Mean percent deviation of 64.62 for the prediction set by the PC-GA-MLR model should be compared with the value of 2.123 for the PC-GA-ANN model. Since the improvement of the results obtained using the non-linear model (PC-GA-ANN) is considerable, it can be concluded that the non-linear characteristics of the PCs on the pK_a values of the compounds in water is serious.

Data and methodology

Acidity constant and theoretical descriptors

The compounds in data sets include: 55 anilines, 77 amines, 82 pyridines, 14 pyrimidines, 26 imidazoles and benzimidazoles, and 28 quinolines. Acidity constants of the compounds were taken from the recently published paper [47]. The molecular models were constructed with

Table 3 Comparison of statistical parameters obtained by the PC-GA-MLR and PC-GA-ANN models for pK_a values of various nitrogen-containing compounds

Model	$\mathrm{RMSE}_{\mathrm{tot}}$	$\mathrm{RMSE}_{\mathrm{train}}$	$\mathrm{RMSE}_{\mathrm{valid}}$	$\mathrm{RMSE}_{\mathrm{pred}}$	$\mathrm{MPD}_{\mathrm{tot}}$	$\mathrm{MPD}_{\mathrm{train}}$	$\mathrm{MPD}_{\mathrm{valid}}$	$\mathrm{MPD}_{\mathrm{pred}}$
PC-GA-MLR	1.3614	1.3693	1.1970	1.4863	70.53	81.27	43.84	64.62
PC-GA-ANN	0.0858	0.0670	0.1336	0.0750	1.591	1.438	1.522	2.123

Subscripts tot, train, valid, and pred refer to the total, training, validation, and prediction sets. RMSE and MPD are root-mean square error and mean percent deviation

HyperChem 7.0, and molecular structures were optimized using the AM1 algorithm [49]. In order to calculate the theoretical descriptors, the Dragon package version 2.1 was used [50]. For this propose, the output of the HyperChem software for each compound fed into the Dragon program and the descriptors were calculated. As a result, a total of 1,481 theoretical descriptors were calculated for each compound in data sets (282 compounds).

Data pretreatment

The theoretical descriptors were reduced by the following procedure:

1. Descriptors that are constant have been eliminated (376 descriptors).
2. In addition, to decrease the redundancy existing in the descriptors data matrix, the correlation of descriptors with each other and with pK_a of the compounds are examined, and collinear descriptors ($R > 0.9$) are detected. Among the collinear descriptors, the one that has the highest correlation with pK_a values is retained, and the others are removed from the data matrix (699 descriptors).
3. Before statistical analysis, the descriptors are scaled to zero mean and unit variance (autoscaling procedure). The data matrix (406 descriptors) is subjected to PCA using Matlab software package [51]. Multiparameter linear regression was obtained using SPSS software [52].

Genetic algorithm

Nowadays, GA is well known as an interesting and more widely used variable selection method. The distinctive aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores different regions in parameter space [53]. To select the most relevant PCs, evolution of population was simulated [30, 54–57]. Each individual of the population defined by a chromosome of binary values represented a subset of PCs. The number of genes at each chromosome was equal to the number of PCs. The population of the first generation was selected randomly. A gene took a value of 1 if its corresponding PC was included in the subset; otherwise, it took a value of zero. The number of genes with a value of 1 was kept relatively low to have a small subset of PCs [57], that is, the probability of generating 0 for a gene was set greater (at least 60%) than the value of 1. The operators used here were crossover and mutation. In the crossover procedure, new chromosomes were generated from a pair of randomly selected chromosomes. Many methods have been proposed for the crossover technique; here the uniform crossover technique was applied to ten pairs of chromosomes in each iteration of generation. In the mutation procedure the binary bit pattern in each chromosome was changed with a small probability. The probability of the application of these operators was varied linearly with generation renewal (0–0.1% for mutation and 60–90% for crossover). The population size was varied between 50 and 250 for different GA runs. For a typical run, the evolution of the generation was stopped when 90% of the generations took the same fitness [27]. The GA program was written in Matlab 6.5. [58].

Artificial neural network

A feed forward ANN with a back-propagation of error algorithm was used to process the non-linear relationship between the selected PCs and the acidity constant. The number of input nodes in the ANN was equal to the number of PCs. The ANN models confined to a single hidden layer, because the network with more than one hidden layer would be harder to train. A three-layer network with a sigmoidal transfer function was designed. The initial weights were randomly selected between 0 and 1. The optimization of the weights and biases was carried out according to Levenberg–Marquardt algorithms for BP of error, which, although requiring far more extensive computer memory, is significantly faster than other algorithms based on gradient descent [59]. The data set was randomly divided into three groups: a training set, a validation set, and a prediction set consisting of 170, 56, and 56 molecules. The training and validation sets were used for the model generation, and the prediction set was used for evaluation of the generated model. The performances of training, validation, and prediction of models are evaluated by the mean percentage deviation (MPD) and root mean square error (RMSE), which are defined as follows:

$$\mathrm{MPD} = \sum_{i=1}^{N} \left| \frac{\left(P_i^{\mathrm{exp}} - P_i^{\mathrm{cal}}\right)}{P_i^{\mathrm{exp}}} \right| \frac{100}{N} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{\left(P_i^{\mathrm{exp}} - P_i^{\mathrm{cal}}\right)^2}{N}} \tag{5}$$

where P_i^{exp} and P_i^{cal} are experimental and calculated values of pK_a with the models, and N denotes the number of data points. Individual percent deviation (IPD) is defined as follows:

$$\mathrm{IPD} = 100 \times \left(\frac{P_i^{\mathrm{cal}} - P_i^{\mathrm{exp}}}{P_i^{\mathrm{exp}}} \right) \tag{6}$$

The processing of the data was carried using Matlab 6.5. The neural networks were implemented using Neural Network Toolbox version 4.0 for Matlab [59].

Acknowledgments The authors wish to acknowledge the vice-presidency of research, University of Mohaghegh Ardabili, for financial support of this work.

References

1. Zhao YH, Yuan X, Yuan LH, Wang LS (1996) Bull Environ Contam Toxicol 57:242
2. Alines P (1996) J Planar Chromatogr Mod TLC 9:52
3. Jover J, Bosque R, Sales J (2007) QSAR Comb Sci 26:385
4. Yao XJ, Wang YW, Zhang XY, Zhang RS, Liu MC, Hu ZD, Fan BT (2002) Chemom Intell Lab Syst 62:217
5. Guha R, Serra JR, Jurs PC (2004) J Mol Graph Model 23:1
6. Krogsgaard-Larsen P, Liljefors T, Madsen U (2002) Textbook of drug design and discovery. Taylor & Francis, London
7. Consonni V, Todeschini R, Pavan M, Gramatica P (2002) J Chem Inf Comput Sci 42:693
8. Karthikeyan M, Glen RC, Bender A (2005) J Chem Inf Model 45:581
9. Melnikov AA, Palyulin VA, Zefirov NS (2007) J Chem Inf Model 47:2077
10. Ajmani S, Rogers SC, Barley MH, Livingstone DJ (2006) J Chem Inf Model 46:2043
11. Katritzky AR, Stoyanova-Slavova IB, Dobchev DA, Karelson M (2007) J Mol Graph Model 26:529
12. Shamsipur M, Siroueinejad A, Hemmateenejad B, Abbaspour A, Sharghi H, Alizadeh K, Arshadi S (2007) J Electranal Chem 600:345
13. Habibi-Yangjeh A, Pourbasheer E, Danandeh-Jenagharad M (2008) Monatsh Chem doi:10.1007/s00706-008-0951-z
14. Avram S, Berner H, Milac AL, Wolschann P (2008) Monatsh Chem 139:407
15. Prakasvudhisarn C, Lawtrakul L (2008) Monatsh Chem 139:197
16. Lawtrakul L, Prakasvudhisarn C (2005) Monatsh Chem 136:1681
17. Todeschini, V. Consonni (2000) Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany
18. Sutter JM, Kalivas JH, Lang PM (1992) J Chemometr 6:217
19. Vendrame R, Braga RS, Takahata Y, Galvao DS (1999) J Chem Inf Comput Sci 39:1094
20. Malinowski ER (2002) Factor analysis in chemistry. Wiley, New York
21. Katritzky AR, Tulp I, Fara DC, Lauria A, Maran U, Acree WE (2005) J Chem Inf Model 45:913
22. Hemmateenejad B, Akhond M, Miri R, Shamsipur M (2003) J Chem Inf Comput Sci 43:1328
23. Hemmateenejad B, Shamsipur M (2004) Internet Electron J Mol Des 3:316
24. Jalali-Heravi M, Kyani A (2004) J Chem Inf Comput Sci 44:1328
25. Hemmateenejad B, Safarpour MA, Miri R, Nesari N (2005) J Chem Inf Model 45:190
26. Hemmateenejad B, Safarpour MA, Miri R, Taghavi F (2004) J Comput Chem 25:1495
27. Depczynski U, Frost VJ, Molt K (2000) Anal Chim Acta 420:217
28. Hemmateenejad B (2005) Chemom Intell Lab Syst 75:231
29. Goldberg DE (2000) Genetic algorithm in search, optimization and machine learning. Addison-Wesley-Longman, Reading
30. Cho SJ, Hermsmeier MA (2002) J Chem Inf Comput Sci 42:927
31. Despagne F, Massart DL (1998) Analyst 123:157
32. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design. Wiley-VCH, Germany
33. Meiler J, Meusinger R, Will M (2000) J Chem Inf Comput Sci 40:1169
34. Habibi-Yangjeh A, Nooshyar M (2005) Phys Chem Liq 43:239
35. Habibi-Yangjeh A, Nooshyar M (2005) Bull Korean Chem Soc 26:139
36. Habibi-Yangjeh A, Danandeh-Jenagharad M, Nooshyar M (2005) Bull Korean Chem Soc 26:2007
37. Habibi-Yangjeh A (2007) Phys Chem Liq 45:471
38. Tabaraki R, Khayamian T, Ensafi AA (2006) J Mol Graph Model 25:46
39. Habibi-Yangjeh A, Danandeh-Jenagharad M, Nooshyar M (2006) J Mol Model 12:338
40. Habibi-Yangjeh A, Esmailian M (2007) Bull Korean Chem Soc 28:1477
41. Habibi-Yangjeh A, Pourbasheer E, Danandeh-Jenagharad M (2008) Bull Korean Chem Soc 29:833
42. Habibi-Yangjeh A, Esmailian M (2008) Chin J Chem 26:875
43. Schuurmann G (1996) Quant Struct Act Relat 15:121
44. Citra MJ (1999) Chemosphere 38:191
45. Liptak MD, Gross KC, Seybold PG, Feldgus S, Shields GC (2002) J Am Chem Soc 124:6421
46. Ma Y, Gross KC, Hollingsworth CA, Seybold PG, Murray JS (2004) J Mol Model 10:235
47. Tehan BG, Lloyd EJ, Wong MG, Pitt WR, Gancia E, Manallack DT (2002) Quant Struct Act Relat 21:473
48. Saiz-Urra L, Perez Gonzalez MP, Teijeira M (2006) Bioorg Med Chem 14:7347
49. HyperChem Release 7, HyperCube, Inc., http://www.hyper.com
50. Todeschini R, Milano Chemometrics and QSPR Group, http://www.disat.unimib.it/chm
51. Matlab 6.5. Mathworks, 1984–2002
52. SPSS for Windows, Statistical Package for IBM PC, SPSS Inc., http://www.spss.com
53. Cartwright HM (1993) Applications of artificial intelligence in chemistry. Oxford University Press, Oxford
54. Baumann K, Albert H, Von Korff M (2002) J Chemometr 16:339
55. Lu Q, Shen G, Yu R (2002) J Comput Chem 23:1357
56. Ahmad S, Gromiha MM (2003) J Comput Chem 24:1313
57. Deeb O, Hemmateenejad B, Jaber A, Garduno-Juarez R, Miri R (2007) Chemosphere 67:2122
58. The Mathworks Inc (2002) Genetic algorithm and direct search toolbox users guide, Massachusetts
59. The Mathworks Inc (2002) Neural network toolbox users guide, Massachusetts