Pharmaceutical Research, Vol. 22, No. 9, September 2005 (© 2005)
DOI: 10.1007/s11095-005-6246-8

Research Paper

Prediction of pK_a for Neutral and Basic Drugs Based on Radial Basis Function Neural Networks and the Heuristic Method

Feng Luan,[1] Weiping Ma,[1] Haixia Zhang,[1,3] Xiaoyun Zhang,[1] Mancang Liu,[1] Zhide Hu,[1] and Botao Fan[2]

Received March 3, 2005; accepted May 31, 2005

Purposes. Quantitative structure–property relationships (QSPR) were developed to predict the pK_a values of a set of neutral and basic drugs via linear and nonlinear methods. The ability of the models to predict pK_a was assessed and compared.
Methods. The descriptors of 74 neutral and basic drugs in this study were calculated by the software CODESSA, which can calculate constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors. Linear and nonlinear QSPR models were developed based on the heuristic method (HM) and radial basis function neural networks (RBFNN), respectively. The heuristic method was also used for the preselection of appropriate molecular descriptors.
Results. The obtained linear model had a correlation coefficient of $r = 0.884$, $F = 37.72$ with a root-mean-squared (RMS) error of 0.482 for the training set, and $r = 0.693$, F = 11.99, and RMS = 0.987 for the test set. The RMS in predicting the overall data set is 0.619. The nonlinear model gave better results; for the training set, $r = 0.886$, $F = 202.314$, and RMS = 0.458, and for the test set $r = 0.737$, $F = 15.41$, and RMS = 0.613. The RMS error in prediction for overall data set is 0.493. Prediction results from nonlinear model are in good agreement with experimental values.
Conclusions. In present study, we developed a QSPR model to predict the important parameter (pK_a) of neutral and basic drugs. The model is useful in predicting pK_a during the discovery of new drugs when experimental data are unknown.

KEY WORDS: neutral and basic drugs; quantitative structure–property relationship; radial basis function neural networks; the heuristic method.

INTRODUCTION

Before a drug can elicit any effect, it usually has to pass through at least one biological membrane by passive diffusion or by carrier-mediated uptake. Many drug molecules contain ionizable groups and hence penetrate across cell membranes, through pores and via active transport mechanism in a pK_a dependent fashion. Hence pK_a is an important factor in estimating the pharmacological behavior of drugs. Because it is not always convenient or practical to perform experimental measurements for pK_a, it is useful to develop easy-to-use and accurate models to predict pK_a values for new compounds not yet synthesized, particularly for drug discovery.

Modeling chemical and biological effects is an important objective in the fields of chemistry and pharmacology today. Chemical and biological effects are closely related to molecular properties, which can be calculated or predicted by types of methods from structure. In 1981 Perrin *et al.* (1) published a book on pK_a prediction, which is widely used but are impractical for large systems, especially for high-throughput virtual screening applications. Fragment methods have proven to be very useful and are available as commercial software. (2). *Ab initio* quantum mechanics and semiempirical quantum mechanics calculations have been used extensively (3,4). In addition, pK_a values can also be calculated by formalisms from statistical thermodynamics, which are based on numerical solutions of the Poisson–Boltzmann equation (5–7). A number of methods have also been developed for prediction of pK_a of amino acid residues in proteins in which the environmental effects are particularly important and difficult to estimate (8).

The expansion of rational techniques, in particular, the quantitative structure activity relationships (QSAR) and all its variants, i.e., quantitative structure–property relationships (QSPR), quantitative structure–retention relationship (QSRR), and finally three-dimensional (3D) or four-dimensional (4D) approaches, have recently become potential methods. Li *et al.* have reported the prediction of pK_a for both acids and bases using a novel tree structured fingerprint describing the ionizing centers (9,10). Multivariate data analysis methods [principal component analysis (PCA) and partial least squares (PLS)] are applied to the analysis of comparative molecular field analysis (CoMFA) data for several nucleic acids components by Gargallo and coworkers (11). Polański *et al.* (12) have predicted this property of benzoic and alkanoic acids by a coupled neural network and PLS system based on the comparison of molecular surfaces. Latter, the same authors conducted a systematic study of the

[1] Department of Chemistry, Lanzhou University, Lanzhou 730000, China.
[2] Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005 Paris, France.
[3] To whom correspondence should be addressed. (e-mail: liumc@lzu.edu.cn)

Table I. Compounds, Experimental and Calculated pK_a

No.	Compounds	Experimental pK_a	Calculated pK_a	
			HM	RBFNN
1	ergotamine	6.30	5.790	6.498
2	nefazodone	6.50	6.135	6.504
3	nizatidine	6.59	7.313	6.352
4	trazodone	6.79	6.766	7.350
5	mirtazapine	7.30	8.473	8.245
6	clozapine	7.63	8.245	7.779
7	domperidone	7.90	8.123	8.821
8	tolamolol	7.90	8.553	7.906
9	lidocaine	7.94	8.134	7.971
10	naloxone	7.94	8.408	8.244
11	quinidine	8.05	8.581	8.835
12	diltiazem	8.06	8.593	8.394
13	nicotine	8.10	8.813	9.182
14	perphenazine	8.11	8.780	7.918
15	butorphanol	8.19	8.458	8.482
16	codeine	8.20	8.352	8.760
17	nebivolol	8.22	8.706	8.564
18	galanthamine	8.32	8.862	8.978
19	fentanyl	8.43	8.508	8.996
20	ranitidine	8.47	8.594	8.628
21	oxycodone	8.53	8.415	8.46
22	cocaine	8.70	8.721	8.358
23	meperidine	8.70	8.629	8.426
24	timolol	8.80	8.77	8.402
25	remoxipride	8.90	8.155	8.688
26	verapamil	8.92	8.837	8.989
27	rivastigmine	8.99	8.185	8.468
28	promethazine	9.10	8.894	9.127
29	mexiletine	9.15	9.239	9.994
30	levomepromazine	9.19	9.402	8.934
31	betaxolol	9.21	9.568	9.758
32	trimipramine	9.24	9.529	9.505
33	chlorpromazine	9.25	9.040	9.063
34	chlorpheniramine	9.26	9.527	9.511
35	propafenone	9.27	9.805	9.012
36	flecainide	9.30	8.621	8.744
37	citalopram	9.38	8.638	8.889
38	clomipramine	9.38	9.215	9.385
39	labetalol	9.40	8.968	8.928
40	amitriptyline	9.40	9.888	9.671
41	propranolol	9.45	9.206	9.707
42	sumatriptan	9.50	9.090	9.569
43	venlafaxine	9.50	9.587	9.583
44	azelastine	9.54	8.815	8.791
45	pindolol	9.54	8.899	9.435
46	bisoprolol	9.57	9.673	9.723
47	alprenolol	9.60	9.451	9.743
48	acebutolol	9.67	9.699	9.103
49	nadolol	9.67	9.327	9.536
50	metoprolol	9.70	9.810	9.839
51	tacrine	9.80	9.768	9.839
52	tolterodine	9.80	9.767	9.269
53	atropine	9.84	9.614	9.367
54	terbutaline	10.00	10.250	9.971
55	atomoxetine	10.10	9.671	9.748
56	nortriptyline	10.10	10.025	9.602
57	desipramine	10.23	9.558	9.330
58	maprotiline	10.50	9.862	9.284
59	amantadine	10.68	10.181	9.881
60	cimetidine	6.97	6.626	7.119
61	sufentanil	7.85	8.744	8.788
62	clonidine	8.05	7.521	7.815
63	morphine	8.18	8.905	8.605

Table I. Continued

No.	Compounds	Experimental pK_a	Calculated pK_a	
			HM	RBFNN
64	risperidone	8.30	7.757	7.534
65	haloperidol	8.65	9.310	8.713
66	azithromycin	8.74	6.279	7.799
67	diphenhydramine	9.10	9.534	9.714
68	procainamide	9.24	7.736	8.431
69	promazine	9.28	9.122	9.120
70	imipramine	9.45	9.648	9.509
71	paroxetine	9.51	9.059	9.162
72	atenolol	9.60	8.773	9.507
73	sotalol	9.76	9.035	9.028
74	quinacrine	10.20	8.816	8.847

Nos. 1–59: training set, 60–74: test set.

performance of the 3D- and 4D-QSAR schemes in modeling steric and electronic effects on benzoic acids. They attempted to predict the pK_a values of (*o*-, *m*-, and *p*-) benzoic acids, which were divided into three subseries to simulate the different levels of steric and electronic control (13). The advantage of QSPR method lies in the fact that it can predict property based on knowledge of the chemical structure alone as soon as the model has been built. Advances in QSAR or QSPR studies have widened the scope of rationalizing drug design and the search for the mechanisms of drug actions.

In the present work, radial basis function neural networks (RBFNN) and the heuristic method (HM) were used for the prediction of pK_a values of 74 neutral and basic drugs using descriptors calculated by the software CODESSA. The HM was also used for the preselection of appropriate molecular descriptors. The principal objective was to explore the possibility of establishing an accurate QSPR model for neutral and basic drugs and to compare the performances of RBFNN and HM. The structural factors affecting the compounds' pK_a values were also investigated.

DATA SET AND MOLECULAR DESCRIPTOR GENERATION

Data Set Description

The studied compounds were a series of neutral and basic drugs whose names and pK_a values are shown in Table I. The table lists a diverse set of 74 drugs, which were taken from Lombardo *et al.* (14). The data set was randomly divided into two subsets in HM and RBFNN: a training set of 59 compounds and a test set of 15 compounds. The training set was used to build the HM and RBFNN model and the test set was used to evaluate its predictive ability in both methods.

Descriptor Generation

The two-dimensional structures of the molecules were drawn with the ISIS DRAW program (MDL Information Systems, Inc., San Leandro, CA, USA) [15]. All molecules were transferred into Hyperchem (Hypercube, Inc., Gainesville, FL, USA) and preoptimized using MM+ molecular mechanics force field (16). A more precise optimization is performed through the semiempirical PM3 method in MOPAC [17]. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient reached 0.05. The resulting geometry was then transferred into CODESSA software (18,19) (developed by the Katritzky group), which can calculate constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors, and has been successfully used in various QSPR and QSAR researches. Constitutional descriptors are related to the number of atoms and bonds in each molecule. Topological descriptors include valence and nonvalence molecular connectivity indices calculated from the hydrogen-suppressed formula of the molecule, encoding information about the size, composition, and the degree of branching of a molecule. Geometrical descriptors are calculated from 3D atomic coordinates of the molecule. These descriptors comprise moments of inertia, shadow indices, molecular volume, molecular surface area, and gravitation indices. Electrostatic descriptors reflect characteristics of the charge distribution of the molecule. Quantum chemical descriptors include information about binding and formation energies, partial atom charge, dipole moment, and molecular orbital energy levels. In the present investigation, about 700 descriptors were provided.

MATERIALS AND METHODS

Heuristic Method

As soon as molecular descriptors are generated, CODESSA uses the heuristic method to preselect descriptors and build the linear model (19–21). Its advantages are is high speed and the absence of software restrictions on the size of the data set. The heuristic method can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. Besides, it can demonstrate which descriptors have bad or missing values, which descriptors are insignificant (from the standpoint of a single-parameter correlation), and which descriptors are highly intercorrelated. The heuristic method of the descriptor selection proceeds with a preselection of descriptors by eliminating (1) descriptors that are not available for each structure, (2) descriptors having a small

variation in magnitude for all structures, (3) descriptors that give an F test's value below 1.0 in the one-parameter correlation, and (4) descriptors whose t values are less than the user-specified value, etc. This procedure orders the descriptors by the decreasing correlation coefficient when used in one-parameter correlations. As a next step, the program calculates the pair correlation matrix of descriptors and further reduces the descriptor pool by eliminating highly correlated descriptors. After the preselection of descriptors, multiple linear regression (MLR) models are developed in a stepwise procedure. Thus, descriptors and correlations are ranked according to the values of the F test and the correlation coefficient. Starting with the top descriptor in the list, two-parameter correlations are calculated.

In the following steps, new descriptors are added one by one until the preselected number of descriptors in the model is achieved. The final result is a list of the ten best models according to the values of the F >test and correlation coefficient. The goodness of the correlation is tested via coefficient regression (r^2), F test (F), and standard deviation (s^2).

Theory of Radial Basis Function Neural Networks

The theory of RBFNN has been extensively presented in some papers (22,23). Here only a brief description of the RBFNN principle is given. Figure 1 shows the basic network architecture. It consists of an input layer, a hidden layer, and an output layer. The input layer does not process the information; it only distributes the input vectors to the hidden layer. The hidden layer of RBFNN consists of a number of RBF units (n_h) and bias (b_k). Each hidden layer unit represents a single radial basis function, with associated center position and width. Each neuron on the hidden layer employs a radial basis function as a nonlinear transfer function to operate on the input data. The most widely used RBF is a Gaussian function characterized by a center (c_j) and a width (r_j). The RBF functions by measuring the Euclidean distance between the input vector (x) and the radial basis function center (c_j), and performs the nonlinear transformation with RBF in the hidden layer as given below

$$h_j(x) = exp\left(-\|x - c_j\|^2 / r_j^2\right) \quad (1)$$

in which h_j is the notation for the output of the jth RBF unit. For the jth RBF, c_j and r_j are the center and the width, respectively. The operation of the output layer is linear, which is given below

$$y_k(x) = \sum_{j=1}^{n_k} w_{kj} h_j(x) + b_k \quad (2)$$

where y_k is the kth output unit for the input vector $\boldsymbol{x}$, w_{kj} is the weight connection between the kth output unit and the jth hidden layer unit, and b_k is the bias. It can be seen from Eqs. (1) and (2) that designing an RBFNN involves selecting centers, number of hidden layer units, widths, and weights. There are various ways of selecting the centers, such as random subset selection, K-means clustering, orthogonal least squares learning algorithm, RBF-PLS, etc. The width of the radial basis function networks can either be chosen to bear the same/different value for all/each unit(s). In this

Table II. Descriptors, Coefficients, Standard Error, and t Test Values for the Linear Model

No.	Descriptor	Coefficient	Standard error	t Test
0	Intercept	13.186	0.406	32.48
1	Relative number of N atoms	−15.970	2.101	−7.600
2	Randic index (order 3)	−0.302	0.031	−9.751
3	RNCG relative negative charge (QMNEG/QTMINUS) [Quantum-Chemical PC]	−12.703	1.942	−6.543
4	RNCS Relative negative charged SA (SAMNEG * RNCG) [Zefirov's PC]	0.127	0.028	4.499
5	Max net atomic charge	0.758	0.023	3.245

N = 59, r = 0.882, F = 37.72, RMS = 0.482.

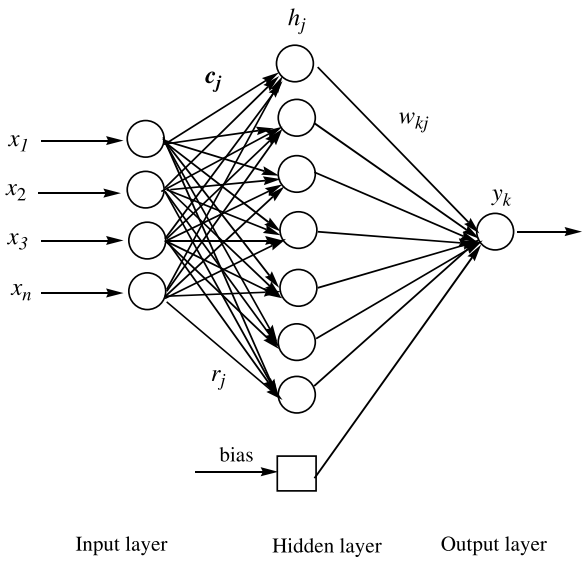

Fig. 1. The structure of radial basis function neural networks.

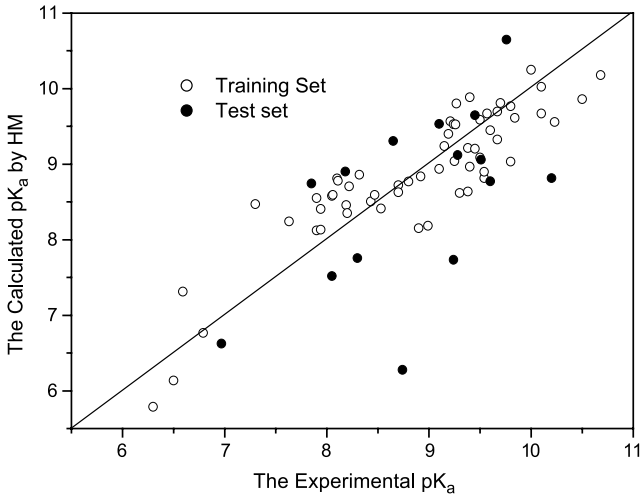

Fig. 2. Predicted vs. experimental pK_a by heuristic method.

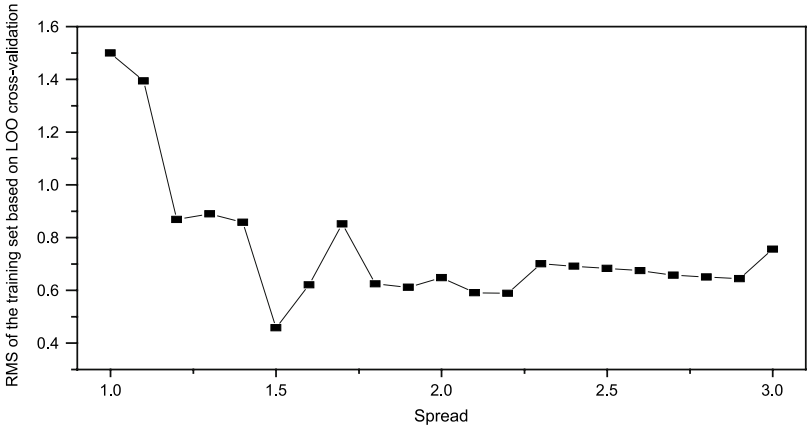

Fig. 3. The spread *vs.* root-mean-squared error of training set based on leave-one-out cross-validation.

paper, considerations were limited to Gaussian functions with a constant width, which was the same for all units. A forward subset selection routine was used to select the centers from training set samples. The adjustment of the connection weight between hidden layer and output layer is performed using a least-squares solution after the selection of centers and width of radial basis functions.

The overall performance of RBFN is evaluated in terms of a root-mean-squared (RMS) error according to the equation below

$$RMS = \sqrt{\frac{\sum_{i=1}^{n_{k_s}} (y_k - \hat{y}_k)^2}{n_k}}$$

where y_k is the desired output, $\hat{y}_k$ is the actual output of the network, and n_k is the number of compounds in analyzed set. The performance of RBFNN is determined by the values of the following parameters: the number n_h of radial basis functions, the center c_j and the width r_j of each radial basis function, the connection weight w_{kj} between the jth hidden layer unit and the kth output unit. The centers of RBFNN are determined via the forward subset selection method proposed by Orr [24,25]. The optimal width was determined by experiments with a number of trials by taking into account the leave-one-out (LOO) cross-validation error. The one that gives a minimum LOO cross-validation error is chosen as the optimal value. After the selection of the centers and number of hidden layer units, the connection weights can be easily calculated via linear least-squares technique.

All calculation programs implementing RBFNN were written in M-file based on a basis MATLAB script for RBFNNs. The scripts were run on a Pentium IV PC with 256 M RAM.

RESULTS AND DISCUSSION

Results of the Heuristic Method

The heuristic method was used to develop the linear model for the prediction of pK_a using all the descriptors.

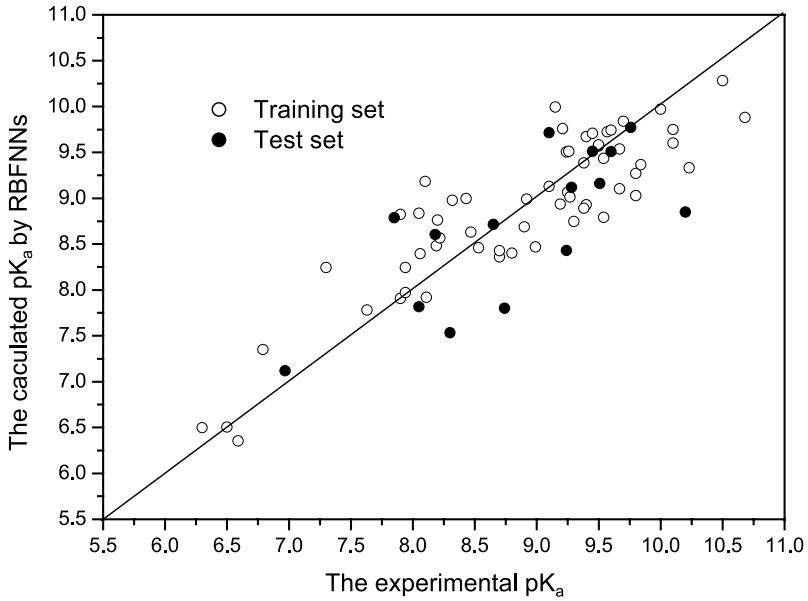

Fig. 4. Predicted *vs.* experimental pK_a by radial basis function neural networks.

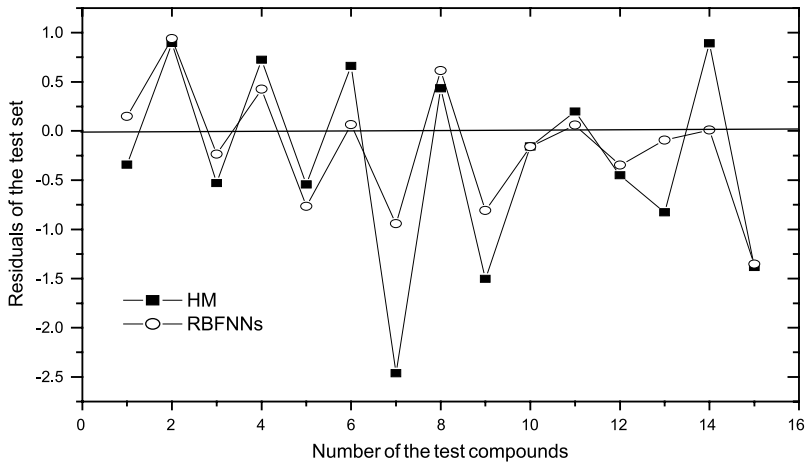

Fig. 5. Comparative residuals *vs.* no. of compounds in the test sets for HM and RBFNN models.

After the heuristic reduction, the pool of descriptors was reduced from 700 to 198. To determine the optimum number of descriptors, various subset sizes were investigated. When adding another descriptor did not improve significantly the statistics of a model, it was determined that the optimum subset size had been achieved. In the present study, five descriptors were eventually selected. A detailed description of the linear model based on compounds in the training set is summarized in Table II.

In the selected linear model, there is a constitutional descriptor: relative number of N atoms (RNN); a topological descriptor: Randic index (order 3); two electrostatic descriptors: relative negative charge (QMNEG/QTMINUS) (Quantum-Chemical PC) (RNCG) and relative negative charged SA (SAMNEG * RNCG) [Zefirov's PC] (RNCS); and one quantum chemical descriptor: max net atomic charge. With the test set, the prediction results were obtained, the statistical parameters were $r = 0.693$, $F = 11.99$, and RMS = 0.987.And the heuristic model produced an RMS error of 0.619 for the whole data set. The predicted vs. observed pK_a based on HM is shown in Table I. Figure 2 shows the predicted vs. observed pK_a values for all of the 74 compounds studied (the training set and the test set).

Results of RBFNNs

After the establishment of a linear model, RBFNNs is used to develop a nonlinear model based on the same subset of descriptors. To obtain better results, the parameters influencing the performance of RBFNN were optimized. The selection of the optimal width value for RBFNN was performed by systemically changing its value in the training step. The value giving the best leave-one-out cross-validation result was used in the model. For this data set, the optimal spread was determined as 1.5 (see Fig. 3.). The corresponding number of centers (hidden layer nodes) of RBFNN is 16. The predicted results of the nonlinear models are shown in Table I and Fig. 4. The obtained model had a correlation coefficient $r = 0.886$, $F = 202.314$, and an RMS error of 0.458 for the training set. The statistical parameters of the test set were 0.737, $F = 15.41$, and RMS = 0.613. The root mean square error in prediction for overall data set is 0.493. Comparative residuals vs. number of compounds in the test sets for HM and RBFNN models are shown in Fig. 5. As shown in Fig. 5, RBFNN performed slightly better than HM.

Discussion of the Input Descriptors

By interpreting the descriptors in the model, it is possible to gain some insight into factors that are likely to influence the pK_a values of these compounds. The relative number of N atoms (RNN) is a constitutional descriptor calculated as the number of N atoms divided by the number of atoms. RNN affects the density of the electron cloud of the molecule. The larger the RNN is, the higher the density of the electron cloud of the molecule becomes as well as the polar separation of positive and negative electric charge in molecular. Thus, an increase in this descriptor leads to a decrease in pK_a of the compound.

Randic index (order 3), which encodes the size, shape, and degree of branching in the compound, also relates to the dispersion interaction among molecules. The larger the molecular size is, the stronger the dispersion interaction becomes. Because of its negative coefficient in the linear model, increasing this descriptor also decreases the pK_a values, indicating that dispersion interaction favors the separation of hydrogen ion from nitrogen atom.

Relative negative charge (RNCG) and relative negative charged SA (RNCS) are electrostatic descriptors. RNCG is defined as the ratio of the maximum (by absolute value) atomic partial negative charge and the sum of similar negative charges in the molecule, whereas RNCS is defined as the solvent-accessible surface area of the most negative atom divided by the relative negative charge. They both represent or directly depend on the quantum-chemically calculated charge distribution in the molecules. Charge distribution is an important factor influencing the polarization of the molecule, which leads to the different dissociation ability of proton. The negative coefficient of RNCG in the model implies that increasing the value of this descriptor can lead to a smaller pK_a value. For RNCS, the larger the

solvent-accessible surface area of the most negative atom is, the lower is the chance for positive ion to replace proton and the larger pK_a value becomes.

This model also contains a quantum chemical descriptor, max net atomic charge (q_{max}). q_{max}, which is obtained from the Mulliken charge distribution scheme of quantum chemical calculations, represents or directly depends on the quantum-chemically calculated charge distribution in the molecules and can also account for the polar interaction of molecule. It receives a positive coefficient in the linear model, indicating that max net atomic charge of the molecule leads to larger pK_a values.

Analysis of the results obtained indicated that the models we proposed correctly represent the structure–property relationships of these compounds, and that molecular descriptors calculated solely from structures can represent the structural features of the compounds responsible for their pK_a values.

CONCLUSION

Quantitative structure–property relationship models that can predict the pK_a values of neutral and basic drugs were developed in this study. The proposed linear model could identify and provide some insights into what structural features are related to the pK_a values for this type of compounds. Nonlinear RBFNN model based on the same sets of descriptors showed better predictive ability. Using the model established, we can predict the pK_a value of neutral and basic drugs whether they have been synthesized or not. Consequently, it is a very helpful tool in designing new drugs, especially for early drug discovery.

ACKNOWLEDGMENTS

The authors thank the National Natural Science Foundation of China (NSFC) Fund (NO.20305008) for supporting this project.

REFERENCES

1. D. D. Perrin, B. Dempsey and E. P. Serjeant. *pKa prediction for organic acids and bases*, Chapman and Hall, New York, 1981.
2. A. T. Santili-Kakoulidou, I. Panderi, F. C. Sizmadia, and F. Darvas. Prediction of distribution coefficient from structure 2. Validation of Prolog D, an expert system. *J. Pharm. Sci.* **86**:1173–1179 (1997).
3. C. O. da Silva, E. C. da Silva, and M. A. C. Nascimento. Ab Initio calculations of absolute pK_a values in aqueous solution I. carboxylic acids. *J. Phys. Chem. A* **103**:11194–11199 (1999).
4. M. J. Citra. Estimating the pK_a of phenols, carboxylic acids and alcohols from semiempirical quantum chemical methods. *Chemosphere* **38**:191–206 (1999).
5. H. Oberoi and N. M. Allewell. Multigrid solution of the nonlinear Poisson–Boltzmann equation and calculation of titration curves. *Biophys. J.* **65**:48–55 (1993).
6. J. Antosiewicz, J. A. McCammon, and M. K. Gilson. Prediction of pH dependent properties of proteins. *J. Mol. Biol.* **238**:415–436 (1994).
7. Y. Y. Sham, Z. T. Chu, and A. Warshel. Consistent calculations of pK_a's of ionizable residues in proteins: semi-microscopic and microscopic approaches. *J. Phys. Chem. B* **101**:4458–4472 (1997).
8. J. Warwicker. Simplified methods for pK_a and acid pH-dependent stability estimation in proteins: removing dielectric and counterion boundaries. *Protein Sci.* **8**:418–425 (1999).
9. X. Li and C. G. Robert. Novel methods for the prediction of log*P*, pK_a, and log*D*. *J. Chem. Inf. Comput. Sci.* **42**:796–805 (2002).
10. X. Li, C. G. Robert, and D. C. Robert. Predicting pK_a by molecular tree structured fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **43**:870–879 (2003).
11. R. Gargallo, C. A. Sotriffer, R. L. Klaus, and M. R. Bernd. Application of multivariate data analysis methods to Comparative Molecular Field Analysis (CoMFA) data: Proton affinities and pK_a prediction for nucleic acids components. *J. Comput. -Aided Mol. Des.* **13**:611–623 (1999).
12. J. Polański, R. Gieleciak, and A. Bark. The Comparative Molecular Surface Analysis (COMSA)—a nongrid 3D QSAR method by a coupled neural network and PLS system: predicting pK_a values of benzoic and alkanoic acids. *J. Chem. Inf. Comput. Sci.* **42**:184–191 (2002).
13. J. Polanski and A. Bak. Modeling steric and electronic effects in 3D- and 4D-QSAR schemes: predicting benzoic pK_a values and steroid CBG binding affinities. *J. Chem. Inf. Comput. Sci.* **43**:2081–2092 (2003).
14. F. Lombardo, R. Scott Obach, M. Y. Shalaeva, and F. Gao. Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics. *J. Med. Chem.* **47**:1242–1250 (2004).
15. ISIS Draw2.3, MDL Information Systems, Inc., 1990–2000.
16. HyperChem 4.0, Hypercube, Inc., 2000.
17. MOPAC, v.6.0 Quantum Chemistry Program Exchange, Program 455, Indiana University, Bloomington, IN.
18. A. R. Katritzky, V. S. Lobanov and M. Karelson. *CODESSA: Training Manual*, University of Florida, Gainesville, FL, 1995.
19. A. R. Katritzky, V. S. Lobanov and M. Karelson. *CODESSA: Reference Manual*, University of Florida, Gainesville, FL, 1994.
20. M. R. M. Oblak and T. Solmajer. Quantitative structure–activity relationship of flavonoid analogues. 3. Inhibition of $p56^{lck}$ protein tyrosine kinase. *J. Chem. Inf. Comput. Sci.* **40**:994–1001 (2000).
21. F. Luan, C. X. Xue, R. S. Zhang, C. Y. Zhao, M. C. Liu, Z. D. Hu, B. T. Fan. Prediction of retention time of a variety of volatile organic compounds based on the heuristic method and support vector machine. *Anal. Chim. Acta.* **537**:101–110 (2005).
22. X. J. Yao, A. Panaye, P. Doucet, R. S. Zhang, H. F. Chen, M. C. Liu, Z. D. Hu, and B. T. Fan. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. *J. Chem. Inf. Comput. Sci.* **44**:1257–1266 (2004).
23. Y. H. Xiang, M. C. Liu, X. Y. Zhang, R. S. Zhang, Z. D. Hu, B. T. Fan, J. P. Doucet, and A. Panaye. Quantitative prediction of liquid chromatography retention of *N*-benzylideneanilines based on quantum chemical parameters and radial basis function neural network. *J. Chem. Inf. Comput. Sci.* **42**:592–597 (2002).
24. M. J. L. Orr. *Introduction to Radial basis function networks,* Centre for Cognitive Science, Edinburgh University, 1996.
25. M. J. L Orr. *MATLAB routines for subset selection and ridge regression in linear neural networks,* Centre for Cognitive Science, Edinburgh University, 1996.