J Mol Model (2006) 12: 338–347
DOI 10.1007/s00894-005-0050-6

ORIGINAL PAPER

Aziz Habibi-Yangjeh · Mohammad Danandeh-Jenagharad · Mahdi Nooshyar

Application of artificial neural networks for predicting the aqueous acidity of various phenols using QSAR

Received: 25 April 2005 / Accepted: 22 July 2005 / Published online: 13 December 2005

Abstract Artificial neural networks (ANNs) have been successfully trained to model and predict the acidity constants (pK_a) of 128 various phenols with diverse chemical structures using a quantitative structure-activity relationship. An ANN with 6-14-1 architecture was generated using six molecular descriptors that appear in the multi-parameter linear regression (MLR) model. The polarizability term (π_I), most positive charge of acidic hydrogen atom (q^+), molecular weight (MW), most negative charge of the phenolic oxygen atom (q^-), the hydrogen-bond accepting ability (ε_B) and partial-charge weighted topological electronic (PCWTE) descriptors are inputs and its output is pK_a. It was found that a properly selected and trained neural network with 106 phenols could represent the dependence of the acidity constant on molecular descriptors fairly well. For evaluation of the predictive power of the ANN, an optimized network was used to predict the pK_as of 22 compounds in the prediction set, which were not used in the optimization procedure. A squared correlation coefficient (R^2) and root mean square error (RMSE) of 0.8950 and 0.5621 for the prediction set by the MLR model should be compared with the values of 0.99996 and 0.0114 by the ANN model. These improvements are due to the fact that the pK_a of phenols shows non-linear correlations with the molecular descriptors.

Keywords Quantitative structure – activity relationship · Artificial neural networks · Acidity constant · Theoretical descriptors · Phenols

A. Habibi-Yangjeh (✉) ·
M. Danandeh-Jenagharad · M. Nooshyar
Department of Chemistry, Faculty of Science,
University of Mohaghegh Ardebili,
P. O. Box 179, Ardebil, Iran
e-mail: habibiyangjeh@yahoo.com
Tel.: +98-451-5512801
Fax: +98-451-5510800

Introduction

The prediction of physicochemical and biological properties/activities of organic molecules is the main objective of quantitative structure-property/activity relationships (QSPRs/QSARs) [1–6]. 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. 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 [7, 8].

Various methods for constructing QSPR/QSAR models have been used including multi-parameter linear regression (MLR), principal component analysis (PCA) and partial least-squares regression (PLS) [9–12]. In addition, artificial neural networks (ANNs) have become popular due to their success where complex non-linear relationships exist amongst data [13–15]. ANNs are biologically inspired computer programs designed to simulate the way in which the human brain processes information. ANNs gather their knowledge by detecting the patterns and relationships in data and learned (or trained) through experience, not from programming. There are many types of neural networks designed by now and new ones are invented every week [15]. The behavior of a neural network is determined by transfer functions of its neurons, by learning rules, and by the architecture itself. An ANN is formed from artificial neurons or processing elements (PE), connected with coefficients (weights), which constitute the neural structure and are organized in layers. The first layer is termed the input layer, and the last layer is the output layer. The layers of neurons between the input and output layers are called hidden layers. The wide applicability of ANNs stems from their flexibility and ability to model non-linear systems without prior knowledge of an empirical model. Neural networks do not need on explicit formulation of the

Table 1 Descriptors, symbols and results of the multi-parameter linear regression (MLR) model[a]

Number	Descriptor	Symbol	Coefficient	β
1	Polarizability term	π_I	−8.413	0.376
2	Most positive charge of acidic hydrogen atom	q^+	−38.289	0.224
3	Molecular weight	MW	0.0040	0.120
4	Most negative charge of the phenolic oxygen atom	q^-	13.034	0.118
5	The hydrogen-bond accepting ability	ε_B	107.868	0.293
6	Partial charge weighted topological electronic	PCWTE	0.0485	0.100
7	Constant		5.133	

[a]The polarizability term (π_I) is obtained by dividing the polarizability volume by the molecular volume. ε_B is equal $0.3-0.01(E_{lw}-E_h)$, in which E_{lw} and E_h are referring to the LUMO energy for water and HOMO energy for the compound, respectively
β is standardized coefficient of descriptors

mathematical or physical relationships of the problem. These give ANNs an advantage over traditional fitting methods for some chemical applications. For these reasons, in recent years, ANNs have been used to a wide variety of chemical problems such as simulation of mass spectra, ion interaction chromatography, aqueous solubility, partition coefficients, simulation of nuclear magnetic resonance spectra, prediction of bioconcentration factors, solvent effects on reaction rates and prediction of normalized polarity parameters in mixed solvent systems [16–23].

Phenol derivatives have been used widely as materials in the manufacture of plastics, as constituents of industrial disinfectants and as chemical reagents in industrial processes and display a variety of biological activities that relate to their acid/base behavior [24]. On the other hand, interpretation and prediction of pK_a values for chemical compounds are of general importance and usefulness for chemists [25]. Although in the last years several theoretical studies have been performed for correlation of pK_a values with structure, linear equations have been used in these studies [25–30].

The main aim of the present work is to develop a QSAR model based on molecular descriptors using ANNs, for the first time, for modeling and predicting pK_a values of various phenols with diverse chemical structures (including 128 phenols). In the first step, a MLR model was constructed. Then, for inspection of non-linear interactions/relation between different molecular parameters in the model, an ANN model was generated for predicting the pK_a values and the results were compared with the experimental and calculated values using the MLR model.

Methods and procedure

Descriptor generation

In order to calculate the theoretical descriptors, *Z*-matrices (molecular models) were constructed with HyperChem 7.0 and molecular structures were optimized using the AM1 Hamiltonian [31]. In order to calculate some of theoretical descriptors, the molecular geometries of the molecules were further optimized with the same Hamiltonian in the MOPAC program version 6.0. The other molecular electronic descriptors were calculated using the *Dragon* package version 2.1 [32]. For this propose the output of the HyperChem software for each compound was fed into the *Dragon* program and the descriptors were calculated. As a result, a total of 18 theoretical descriptors were calculated for each compound in the data set (128 phenols).

Linear correlations

The acidity constant of the phenols are literature values at 25°C [33]. An MLR model was developed for predicting pK_a values by molecular descriptors. The method of stepwise multi-parameter linear regression was used to select the most important descriptors and to calculate the coefficients relating the pK_a to the descriptors. The MLR models were generated using spss/pc software package release 9.

Table 2 Correlation coefficients between various theoretical descriptors that have been used in the multi-parameter linear regression (MLR) and artificial neural network (ANN) models

Descriptor	π_I	q^+	q^-	ε_B	MW	PCWTE
π_I	1	0.633	0.731	0.557	0.159	0.487
q^+	0.633	1	0.610	0.661	0.249	0.130
q^-	0.731	0.610	1	0.607	0.081	0.306
ε_B	0.557	0.661	0.607	1	0.296	0.132
MW	0.159	0.249	0.081	0.296	1	0.285
PCWTE	0.487	0.130	0.306	0.132	0.285	1

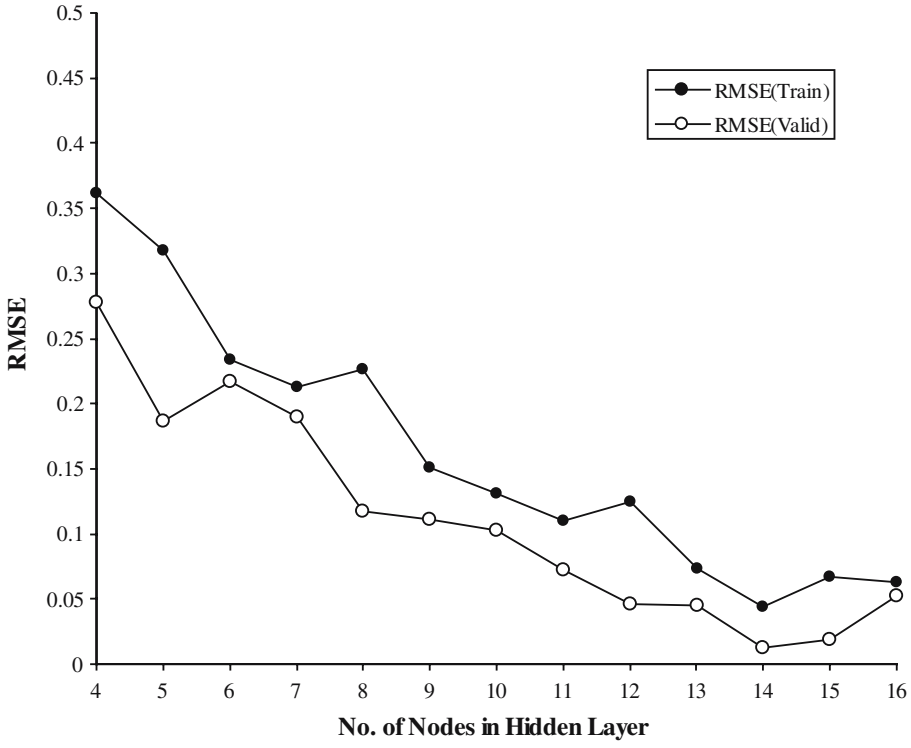

Fig. 1 Plot of RMSE for training and validation sets versus the number of nodes in hidden layer

Neural network generation

The specification of a typical neural network model requires the choice of the type of inputs, the number of hidden layers, the number of neurons in each hidden layer and the connection structure between the inputs and the output layers. The number of input nodes in the ANNs was equal to the number of molecular descriptors in the MLR model. A three-layer network with a sigmoidal transfer function was designed. The initial weights were randomly selected

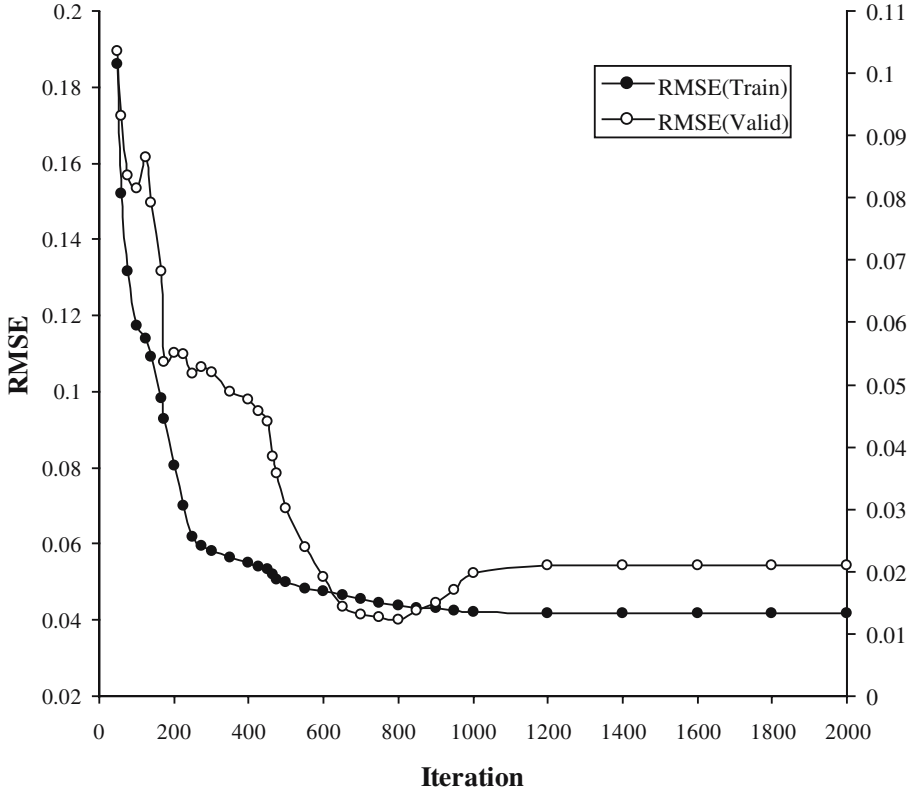

Fig. 2 Plot of RMSE for training and validation sets versus the number of iterations

Table 3 Experimental and calculated values of pK_a for various phenols in water at 25°C for training, validation and prediction sets by multi-parameter linear regression (MLR) and artificial neural network (ANN) models along with individual percent deviation (IPD)[a]

Number	Compound	Exp.	MLR	IPD_{MLR}	ANN	IPD_{ANN}
Training set						
1	2-acetylphenol	9.19	9.036	−1.67	9.195	0.06
2	2-allylphenol	10.28	9.857	−4.11	10.263	−0.17
3	2-bromophenol	8.45	8.610	1.87	8.456	0.04
4	4-bromophenol	9.34	8.759	−6.22	9.356	0.17
5	2,6-di-*tert*-butyl-4-bromophenol	10.83	10.608	−2.05	10.832	0.02
6	2,6-di-*tert*-butyl-4-methoxyphenol	12.15	11.771	−3.12	12.142	−0.07
7	2,4-di-*tert*-butylphenol	11.64	11.569	−0.61	11.630	−0.08
8	2-*tert*-butylphenol	11.24	10.959	−2.50	11.275	0.31
9	3-*tert*-butylphenol	10.10	10.825	7.17	10.117	0.17
10	1-chloro-2,6-dimethyl-4-hydroxybenzene	9.55	9.734	1.93	9.542	−0.07
11	4-chloro-2,6-dinitrophenol	2.97	2.974	0.12	2.964	−0.20
12	4-chloro-2-nitrophenol	6.48	6.298	−2.82	6.481	0.01
13	2-chlorophenol	8.55	8.885	3.92	8.533	−0.20
14	4-chlorophenol	9.43	9.081	−3.70	9.469	0.41
15	*o*-cresol	10.26	10.119	−1.38	10.183	−0.75
16	*p*-cresol	10.26	10.137	−1.20	10.306	0.45
17	4-cyano-2,6-dimethylphenol	8.27	8.235	−0.43	8.259	−0.14
18	3-cyanophenol	8.61	8.205	−4.70	8.605	−0.06
19	3,5-dibromophenol	8.06	7.430	−7.77	8.060	0.04
20	2,3-dichlorophenol	7.44	8.081	8.61	7.434	−0.08
21	2,4-dichlorophenol	7.85	8.131	3.58	7.878	0.36
22	3,4-dichlorophenol	8.63	8.315	−3.65	8.632	0.02
23	3,5-dichlorophenol	8.18	8.010	−2.07	8.211	0.39
24	3-(diethoxyphosphinyl)phenol	8.68	9.157	5.50	8.687	0.08
25	4-(diethoxyphosphinyl)phenol	8.28	8.508	2.76	8.287	0.08
26	1,2-dihydroxybenzene	9.36	9.720	3.90	9.351	−0.05
27	1,3-dihydroxybenzene	9.44	9.595	1.64	9.457	0.18
28	1,4-dihydroxy-2,6-dinitrobenzene	4.42	3.667	−17.04	4.421	0.01
29	1,3-dihydroxy-2-methylbenzene	10.05	9.773	−2.75	9.963	−0.87
30	3,5-diiodophenol	8.10	6.947	−14.26	8.118	0.18
31	3,5-dimethoxyphenol	9.35	9.460	1.23	9.344	−0.01
32	3,5-dimethyl-4-nitrophenol	8.25	7.826	−5.09	8.247	0.02
33	2,3-dimethylphenol	10.50	10.346	−1.47	10.301	−1.89
34	2,5-dimethylphenol	10.22	10.365	1.42	10.303	0.81
35	2,6-dimethylphenol	10.59	10.267	−3.05	10.505	−0.81
36	3,5-dimethylphenol	10.15	10.309	1.56	10.269	1.17
37	2,4-dinitrophenol	4.08	4.577	12.18	4.078	−0.05
38	3,4-dinitrophenol	5.42	6.905	27.31	5.438	0.25
39	3,5-dinitrophenol	6.73	5.339	−20.70	6.736	0.06
40	2-ethoxyphenol	10.11	10.107	−0.02	10.107	−0.02
41	3-ethoxyphenol	9.66	9.903	2.57	9.659	0.04
42	3-ethylphenol	10.07	10.314	2.43	10.047	−0.23
43	4-ethylphenol	10.00	10.359	3.59	10.212	2.12
44	3-fluorophenol	9.29	9.162	−1.38	9.291	0.01
45	4-fluorophenol	9.89	9.384	−5.11	9.847	−0.43
46	3′-hydroxyacetophenone	9.19	9.261	0.77	9.189	−0.01
47	4′-hydroxyacetophenone	8.05	8.939	11.04	8.053	0.04
48	3-hydroxybenzaldehyde	9.00	8.751	−2.77	9.006	0.07
49	4-hydroxybenzaldehyde	7.62	8.465	11.08	7.620	0.00
50	3-hydroxybenzyl alcohol	9.83	9.834	0.04	9.886	0.57
51	4-hydroxybenzyl alcohol	9.82	9.911	0.92	9.833	0.13
52	1-hydroxy-2,4-xdihydroxymethylbenzene	9.79	9.803	0.14	9.796	0.06

Table 3 (continued)

Number	Compound	Exp.	MLR	IPD_{MLR}	ANN	IPD_{ANN}
53	2-hydroxy-3-methoxybenzaldehyde	7.91	7.915	0.04	7.914	0.03
54	4-hydroxy-3-methoxybenzaldehyde	7.40	8.864	19.85	7.392	−0.05
55	(2-hydroxy-5-methylbenzene)-methanol	10.15	9.916	−2.31	10.148	−0.02
56	1-hydroxy-2-propylbenzene	10.50	10.532	0.31	10.425	−0.72
57	4-hydroxy-α,α,α-trifluorotoluene	8.68	8.743	0.79	8.673	−0.02
58	4-indanol	10.32	10.175	−1.40	10.316	−0.04
59	2-iodophenol	8.46	8.229	−2.78	8.466	0.03
60	4-iodophenol	9.20	8.347	−9.27	9.210	0.11
61	2,6-di-iodo-4-nitrophenol	3.32	4.340	30.74	3.325	0.14
62	3-methoxyphenol	9.65	9.641	−0.12	9.732	0.83
63	4-methoxyphenol	10.20	10.027	−1.69	10.163	−0.37
64	6-methyl-2-butylphenol	11.72	10.737	−8.39	11.724	0.04
65	2-methyl-4-*tert*-butylphenol	10.59	11.024	4.10	10.606	0.15
66	2,2′-methylene-bis(4,6-dichlorophenol)	5.60	6.998	24.96	5.601	0.03
67	4-methylsulfonyl-3,5-dimethylphenol	8.13	8.229	1.22	8.130	0.00
68	4-methylsulfonylphenol	7.83	7.692	−1.76	7.840	0.13
69	3-(*s*-methylthio)phenol	9.53	9.495	−0.37	9.550	0.20
70	2-nitrohydroquinone	7.63	7.076	−7.26	7.632	0.03
71	2-nitrophenol	7.22	6.837	−5.33	7.225	0.04
72	4-nitrophenol	7.15	6.969	−2.53	7.144	−0.08
73	4-nitrosophenol	6.48	7.757	19.71	6.480	0.00
74	2-phenylphenol	9.55	9.122	−4.48	9.556	0.06
75	3-phenylphenol	9.63	8.944	−7.13	9.629	−0.01
76	5,6,7,8-tetrahydro-1-naphthol	10.28	10.365	0.83	10.284	0.04
77	5,6,7,8-tetrahydro-2-naphthol	10.48	10.443	−0.35	10.457	−0.22
78	2,4,5-trichlorophenol	7.37	7.398	0.38	7.363	−0.09
79	3,4,5-trichlorophenol	7.84	7.577	−3.34	7.829	−0.13
80	1,2,3-trihydroxybenzene	9.03	8.792	−2.63	9.033	0.03
81	1,3,5-trihydroxybenzene	8.45	9.127	8.01	8.461	0.13
82	2,4,5-trimethylphenol	10.57	10.616	0.43	10.603	0.31
83	3,4,5-trimethylphenol	10.25	10.633	3.74	10.197	−0.52
84	2,4,6-tripropylphenol	11.47	11.134	−2.93	11.481	0.10
Validation set						
85	4-acetylphenol	8.05	8.992	11.70	8.049	−0.01
86	2,6-di-*tert*-Butyl-4-methylphenol	12.23	11.646	−4.78	12.261	0.25
87	4-*tert*-butylphenol	10.31	10.871	5.44	10.286	−0.24
88	*m*-cresol	10.00	10.068	0.68	10.002	0.02
89	2,6-dichlorophenol	6.78	7.684	13.33	6.773	−0.11
90	3,4-dihydroxybenzaldehyde	7.55	8.008	6.07	7.551	0.01
91	1,2-dihydroxy-3-nitrobenzene	6.68	5.752	−13.88	6.677	−0.04
92	1,4-dihydroxy-2,3,5,6-tetramethylbenzene	11.25	10.826	−3.77	11.269	0.17
93	2,6-dimethyl-4-nitrophenol	7.19	7.731	7.53	7.175	−0.21
94	3,4-dimethylphenol	10.32	10.387	0.65	10.302	−0.18
95	2,6-dinitrophenol	3.71	3.359	−9.55	3.713	−0.01
96	2-fluorophenol	8.73	9.078	3.98	8.732	0.02
97	2-hydroxybenzaldehyde	8.34	8.523	2.19	8.336	−0.05
98	4-hydroxybenzonitrile	7.95	8.041	1.14	7.969	0.24
99	1-hydroxy-2,4,6-trihydroxymethylbenzene	9.56	9.566	0.07	9.563	0.03
100	2-methoxy-4-(2-propenyl)phenol	10.00	9.867	−1.33	10.000	0.00
101	3-methylsulfonylphenol	9.33	8.370	−10.29	9.328	−0.03
102	4-(*s*-methylthio)phenol	9.53	9.739	2.19	9.513	−0.18
103	phenol	9.99	9.841	−1.49	9.998	0.08
104	4-phenylphenol	9.55	9.029	−5.46	9.548	−0.02
105	3-trifluoromethylphenol	8.95	8.906	−0.49	8.956	0.07
106	2,4,6-trimethylphenol	10.88	10.524	−3.27	10.874	−0.05

Table 3 (continued)

Number	Compound	Exp.	MLR	IPD_{MLR}	ANN	IPD_{ANN}
Prediction set						
107	3-bromophenol	9.03	8.593	−4.85	9.013	−0.20
108	2,6-di-*tert*-butylphenol	11.70	11.424	−2.36	11.682	−0.16
109	4-chloro-3-methylphenol	9.55	9.424	−1.31	9.520	−0.30
110	3-chlorophenol	9.10	8.902	−2.18	9.088	−0.13
111	4-cyano-3,5-dimethylphenol	8.21	8.896	8.36	8.203	−0.09
112	1,3-dichloro-2,5-dihydroxybenzene	7.30	7.814	7.05	7.294	−0.08
113	3,5-diethoxyphenol	9.37	9.861	5.24	9.378	0.09
114	1,4-dihydroxybenzene	9.91	9.980	0.71	9.901	−0.09
115	1,2-dihydroxy-4-nitrobenzene	6.70	6.935	3.49	6.694	−0.10
116	2,4-dimethylphenol	10.58	10.383	−1.86	10.570	−0.10
117	2,5-dinitrophenol	5.22	4.726	−9.39	5.218	0.04
118	2-ethylphenol	10.20	10.375	1.72	10.221	0.21
119	2′-hydroxyacetophenone	9.90	9.170	−7.37	9.909	0.09
120	2-hydroxybenzyl alcohol	9.92	9.647	−2.76	9.920	−0.01
121	3-hydroxy-4-methoxybenzaldehyde	8.89	8.136	−8.47	8.890	0.01
122	3-hydroxy-4-nitrotoluene	7.41	6.084	−17.89	7.407	−0.05
123	3-iodophenol	8.88	8.328	−6.21	8.868	−0.12
124	2-methoxyphenol	9.99	9.811	−1.79	9.996	0.06
125	2,2′-methylene-bis(4-chlorophenol)	7.60	8.563	12.67	7.605	0.06
126	3-nitrophenol	8.36	7.350	−12.08	8.355	−0.06
127	2,4,6-tri-*tert*-butylphenol	12.19	11.933	−2.110	12.183	−0.06
128	2,3,4-trimethylphenol	10.59	10.625	0.33	10.590	0.00

[a] *MLR* and *ANN* refer to multi-parameter linear regression and artificial neural network calculated values of pK_a, respectively
Exp. is referring to the experimental values of pK_a

between 0 and 1. Before training, the input and output values were normalized between 0.1 and 0.9. Optimization of the weights and biases was carried out according to the Levenberg–Marquardt algorithm for back-propagation of errors, which, although requiring far more extensive computer memory, is significantly faster than other algorithms based on gradient descent [34]. The data set was randomly divided into three groups: a training set, a validation set and a prediction set consisting of 84, 22 and 22 compounds, respectively. The training and validation sets were used for the model generation and the prediction set was used for evaluation of the model generated, because a prediction set is a better estimator of the ANN generalization ability than a validation (monitoring) set [35].

The performances of training, validation and prediction of ANNs are evaluated by the mean percentage deviation (MPD) and root-mean square error (RMSE), which are defined as follows:

$$\mathrm{MPD} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{P_i^{calc} - P_i^{\exp}}{P_i^{\exp}}\right| \quad (1)$$

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

where $P_i^{\exp}$ and P_i^{calc} are experimental and calculated values of pK_a with the models and N denote the number of data points.

The individual percent deviation (IPD) is defined as follows:

$$\mathrm{IPD} = 100 \times \left(\frac{P_i^{calc} - P_i^{\exp}}{P_i^{\exp}}\right) \quad (3)$$

The data was processed using Matlab 6.5 [36]. The neural networks were implemented using Neural Network Toolbox Ver. 4.0 for Matlab [34].

Results and discussion

Multi-parameter linear correlation of the pK_a values of 84 phenols in the training set versus the molecular descriptors gives the results shown in Table 1. It can be seen from this table that six descriptors appear in the MLR model. These descriptors are: polarizability index (π_I), most positive charge of acidic hydrogen atom (q^+), molecular weight (MW), most negative charge of phenolic oxygen atom (q^-), the hydrogen-bond accepting ability (ε_B) and partial charge weighted topological electronic (PCWTE) descriptors.

As can be seen, the acidity of phenols increases with increasing π_I, q^+ and MW. With increasing q^+, interactions

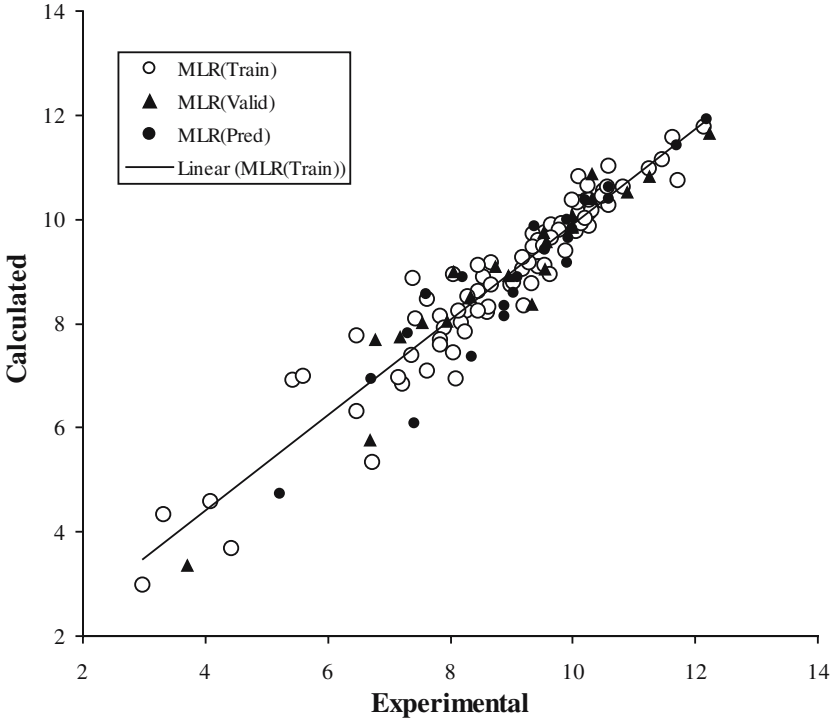

Fig. 3 Plot of the calculated values of pK_a from the MLR model versus the experimental values of it for training, validation and prediction sets

of water with the acidic hydrogen of phenols increases, so that it can be easily removed from the compounds. Polarizability and then the dipole–induced dipole interactions increase with increasing π_I and MW, as a result the acidity of phenols increases with increasing values of these descriptors [37]. The acidity constant of the compounds decrease with increasing q^-, ε_B and PCWTE descriptors because the basicity of the phenolic oxygen atom increases with increasing values of these descriptors. The effects of π_I, q^+ and MW are larger than those of the other descriptors because the standardized coefficients of π_I, q^+ and MW are higher than those of the other descriptors.

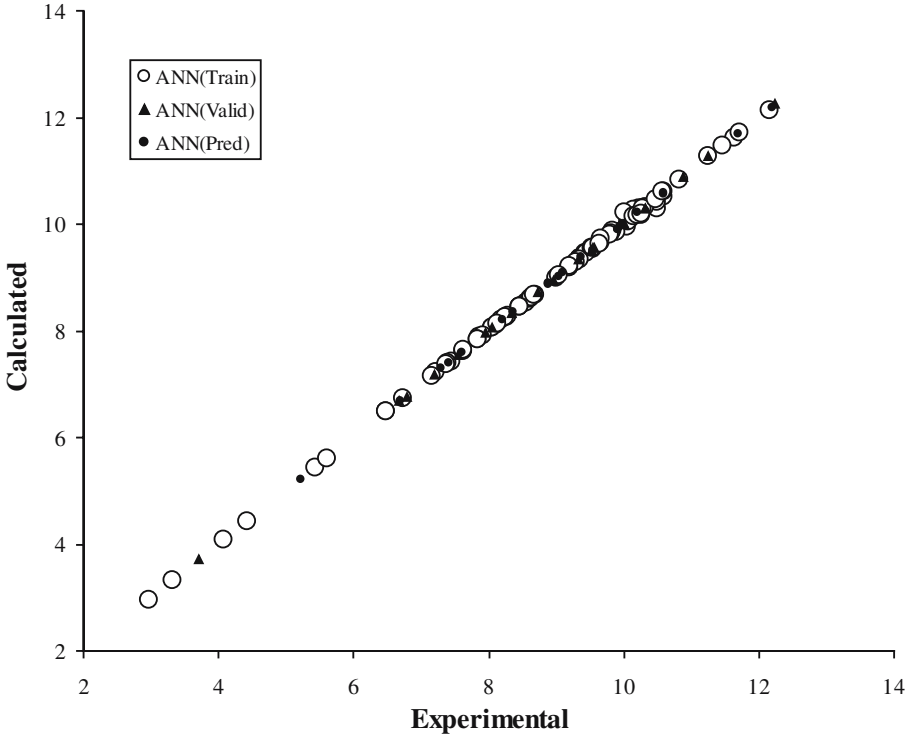

Fig. 4 Plot of the calculated values of pK_a from the ANN model versus the experimental values of it for training, validation and prediction sets

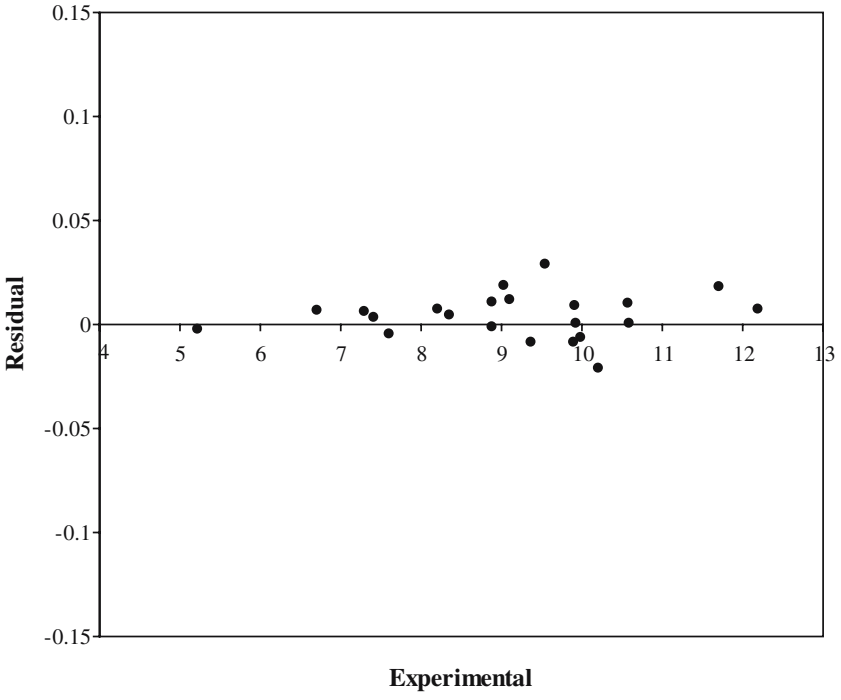

Fig. 5 Plot of the residual for calculated values of pK_a from the ANN model versus the experimental values of it for prediction set

Table 2 demonstrates that all of the descriptors are strongly orthogonal, which reflects the statistical reliability of the model.

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 [22]. 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 four to sixteen). The root mean squared 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. 1. It is clear that 14 nodes in the hidden layer is the optimum value.

This network consists of six inputs (including π_I, q^+, MW, q^-, ε_B, and PCWTE descriptors), the same descriptors in the MLR model, and one output for the pK_a values. Then an ANN with architecture 6-14-1 was generated. It is noteworthy that training the network was stopped when the RMSEV started to increases i.e. when overtraining begins. The overtraining causes the ANN to loose its prediction power [35]. Therefore, during training of the networks, it is desirable that iterations are stopped when overtraining begins. To control overtraining of the network during the training procedure, the values of RMSET and RMSEV were calculated and recorded to monitor the extent of the training in various iterations. Results obtained showed that after 800 iterations the value of RMSEV started to increase and overfitting began (Fig. 2).

The generated ANN was then trained using the training and validation sets for optimization of the weights and biases. For evaluation of the predictive power of the generated ANN, an optimized network was used to predict the pK_a values of various phenols in the prediction set, which were not used in the modeling procedure (Table 3). The calculated values of pK_a for the compounds in training, validation and prediction sets using the ANN model are plotted versus the experimental values in Fig. 3.

Table 4 Comparsion of statistical parameters obtained by the MLR and ANN models for correlation acidity constant of phenols with molecular descriptors[a]

Model	R^2_{tot}	R^2_{train}	R^2_{valid}	R^2_{pred}	$RMSE_{tot}$	$RMSE_{train}$	$RMSE_{valid}$	$RMSE_{pred}$
MLR	0.91042	0.91328	0.92313	0.89500	0.5171	0.5233	0.5048	0.5621
ANN	0.99958	0.99940	0.99996	0.99996	0.0361	0.0437	0.0123	0.0114

[a]Subscript train is referring to the training set, valid is referring to the validation set and pred is referring to the prediction set, tot is referring to the total data set, R is the correlation coefficient

As expected, the calculated values of pK_a are in excellent agreement with the experimental values. The correlation equation for all of the calculated values of pK_a from the ANN model and the experimental values is as follows:

$$\begin{aligned} &pK_a(\text{calc}) = 0.99956\, pK_a(\text{exp}) + 0.00456 \\ &(R^2 = 0.99958;\ \text{MPD} = 0.1821; \\ &\text{RMSE} = 0.0361;\ F = 298603.39) \end{aligned} \quad (4)$$

Similarly, the correlation of pK_a (calc) versus pK_a (exp) values in the prediction set gives Eq. 5:

$$\begin{aligned} &pK_a(\text{calc}) = 0.99937\, pK_a(\text{exp}) + 0.00159 \\ &(R^2 = 0.99996;\ \text{MPD} = 0.0960; \\ &\text{RMSE} = 0.0114;\ F = 454972.29) \end{aligned} \quad (5)$$

A plot of IPD for pK_a values in the prediction set versus the experimental values is illustrated in Fig. 4. As can be seen, the model dos not show proportional and systematic errors because the slope (a=0.99937) and intercept (b=0.00159) of the correlation equation are not significantly different from unity and zero, respectively, and the propagation of errors in both sides of zero is random (Fig. 5).

Table 4 compares the results obtained using the MLR and ANN models. The squared correlation coefficient (R^2) and RMSE of the models for total, training, validation and prediction sets show the potential of the ANN model for predicting pK_a values of various phenols in water.

As a result, it was found that properly selected and trained neural network can represent the dependence of the acidity constant of phenols in water on the molecular descriptors well. Then the optimized neural network could simulate the complicated nonlinear relationship between pK_a values and the molecular descriptors. The squared correlation coefficients (R^2) and RMSE of 0.8950 and 0.5621 for the prediction set by the MLR model should be compared with the values of 0.99996 and 0.0114, respectively, for the ANN model. It can be seen from Table 4 that although the parameters appearing in the MLR model are used as inputs for the ANN, the statistics show a large improvement. These improvements are due to the fact that pK_a values of phenols show non-linear correlations with the molecular descriptors.

Conclusions

A six-descriptor nonlinear computational neural network model has been developed for predicting acidity constants (pK_a) of various phenols in water using a quantitative-structure–activity relationship. Comparison of the values of RMSE for training, validation and prediction sets (and other statistical parameters in Table 4) for the MLR and ANN models show the superiority of the ANN model over the regression model. The root-mean square error of 0.5621 for the prediction set by the MLR model should be compared with the value of 0.0114 for the ANN model. Since improvement of the results obtained using a nonlinear model (ANN) is considerable, it can be concluded that the nonlinear characteristics of the molecular descriptors on the pK_a values of phenols in water is serious.

Acknowledgement The Authors wish to acknowledge the vice-presidency of research, university of Mohaghegh Ardebili, for financial support of this work.

References

1. Cronce DT, Famini GR, Soto JAD, Wilson LY (1998) J Chem Soc Perkin Trans 2:1293–1301
2. Engberts JBFN, Famini GR, Perjessy A, Wilson LY (1998) J Phys Org Chem 11:261–272
3. McClelland HE, Jurs PC (2000) J Chem Inf Comput Sci 40: 967–975
4. Hiob R, Karelson M (2000) J Chem Inf Comut Sci 40:1062–1071
5. Habibi-Yangjeh A (2004) Indian J Chem B 43:1504–1526
6. Ma Y, Gross KC, Hollingsworth CA, Seybold PG, Murray JS (2004) J Mol Model 10:235–239
7. Karelson M, Lobanov VS (1996) Chem Rev 96:1027–1043
8. Todeschini R, Consonni V (2000) Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, Germany
9. Kramer R (1998) Chemometric techniques for quantitative analysis. Marcel Dekker, New York
10. Kuzmanovski I, Aleksovska S (2003) Chemom Intell Lab Syst 67:167–174
11. Barros AS, Rutledge DN (1998) Chemomet Intell Lab Syst 40:65–72
12. Garkani-Nejad Z, Karlovits M, Demuth W, Stimpfl T, Vycudilik W, Jalali-Heravi M, Varmuza K (2004) J Chromatogr A 1028:287–295
13. Patterson DW (1996) Artificial neural networks: theory and applications. Simon and Schuster, New York
14. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design. Wiley-VCH, Weinheim
15. Agatonovic-Kustrin S, Beresford R (2000) J Pharm Biomed Anal 22:717–727
16. Fatemi MH (2002) J Chromatogr A 955:273–280
17. Bunz AP, Braun B, Janowsky R (1999) Fluid Phase Equilib 158:367–374
18. Homer J, Generalis SC, Robson JH (1999) Phys Chem Chem Phys 1:4075–4081
19. Urata S, Takada A, Uchimaru T, Chandra AK, Sekiya A (2002) J Fluorine Chem 116:163–171
20. Koziol J (2002) Internet Electron J Mol Des 1:80–93
21. Habibi-Yangjeh A, Nooshyar M (2005) Bull Korean Chem Soc 26:139–145
22. Habibi-Yangjeh A, Nooshyar M (2005) Physics and Chemistry of Liquids 43:239–247
23. Jalali-Heravi M, Masoum S, Shahbazikhah P (2004) J Magn Reson 171:176–185
24. Selassie CD, DeSoyza TV, Rosario M, Gao H, Hansch C (1998) Chemico-Biological Interaction 113:175–182
25. Gruber C, Buss V (1989) Chemosphere 19:1595–1609
26. Citra MJ (1999) Chemosphere 38:191–206
27. Schuurmann G (1996) Quant Struct Act Relat 15:121–132
28. Gross KC, Seybold PG (2001) Int J Quant Chem 85:569–579

29. Liptak MD, Gross KC, Seybold PG, Feldgus S, Shields GC (2002) J Am Chem Soc 124:6421–6427
30. Hanai T, Koizumi K, Kinoshita T (2000) J Liq Chromatogr Relat Technol 23:363–385
31. HyperChem, Release 7.0 for Windows (2002) Molecular Modeling System, Hypercube Inc
32. Todeschini R, Consonni V, Pavan M (2002) Dragon Software Version 2.1
33. Dean JA (1999) Lange's Handbook of Chemistry, 15th edn. McGraw-Hill Inc.
34. Demuth H, Beale M (2000) Neural network toolbox. Mathworks, Natick MA
35. Despagne F, Massart DL (1998) Analyst 123:157R–178R
36. Matlab 6.5. (1984–2002) Mathworks
37. Famini GR, Wilson LY (1999) J Phys Org Chem 12:645–652