Monatsh Chem 139, 1423–1431 (2008)
DOI 10.1007/s00706-008-0951-z
Printed in The Netherlands

Monatshefte für Chemie
Chemical Monthly

Prediction of basicity constants of various pyridines in aqueous solution using a principal component-genetic algorithm-artificial neural network

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

Department of Chemistry, Faculty of Science, University of Mohaghegh Ardabili, Ardabil, Iran

Received 7 April 2008; Accepted 2 May 2008; Published online 19 June 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 of the basicity constant (pK_b) for various pyridines (91 compounds) dissolved in water at 25°C. A large number of theoretical descriptors were calculated for each compound. The first 54 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 eight principal components as variables. For evaluation of the predictive power of the models, pK_b values of 18 compounds in the prediction set were calculated. Mean percentage deviation (MPD) for PC-GA-MLR and PC-GA-ANN models are 21.096 and 3.541. Comparison of the results obtained by the models reveals superiority of the PC-GA-ANN model relative to the PC-GA-MLR model. The improvements are due to the fact that the pK_b of the pyridines demonstrate non-linear correlations with the principal components.

Keywords QSPR; pK_b; Pyridines; Genetic algorithm; Artificial neural network.

Introduction

The quantitative structure-property/activity relationship (QSPR/QSAR) models now correlate chemical structure to a wide variety of physical, chemical, biological (including biomedical, toxicological, ecotoxicological), and technological properties [1–9]. QSPR/QSAR models are essentially calibration models in which the independent variables are molecular descriptors that describe the structure of molecules and the dependent variable is the property/activity of interest. Since these theoretical descriptors are determined solely from computational methods, *a priori* predictions of the properties/activities of compounds are possible, no laboratory measurements are needed thus saving time, space, materials, equipment, and alleviating safety (toxicity) and disposal concerns. To obtain a significant correlation, it is crucial that appropriate descriptors be employed [6]. A wide variety of molecular descriptors has been reported for using in QSPR/QSAR models [10]. 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) [11–14]. The PCA was used to compress the

Correspondence: Eslam Pourbasheer, Department of Chemistry, Faculty of Science, University of Mohaghegh Ardabili, Ardabil, Iran. E-mail: ehsan@khayam.ut.ac.ir

descriptor groups into principal components (PCs). In order to reduce the dimensionality of the independent variable space, a limited number of PCs are used [15–18]. 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 [15–21]. 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 PCs with highest eigenvalue is considered as the most significant one and, 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 [18]. 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 [15, 16]. Better results are often achieved by this method. Recently, a genetic algorithm (GA) has been applied for the selection of the most relevant PCs instead of the older methods [19, 20]. Comparison of the results obtained using GA principal component selection with the two abovementioned methods shows that GA gives a better result and close to the correlation ranking [19–21]. GA is a stochastic method to solve optimization problems applying evolution hypothesis of Darwin and different genetic functions, *i.e.*, crossover and mutation [22, 23]. A genetic algorithm 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 [22].

Artificial neural networks (ANNs) have become popular in QSPR/QSAR models due to their success where complex non-linear relationships exist amongst data [24, 25]. An ANN is formed from 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 [26–34].

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. Experimentally determined pK_b 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 [35].

In the present work, 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 principal components and aqueous basicity (pK_b) of 91 various pyridines with diverse chemical structures at 25°C and the results were compared with each other and the experimental values. To the best of our knowledge, there is no report on prediction of basicity constant for organic bases using these nonlinear methods.

Results and discussion

Principal component analysis

After the elimination of the constant and one of the collinear ones, 303 descriptors remained from 1481 theoretical descriptors calculated for the compounds. The results of application of PCA on the descriptors data matrix are shown in Table 1 for the first 54 PCs. The logarithm of eigen-value (log EV), the percent of variance, which can be explained by each PC (*%V*), and the cumulative percent of variances (*C%V*) are included in this table. As is shown, 99.9% of the variances in the descriptors data matrix are explained by 54 first PCs. Therefore, we focused our analysis on these PCs, and the reminders, which are noisy factors, were not considered.

Table 1 The results for application of PCA on the descriptors data matrix

No.	log EV	%*V*	*C*%*V*	No.	log EV	%*V*	*C*%*V*
1	1.759	8.338	8.338	28	0.928	1.232	82.745
2	1.676	6.894	15.231	29	0.908	1.177	83.921
3	1.641	6.365	21.596	30	0.889	1.126	85.047
4	1.524	4.860	26.456	31	0.853	1.037	86.084
5	1.484	4.436	30.892	32	0.831	0.985	87.070
6	1.457	4.167	35.059	33	0.821	0.963	88.033
7	1.422	3.845	38.904	34	0.786	0.887	88.920
8	1.376	3.457	42.361	35	0.756	0.829	89.749
9	1.358	3.312	45.673	36	0.743	0.804	90.554
10	1.338	3.170	48.842	37	0.727	0.776	91.330
11	1.327	3.088	51.930	38	0.720	0.763	92.093
12	1.280	2.772	54.702	39	0.694	0.719	92.813
13	1.224	2.435	57.137	40	0.675	0.689	93.501
14	1.210	2.358	59.495	41	0.659	0.663	94.164
15	1.185	2.228	61.724	42	0.626	0.614	94.778
16	1.168	2.140	63.864	43	0.615	0.599	95.377
17	1.139	2.002	65.866	44	0.593	0.569	95.947
18	1.121	1.920	67.786	45	0.567	0.536	96.483
19	1.101	1.833	69.619	46	0.499	0.459	96.942
20	1.084	1.763	71.382	47	0.487	0.446	97.388
21	1.045	1.612	72.994	48	0.482	0.441	97.830
22	1.027	1.547	74.541	49	0.456	0.416	98.245
23	1.006	1.475	76.017	50	0.451	0.411	98.656
24	0.991	1.423	77.440	51	0.379	0.348	99.004
25	0.982	1.396	78.835	52	0.357	0.331	99.335
26	0.973	1.366	80.202	53	0.314	0.300	99.635
27	0.955	1.311	81.513	54	0.268	0.269	99.904

Principal component-genetic algorithm-multiparameter linear regression

Obtaining the number of significant principal components is the main problem in the PCA-based methods. The first 54 principal components (PCs) were found to explain more than 99.9% of variances in the original data matrix. As noted previously, not all of the PCs is informative for QSAR/QSPR modeling [18–21]. Then, we used GA for the selection of the most relevant PCs instead of the older methods. The selected PCs are PC3, PC4, PC6, PC7, PC10, PC11, PC12, and PC35. As can be seen the selected principal components are not based on their eigen value. For example, the third and forth PCs are selected and the first and second ones 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. Multiparameter linear correlation of pK_b values for 55 various pyridines in the training set was obtained using eight principal components selected by GA. The calculated values of pK_b for the compounds in training, validation, and prediction sets using the PC-GA-MLR model

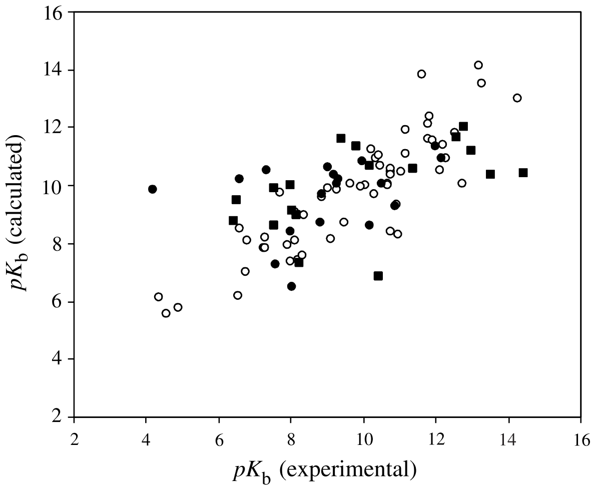

Fig. 1 Plot of the calculated values of pK_b from the PC-GA-MLR model *versus* the experimental values of it for training (○), validation (■), and prediction (●) sets

have been plotted versus the experimental values of it (Fig. 1).

Principal component-genetic algorithm-artificial neural network

To process the non-linear relationships exists between the basicity and the PCs, the ANN modeling method combined with PCA for dimension reduction and GA for feature selection was employed. A principal component-genetic algorithm-artificial neural network (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 [19–21]. The input vectors were the set of PCs, which were selected by GA, and therefore, the number of nodes in the input layer was dependent on the number of selected PCs. In the PC-GA-MLR model it is assumed that the PCs are independent of each other and truly additive relevant to the property under study. ANNs are particularly well-suited for QSAR/QSPR models because of their ability to extract non-linear information present in the data matrix. For this reason the next step in this work was generation of the ANN model. 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 [28–31]. Before training the network, the number of nodes in the hidden layer was optimized. The number of nodes in the hidden layer was optimized by several training sessions with different numbers of hidden nodes (from one to eighteen). 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. A plot of *RMSET* and *RMSEV versus* the number of nodes in the hidden layer is shown in Fig. 2. It is clear that the twelve nodes in the hidden layer is the optimum value.

This network consists of eight inputs (including PC3, PC4, PC6, PC7, PC10, PC11, PC12, and PC35), the same PCs in the PC-GA-MLR model, and one output for pK_b. Then an ANN with architecture 8-12-1 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 loose its prediction power [25]. 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 was not seen 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_b values

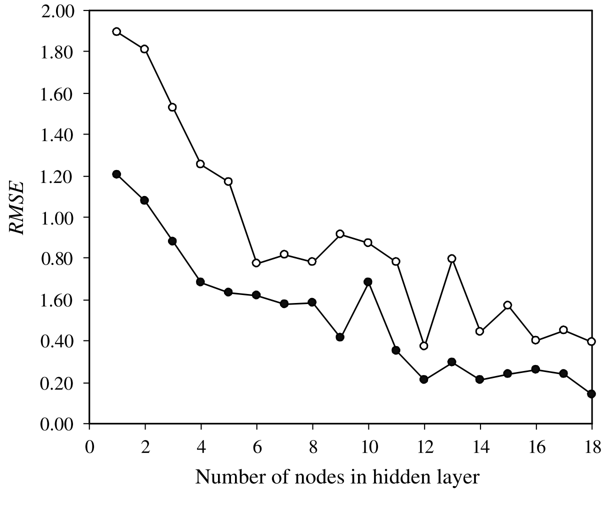

Fig. 2 Plot of *RMSE* for training (●) and validation (○) sets *versus* the number of nodes in hidden layer

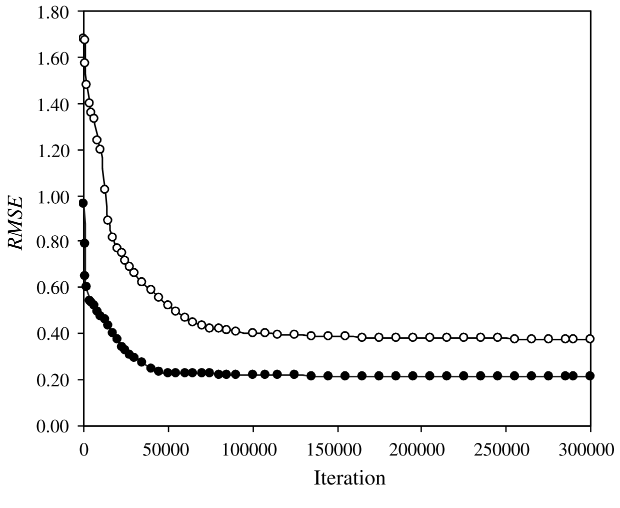

Fig. 3 Plot of *RMSE* for training (●) and validation (○) sets *versus* the number of iterations

Table 2 Experimental and calculated values of pK_b for various pyridines in water at 25°C for training, validation and prediction sets by PC-GA-MLR and PC-GA-ANN models along with individual percent deviation (*IPD*)[a]

No.	Compound	Exp.	Calculated 1	*IPD*1	Calculated 2	*IPD*2
training						
1	3-Acetamidopyridine	9.63	10.042	−4.278	9.607	−0.235
2	3-Acetylpyridine	10.74	10.490	2.366	10.866	1.132
3	4-Amino-3-bromomethylpyridine	6.53	6.195	5.127	6.560	0.464
4	2-Amino-3-methylpyridine	6.76	6.998	−3.520	6.602	−2.339
5	4-Amino-3-methylpyridine	4.57	5.544	−21.320	4.444	−2.761
6	2-Amino-5-methylpyridine	6.78	8.121	−19.779	6.993	3.139
7	2-Aminopyridine	7.29	8.186	−12.294	7.402	1.532
8	4-Aminopyridine	4.89	5.776	−18.214	4.942	1.142
9	2-Bromopyridine	13.29	13.508	−1.640	13.309	0.141
10	3-Bromopyridine	11.15	11.918	−6.886	11.055	−0.850
11	4-Bromopyridine	10.29	9.690	5.828	10.251	−0.383
12	3-tert-Butylpyridine	8.18	7.435	9.112	8.078	−1.249
13	3-Carbamoylpyridine	10.67	10.054	5.791	10.701	0.272
14	3-Chloropyridine	11.16	11.072	0.784	11.385	2.018
15	2-Cyanopyridine	14.26	13.005	8.803	14.196	−0.450
16	3-Cyanopyridine	12.55	11.828	5.751	12.891	2.720
17	4-Cyanopyridine	12.10	10.549	12.814	12.001	−0.815
18	2,4-Dimethylpyridine	7.26	7.824	−7.774	7.187	−1.010
19	2,6-Dimethylpyridine	7.29	7.861	−7.836	7.131	−2.177
20	3,5-Dimethylpyridine	7.91	7.934	−0.308	7.934	0.300
21	3-Ethyl-2-hydroxypyridine	9.00	9.893	−9.919	8.995	−0.059
22	2-Ethylpyridine	8.11	9.090	−12.083	7.990	−1.480
23	4-Ethylpyridine	8.13	8.110	0.250	8.191	0.744
24	3-Fluoropyridine	11.03	10.451	5.252	10.674	−3.224
25	2-Hydroxy-4-methylpyridine	9.47	8.725	7.880	9.661	2.001
26	2-Hydroxypyridine	12.75	10.058	21.117	12.859	0.858
27	4-Hydroxypyridine	10.77	8.397	22.031	10.552	−2.025
28	2-Iodopyridine	12.18	11.415	6.278	12.536	2.923
29	3-Iodopyridine	10.75	10.562	1.746	10.548	−1.880
30	4-Isopropylpyridine	7.98	7.395	7.326	7.936	−0.548
31	3-(N-Methoxyacetamido)pyridine	10.48	10.700	−2.102	10.295	−1.768
32	2-Methoxycarbonylpyridine	11.79	11.607	1.551	11.864	0.628
33	4-Methoxycarbonylpyridine	10.74	10.390	3.263	10.815	0.698
34	2-Methoxypyridine	10.94	9.350	14.537	11.132	1.752
35	3-Methoxypyridine	9.09	8.169	10.134	8.713	−4.150
36	2-(Methylaminomethyl)6-methylpyridine	10.97	8.322	24.134	10.948	−0.201
37	4-(Methylamino)pyridine	4.35	6.148	−41.326	4.161	−4.349
38	3-(N-Methylbenzamido)pyridine	10.34	10.947	−5.870	10.294	−0.443
39	2-(N-Ethylmethanesulfonamido)pyridine	12.27	10.925	10.959	12.383	0.923
40	3-(N-Ethylmethanesulfonamido)pyridine	10.06	9.988	0.711	9.604	−4.536
41	4-(N-Ethylmethanesulfonamido)pyridine	8.86	9.586	−8.199	9.654	8.964
42	3-Methylpyridine	8.32	7.596	8.701	8.516	2.352
43	3-Nitropyridine	13.21	14.141	−7.050	13.094	−0.881
44	2-Propylpyridine	7.70	9.730	−26.360	7.700	0.000
45	2-Pyridinealdoxime	10.44	11.050	−5.840	10.140	−2.876
46	3-Pyridinealdoxime	9.93	9.978	−0.482	10.345	4.174
47	4-Pyridinealdoxime	9.27	9.838	−6.128	9.141	−1.397
48	3-Pyridinecarbaldehyde	10.20	11.229	−10.084	10.051	−1.461
49	3-Pyridinecarbamide(nicotinamide)	10.67	10.017	6.124	10.596	−0.697
50	Pyridine-3-carboxylic acid	11.93	11.557	3.128	12.301	3.109
51	Pyridine-2,3-dicarboxylic acid	11.64	13.845	−18.940	11.588	−0.447
52	Pyridine-2,4-dicarboxylic acid	11.77	12.115	−2.928	11.760	−0.089

(*continued*)

Table 2 (*continued*)

No.	Compound	Exp.	Calculated 1	*IPD*1	Calculated 2	*IPD*2
53	Pyridine-2,6-dicarboxylic acid	11.84	12.400	−4.732	11.574	−2.248
54	2,4,6-Trimethylpyridine	6.57	8.502	−29.406	6.638	1.040
55	4-Vinylpyridine	8.38	8.976	−7.115	8.591	2.514
Validation						
56	2-Acetylpyridine	11.36	10.596	6.699	10.767	−5.196
57	2-Amino-4-methylpyridine	6.52	9.471	−45.262	6.338	−2.787
58	3-Aminopyridine	7.97	10.012	−25.621	7.999	0.364
59	2-tert-Butylpyridine	8.24	7.337	10.963	8.239	−0.011
60	2-Chloropyridine	13.51	10.367	23.267	13.273	−1.755
61	2,6-Di-tert-butylpyridine	10.42	6.835	34.406	10.577	1.508
62	3,4-Dimethylpyridine	7.53	8.603	−14.246	7.322	−2.764
63	2-Fluoropyridine	14.45	10.415	27.909	13.179	−8.776
64	2-(2-Hydroxyphenyl)pyridine	9.81	11.328	−15.472	9.795	−0.151
65	2-Isopropylpyridine	8.17	8.979	−9.896	8.133	−0.454
66	4-(N-Methoxyacetamido)pyridine	9.38	11.606	−23.731	9.365	−0.164
67	4-Methoxypyridine	7.53	9.898	−31.454	7.426	−1.376
68	2-(N-Methylbenzamido)pyridine	12.56	11.681	7.001	12.440	−0.959
69	2-Methylpyridine	8.04	9.155	−13.866	7.668	−4.631
70	4-Nitropyridine	12.77	12.015	5.909	12.518	−1.971
71	2-Pyridinecarbaldehyde	10.16	10.658	−4.899	10.191	0.304
72	Pyridine-2-carboxylic acid	12.99	11.215	13.661	13.485	3.814
73	2,3,6-Trimethylpyridine	6.40	8.753	−36.769	6.478	1.225
Prediction						
74	4-Acetylpyridine	10.50	10.055	4.190	10.264	−2.200
75	2-Amino-6-methylpyridine	6.59	10.213	−54.972	6.432	−2.396
76	2-Benzylpyridine	8.87	9.698	−9.338	8.355	−5.808
77	4-tert-Butylpyridine	8.01	6.520	18.601	7.990	−0.246
78	4-Chloropyridine	10.17	8.590	15.540	10.263	0.914
79	2,5-Dimethylpyridine	7.57	7.267	4.007	7.402	−2.221
80	4-Ethoxypyridine	7.33	10.514	−43.438	7.698	5.022
81	4-Formyl-3-hydroxypyridine	9.95	10.833	−8.872	10.012	0.624
82	3-Hydroxypyridine	9.20	10.368	−12.699	9.627	4.640
83	2-(N-Methoxyacetamido)pyridine	11.99	11.325	5.547	12.412	3.517
84	3-Methoxycarbonylpyridine	10.87	9.309	14.362	10.781	−0.817
85	4-Methylamino-3-methylpyridine	4.17	9.859	−136.423	4.813	15.408
86	4-(N-Methylbenzamido)pyridine	9.32	10.221	−9.667	9.407	0.931
87	4-Methylpyridine	8.00	8.387	−4.839	8.388	4.854
88	Pyridine	8.83	8.737	1.051	8.852	0.246
89	4-Pyridinecarbaldehyde	9.26	10.041	−8.435	9.876	6.654
90	Pyridine-4-carboxylic acid	12.16	10.918	10.210	11.974	−1.532
91	2-Vinylpyridine	9.02	10.602	−17.539	9.535	5.707

[a] Experimental values of pK_b are from Ref. [36]. 1 and 2 denote to the values obtained by PC-GA-MLR and PC-GA-ANN models

in the prediction set, which were not used in the modeling procedure (Table 2). The calculated values of pK_b 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 expected, the calculated pK_b values are in good agreement with those of the experimental values. The correlation equation for all of the calculated values of pK_b from the ANN model and the experimental values is given by Eq. (1):

$$pK_b(\mathrm{cal}) = 0.9814pK_b(\mathrm{exp}) + 0.1765 \qquad (1)$$

($R = 0.9921$; $MPD = 2.138$; $RMSE = 0.2819$; $F = 5559.85$)

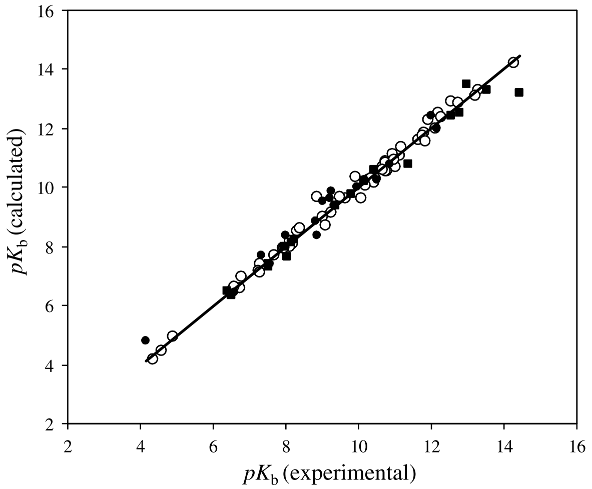

Fig. 4 Plot of the calculated values of pK_b from the PC-GA-ANN model *versus* the experimental values of it for training (○), validation (■), and prediction (●) sets

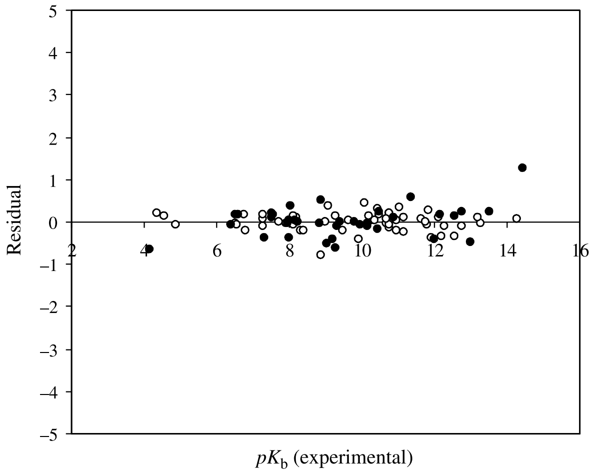

Fig. 5 Plot of the residual for calculated values of pK_b from the PC-GA-ANN model *versus* the experimental values of it for training (○), validation (■), and prediction (●) sets

Similarly, the correlation of pK_b (cal) *versus* pK_b (exp) values in the prediction set gives Eq. (2):

$$pK_b(\text{cal}) = 0.9562pK_b(\text{exp}) + 0.520 \qquad (2)$$

($R = 0.9853$; $MPD = 3.541$; $RMSE = 0.3448$; $F = 533.69$)

The plot of residual for pK_b values in the training, validation, and prediction sets *versus* the experimental values of it has been illustrated in Fig. 5. As can be seen, the propagation of errors in both sides of zero is random.

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_b values of various pyridines in aqueous solution.

Table 3 Comparison of statistical parameters obtained by the PC-GA-MLR and PC-GA-ANN models for pK_b values of pyridines[a]

Model	$RMSE_{tot}$	$RMSE_{train}$	$RMSE_{valid}$	$RMSE_{pred}$
PC-GA-MLR	1.5317	1.0623	2.1010	1.9995
PC-GA-ANN	0.2819	0.2130	0.3797	0.3448

[a] Subscript train is referring to the training set, valid is referring to the validation set and the pred is referring to the prediction set, tot is referring to the total data set, and *RMSE* is the root mean square error

As a result, it was found that properly selected and trained neural network could fairly represent dependence of the basicity constant of pyridines in aqueous solution on the PCs. Then the optimized neural network could simulate the complicated nonlinear relationship between pK_b value and the PCs. The *MPD* and *RMSE* are 21.096 and 1.9995 for the prediction set by the PC-GA-MLR model should be compared with the values of 3.541 and 0.3448, 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 has shown a large improvement. These improvements are due to the fact that pK_b values of pyridines shows non-linear correlations with the principal components.

Conclusions

Quantitative-structure activity relationships were applied on the basicity constant of 91 various pyridines in water at 25°C by using the principal component-genetic algorithm-multi parameter linear regression (PC-GA-MLR) and principal component-genetic algorithm-artificial neural network (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 demonstrate superiority of the PC-GA-ANN model over the PC-GA-MLR model. A mean percent deviation of 21.096 for the prediction set by the PC-GA-MLR model should be compared with the value of 3.541 for the PC-GA-ANN model. Since the improvement of the results obtained using non-linear model (PC-GA-ANN) is considerable, it can be concluded that the non-linear characteristics of the principal components on the pK_b values of the compounds in water is serious.

Data and methodology

Basicity constant and theoretical descriptors

Basicity constant of pyridines are literature values at 25°C [39]. The *z*-matrices (molecular models) were constructed with HyperChem 7.0 and molecular structures were optimized using the AM1 algorithm [40]. In order to calculate the theoretical descriptors, the Dragon package version 2.1 was used [41]. 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 1481 theoretical descriptors were calculated for each compound in data sets (91 compounds).

Principal component analysis (PCA)

The theoretical descriptors were reduced by the following procedure:

1) descriptors that are constant were eliminated (436 descriptors).
2) in addition, to decrease the redundancy existing in the descriptors data matrix, the correlation of descriptors with each other and with the pK_b of the molecules are examined, and collinear descriptors ($R > 0.9$) are detected. Among the collinear descriptors, one that has the highest correlation with pK_b values is retained, and the others are removed from the data matrix (742 descriptors).
3) before statistical analysis, the descriptors are scaled to zero mean and unit variance (autoscaling procedure). The data matrix (303 descriptors) is subjected to principal component analysis using Matlab software package [42]. Multiparameter linear regression was obtained using spss software [43].

Genetic algorithm (GA)

To select the most relevant principal components, the evolution of a population was simulated [44–48]. Each individual of the population defined by a chromosome of binary values represented a subset of principal components. The number of genes at each chromosome was equal to the number of principal components. The population of the first generation was selected randomly. A gene took a value of 1 if its corresponding principal component 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 principal components [48], 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. 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 [20]. The GA program was written in Matlab 6.5 [49].

Artificial neural network (ANN)

A feed forward artificial neural network with a back-propagation of error algorithm was used to process the non-linear relationship between the selected principal components and the basicity 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. Optimization of the weights and biases was carried out according to the resilient back-propagation algorithm. The data set was randomly divided into three groups: a training set, a validation set and a prediction set consisting of 55, 18, and 18 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:

$$MPD = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{(P_i^{\exp} - P_i^{\mathrm{cal}})}{P_i^{\exp}}\right| \tag{3}$$

$$RMSE = \sqrt{\sum_{i=1}^{N}\frac{(P_i^{\exp} - P_i^{\mathrm{cal}})^2}{N}} \tag{4}$$

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

$$IPD = 100 \times \left(\frac{P_i^{\mathrm{cal}} - P_i^{\exp}}{P_i^{\exp}}\right) \tag{5}$$

The processing of the data was carried using Matlab 6.5 [42]. The neural networks were implemented using Neural Network Toolbox Ver. 4.0 for Matlab [50].

Acknowledgements

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

References

1. Yao XJ, Wang YW, Zhang XY, Zhang RS, Liu MC, Hu ZD, Fan BT (2002) Chemom Intell Lab Syst 62:217
2. Guha R, Serra JR, Jurs PC (2004) J Mol Graph Model 23:1
3. Krogsgaard-Larsen P, Liljefors T, Madsen U (2002) Textbook of Drug Design and Discovery. Taylor & Francis, London
4. Consonni V, Todeschini R, Pavan M, Gramatica P (2002) J Chem Inf Comput Sci 42:693
5. Karthikeyan M, Glen RC, Bender A (2005) J Chem Inf Model 45:581

6. Melnikov AA, Palyulin VA, Zefirov NS (2007) J Chem Inf Model 47:2077
7. Ajmani S, Rogers SC, Barley MH, Livingstone DJ (2006) J Chem Inf Model 46:2043
8. Katritzky AR, Stoyanova-Slavova IB, Dobchev DA (2007) J Mol Graph Model 26:529
9. Shamsipur M, Siroueinejad A, Hemmateenejad B, Abbaspour A, Sharghi H, Alizadeh K, Arshadi S (2007) J Electranal Chem 600:345
10. Todeschini R, Consonni V (2000) Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, Germany
11. Sutter JM, Kalivas JH, Lang PM (1992) J Chemometr 6:217
12. Vendrame R, Braga RS, Takahata Y, Galvao DS (1999) J Chem Inf Comput Sci 39:1094
13. Malinowski ER (2002) Factor Analysis in Chemistry. Wiley-Interscience, New York
14. Katritzky AR, Tulp I, Fara DC, Lauria A, Maran U, Acree WE (2005) J Chem Inf Model 45:913
15. Hemmateenejad B, Akhond M, Miri R, Shamsipur M (2003) J Chem Inf Comput Sci 43:1328
16. Hemmateenejad B, Shamsipur M (2004) Internet Electron, J Mol Des 3:316
17. Jalali-Heravi M, Kyani A (2004) J Chem Inf Comput Sci 44:1328
18. Hemmateenejad B, Safarpour MA, Miri R, Nesari N (2005) J Chem Inf Model 45:190
19. Hemmateenejad B, Safarpour M, Miri R, Taghavi F (2004) J Comput Chem 25:1495
20. Depczynski U, Frost VJ, Molt K (2000) Anal Chim Acta 420:217
21. Hemmateenejad B (2005) Chemom Intell Lab Syst 75:231
22. Goldberg DE (2000) Genetic Algorithm in Search, Optimization and Machine Learning. Addison-Wesley-Longman, Reading, MA, USA
23. Cho SJ, Hermsmeier MA (2002) J Chem Inf Comput Sci 42:927
24. Despagne F, Massart DL (1998) Analyst 123:157
25. Zupan J, Gasteiger J (1999) Neural Networks in Chemistry and Drug Design. Wiley-VCH, Germany
26. Meiler J, Meusinger R, Will M (2000) J Chem Inf Comput Sci 40:1169
27. Habibi-Yangjeh A, Nooshyar M (2005) Phys Chem Liq 43:239
28. Habibi-Yangjeh A, Nooshyar M (2005) Bull Korean Chem Soc 26:139
29. Habibi-Yangjeh A, Danandeh-Jenagharad M, Nooshyar M (2005) Bull Korean Chem Soc 26:2007
30. Habibi-Yangjeh A (2007) Phys Chem Liq 45:471
31. Tabaraki R, Khayamian T, Ensafi AA (2006) J Mol Graph Model 25:46
32. Habibi-Yangjeh A, Danandeh-Jenagharad M (2007) Indian J Chem 46B:478
33. Habibi-Yangjeh A, Esmailian M (2007) Bull Korean Chem Soc 28:1477
34. Habibi-Yangjeh A, Pourbasheer E, Danandeh-Jenagharad M (2008) Bull Korean Chem Soc 29:833
35. Jover J, Bosque R, Sales J (2007) QSAR Comb Sci 26:385
36. Ivanova AA, Baskin II, Palyulin VA, Zefirov ANS (2007) Doklady Chem 413:90
37. Luan F, Ma W, Zhang H, Zhang X, Liu M, Hu Z, Fan B (2005) Pharmaceut Res 22:1454
38. Habibi-Yangjeh A, Danandeh-Jenagharad M, Nooshyar M (2006) J Mol Model 12:338
39. Dean JA (1999) Lange's Handbook of Chemistry, 15th edn. McGraw-Hill Inc
40. HyperChem Release 7, HyperCube Inc., http://www.hyper.com.
41. Todeschini R, Milano Chemometrics and QSPR Group, http://www.disat.unimib.it/vhm.
42. Matlab 6.5. Mathworks, 1984–2002
43. SPSS for Windows, Statistical Package for IBM PC, SPSS Inc., http://www.spss.com
44. Cho SJ, Hermsmeier MA (2002) J Chem Inf Comput Sci 42:927
45. Baumann K, Albert H, Von Korff M (2002) J Chemometr 16:339
46. Lu Q, Shen G, Yu R (2002) J Comput Chem 23:1357
47. Ahmad S, Gromiha MM (2003) J Comput Chem 24:1313
48. Deeb O, Hemmateenejad B, Jaber A, Garduno-Juarez R, Miri R (2007) Chemosphere 67:2122
49. The Mathworks Inc (2002) Genetic Algorithm and Direct Search Toolbox User's Guide, Massachusetts
50. The Mathworks Inc (2002) Neural Network Toolbox User's Guide, Massachusetts