Asian Journal of Chemistry; Vol. 24, No. 9 (2012), 3842-3848

ASIAN JOURNAL OF CHEMISTRY

www.asianjournalofchemistry.co.in

Prediction of the Odor Thresholds of Oxygen and Nitrogen Containing Heterocycles Using the Quantitative Structure-Property Relationship Approach

XUAN XU, FENG LUAN*, HUITAO LIU and YUAN GAO

Department of Applied Chemistry, Yantai University, Yantai 264005, P.R. China

*Corresponding author: Fax: +86 535 6902063; Tel: +86 535 6902063; E-mail: fluan@sina.com

(*Received*: 25 May 2011; *Accepted*: 14 April 2012) AJC-11252

Quantitative structure-property relationship (QSPR) models are developed to correlate the odor thresholds of 50 oxygen and nitrogen containing heterocycles from their molecular structures. Three statistic methods including multiple linear regression (MLR), non-linear radial basis function neural network (RBFNN) and support vector machine (SVM) are performed to build the models. A six-descriptor equation with the squared correlation coefficient (R^2) of 0.8012 and root mean square error (RMS) of 1.0011 were obtained for the training set and R^2 = 0.648, RMS = 1.7165 for the external test set. The radial basis function neural network model gave better results: R^2 = 0.8767, RMS = 0.7165 for the training set and R^2 = 0.7746, RMS = 1.3570 for the external test set. The SVM model gave similar results to multiple linear regression, that is, R^2 = 0.8023, RMS = 0.9271 for the training set and R^2 = 0.7033 and RMS = 1.5888 for the test set. The aim of the paper is to provide an easy, direct and relatively accurate way to estimate the odor thresholds.

Key Words: Odor thresholds, Quantitative structure-property relationship, Multiple linear regression, Radial basis function neural network, Support vector machines.

INTRODUCTION

In the flavour and fragrance industry, heterocyclic compounds are of great interest because of their widespread occurrence in food flavours and their valuable organoleptic characteristics. Even though these heterocyclic aroma chemicals are found only in tiny amounts in foods, their powerful odors and low odor thresholds make them key in boosting flavours and fragrances. The main heterocyclic aroma chemicals are almost three groups, which include oxygen-, sulfur- and nitrogen containing rings, respectively. The oxygen-containing heterocyclic aroma chemicals belong to the oxirane, furan, pyran and oxepine groups. The sulfur-containing aroma chemicals belong to the thiophene family and, together with nitrogen, to the thiazole and dithiazine systems. The nitrogen-containing aroma chemicals belong to pyrrole, indole, pyridine, quinoline, pyrazine and quinoxaline systems and, together with sulfur, as mentioned above, the thiazole and dithiazine families[1].

Over the years, a lot of work has been done to explore the mysteries of the sense of odors since humans and animals can sense by smell different odors. As is well known, both of the theory of stereochemistry and the method of structure-functional groups are widely supported. They hold the view that odors are not only associated with the shape or the size, but also with the nature and content of functional groups and their location in the whole molecule. So, to understand the human chemosensory perception, it is necessary not only to have the knowledge of the relevant structural and physicochemical properties of chemicals, but also necessary to explore the relationships between the characteristics of compounds and their properties. Among the different types of properties, odor thresholds value is an important biological property of odorant molecules. Further, structure-odor relationships are of great help to screen out new fragrance or synthesize new ones artificially.

Until now, there have been a number of studies on the correlations of odor detection thresholds (ODT) with various properties of odorant. The study by Laffort and Patte[2] was the first to employ a physicochemical analysis of these compounds. Mihara and Masuda[3] used a two-term regression equation to model the logarithm of the odor threshold of 60 di-substituted pyrazines. Seeman *et al.*[4] studied the odor profile of structurally similar pairs of 1,3-dialkylbenzenes and 2,6-dialkylpyridines as a function of the accessibility of the nitrogen atom and steric hindrance. Winter related the activity of a series of ambergris-type odorants to a minimum accessible surface area about the ether oxygen in the molecule[5]. Edwards and Jurs[6] also used discriminant analysis to study the ability of odorant molecules to stimulate activity of the enzyme adenylate cyclase. Latter,

computer assisted statistical methods have been used to study the odor thresholds of two sets of odor active molecules by the same authors. One data set included 53 aliphatic alcohols. The other data set included 74 mono and di-substituted pyrazine derivatives[7]. Chastrette[8] has reviewed the former works up to 1996. Then, Yamanaka[9] showed that odor thresholds for several homologous series could be correlated with the odorant activity coefficient in water. Abraham *et al.*[10] performed a model for odor thresholds for a series of 64 compounds, including esters, aldehydes, ketones, alcohols, carboxylic acids, aromatic hydrocarbons, terpenes and some of other volatile organic compounds. Ivanciuc[11] has investigated the application of support vector machines (SVM) to the classification of 98 tetra-substituted pyrazines by five theoretical descriptors. Hau *et al.*[12] studied the odor thresholds of volatile organic compounds by QSAR approach. Tan and Siebert[13] also gave a QSAR study on flavour thresholds in beer of different organic compounds such as alcohol, ester, aldehyde and ketone. The aroma quality and the threshold values of some pyrazines was predicted using artificial neural networks by Wailzer *et al.*[14]. Latter, threshold of pyrazine derivatives were also studied by Zakarya *et al.*[15]. In our previous study, we have given a QSPR study on 74 pyrazine derivatives using different statistical methods, such as MLR, RBFNN and SVM[16] and we also performed a classification study of the fragrant properties of chemical compounds based on the support vector machine and linear discriminant analysis[17].

All of the former studies prompted us to go on carrying out a theoretical study on the odor threshold of the very important oxygen and nitrogen containing heterocyclic compounds in the flavour and fragrance industry. To the best of our knowledge, there are no general QSPR studies on this topic of these special kinds of compounds. The aim of the present work is to devise quantitative structure-property relationships that could be used to correlate odor thresholds with relevant physicochemical properties and thereby to perform prediction of such thresholds. The structural factors affecting the compounds' odor thresholds values are also investigated.

EXPERIMENTAL

Data set: The experimental value of the odor thresholds is not so many. The data set of the 50 oxygen and nitrogen containing heterocycles was collected from a handbook[18]. Concentration unit of the experimental odor threshold is ppm and it is by volume. Of these compounds, 32 are nitrogen-containing and 18 are oxygen-containing heterocycles. A complete list of the compounds and their corresponding odor thresholds is in Table-1. As usually did by QSPR study, the entire set of compounds was divided into two subsets: a training set, whose information was used to build the models and an external test set, consisting of molecules not found in the training set, which was used to validate the models once they were built. Members of each set were assigned randomly. In Table-1, the serial number 1-26 in the training set and 1-6 in the external test set are nitrogen-containing heterocycles; the serial number 27-40 in the training set and 7-10 in the external test set are oxygen-containing ones. The training set consisted of 40 compounds (80 %) and the test set contained 10 compounds (20 %). In addition, each set contained roughly the same percentage of oxygen-containing compounds (training set = 25.0 %, test set = 28.6 %).

Molecular structure optimization and descriptor generation: To obtain a QSPR model, the compounds must be represented by molecular descriptors that retain as much structure information as possible. Here five classes of descriptors *i.e.*, constitutional, topological, geometrical, electrostatic and quantum chemical descriptors, were calculated. The descriptors were generated as follows: The compounds were drawn using ISIS Draw 2.4[19] and pre-optimized using the molecular mechanics force field method (MM+) available in HyperChem 7.0[20]. The molecular structures were then optimized using the Polak-Ribiere algorithm until the root mean square gradient was equal to or less than 0.001. A more precise optimization was done with a semi-empirical PM3[21] method in MOPAC[22]. Thereafter, CODESSA PRO[23,24] was used to calculate the above five types of molecular descriptors. Altogether, 480 descriptors were calculated for each of the 50 heterocyclic compounds studied.

Selection of molecular descriptors: A successful QSPR model depends on suitable descriptors selection. If molecular structures are represented by improper descriptors, they will not lead to reasonable predictions. The process of features selection entails pruning the descriptors pool through the heuristic method (HM) available in the framework of the CODESSA program[23,24]. 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 will demonstrate which descriptors have bad or missing values, or are insignificant (from the standpoint of a single-parameter correlation) or are highly inter-correlated. The detailed discussion about the heuristic method can be found in Ref[24]. Here, only the main steps of this method are given in the following: The heuristic method of the descriptor selection proceeds with a pre-selection of descriptors by eliminating those descriptors that are not available for each structure; descriptors having a small variation in magnitude for all structures; descriptors that give a F-test's value below 1.0 in the one-parameter correlation; and descriptors whose *t*-values are less than the user-specified value, *etc*. This procedure orders the descriptors by decreasing correlation coefficient when used in one-parameter correlations. Following the pre-selection of descriptors, multiple linear regression (MLR) models are developed in a stepwise procedure.

Methodology of modeling

Theory of MLR and RBFNN: MLR analysis and RBFNN artificial neural networks were used to correlate the descriptors and the odor thresholds values of the 50 heterocyclic compounds. The forward stepwise multiple regression analysis, a commonly used method in QSPR study, was employed to establish the quantitative regression models[25,26]. The general purpose of it is to obtain a mathematical function (eqn. 1) that best describes the desired activity, Y, as a linear combination of the X variables (the molecular descriptors), with the regression coefficients bn. Such coefficients are to be optimised by means of MLR analysis using the chosen training set compounds.

TABLE-1
EXPERIMENTAL AND CALCULATED log T FOR TRAINING AND TEST SET

No.	Name	Experimental	Calculated log T		
	Training set	log T	MLR	RBFNN	SVM
1	4-Butyl-5-propylthiazole	-2.523	-2.1894	-2.0624	-1.6960
2	2-Ethyl-3-methoxypyrazine	-0.398	0.2316	0.0455	0.2291
3	2-Pentylpyridine	-0.222	0.2800	0.1294	0.3721
4	1-(4,5-Dihydrothiazol-2-yl)ethanone	0.114	-0.2674	0.0408	-0.0501
5	2-Isobutylthiazole	0.301	-0.2069	-0.1138	-0.1082
6	2-(Methylthio)benzo[d]thiazole	0.699	0.0518	0.5699	-0.1273
7	2,6-Diethylpyrazine	0.778	2.3299	2.0988	2.3345
8	2-Isobutyl-5-methoxypyrazine	1	0.6085	0.9025	0.8112
9	2-Isopropyl-5-methoxypyrazine	1	0.8946	1.037	0.9701
10	2-Ethyl-3-methoxypyrazine	1.041	1.3735	0.8436	1.2604
11	Pyridine	1.477	0.9529	1.4045	0.6510
12	2-Isobutyl-3-methylpyrazine	1.544	1.3633	0.7458	1.3498
13	2-Methyl-5-vinylpyridine	1.602	0.9217	0.3753	0.7950
14	2-Ethyl-5-methylpyrazine	2	2.6954	3.2579	2.7642
15	2-Ethyl-3-methylpyrazine	2.114	1.6747	1.4365	1.6059
16	2,3-Dimethylpyrazine	2.602	2.3363	2.2427	2.1992
17	2,3,5-Trimethylpyrazine	2.602	1.2942	1.368	1.3973
18	2-Methoxypyrazine	2.845	4.187	4.6367	3.7426
19	2,3,5,6-Tetramethylpyrazine	3	2.3939	2.8236	2.1744
20	2-Pentylpyrazine	3	3.3453	2.8891	3.2499
21	2,5-Dimethylpyrazine	3.255	3.7805	3.6651	3.6848
22	1-Ethyl-1*H*-pyrrole-2-carbaldehyde	3.301	2.5759	3.4494	2.4750
23	2-Ethylpyrazine	4.778	4.1451	3.981	3.8563
24	2-Methylpyrazine	5	4.6999	4.5243	4.3308
25	5-Methyl-1*H*-pyrrole-2-carbaldehyde	5.041	4.4891	3.9481	4.2452
26	Pyrazine	5.699	5.5040	5.8141	4.8823
27	(R)-4-Hydroxy-2,5-dimethylfuran-3(2*H*)-one	-1.4	0.9137	-0.4008	1.2062
28	2-(Methylthio)furan	-1.4	0.9070	-0.3298	0.6039
29	3-Methyl-2-vinylfuran	-0.3	1.1665	0.7579	1.2025
30	3,4-Dimethylthiophene	0.11	-0.2266	-0.1753	-0.4873
31	2-Vinylfuran	3	4.0805	3.5843	3.8263
32	2-Pentylfuran	3.18	2.0622	3.2531	2.3531
33	4,4-Dimethyl-1,3-dioxolane	3.4	3.6946	3.1843	3.0484
34	(Furan-2-yl)methanol	3.48	2.8324	3.1658	2.7804
35	2-Propylfuran	3.78	2.8309	3.4524	2.9939
36	2-Ethylfuran	3.9	2.8358	2.9733	2.9171
37	5-Methylfuran-2-carbaldehyde	4	3.4723	4.1866	3.1725
38	3-Hydroxy-2-methyl-4*H*-pyran-4-one	4.54	2.8506	4.001	2.8936
39	5-(Hydroxymethyl)furan-2-carbaldehyde	5.3	5.6319	4.5598	5.2853
40	1-(Furan-2-yl)-2-hydroxyethanone	5.3	6.0226	6.2916	5.6291
	Test set				
1	2-Hexyl-3-methoxypyrazine	-3	0.1119	-0.7148	0.3403
2	3-Isobutyl-2-methoxy-5-methylpyrazine	0.415	-0.8316	-1.1538	-0.5038
3	2,5-Diethylpyrazine	1.301	2.6981	2.9093	2.7731
4	2-Methoxy-3-methylpyrazine	2.176	2.0006	2.0456	1.8304
5	2,6-Dimethylpyrazine	3.176	2.5975	2.5491	2.4831
6	1-(1*H*-Pyrrol-2-yl)ethanone	5.301	5.4257	5.8248	4.9989
7	4-Methoxy-2,5-dimethylfuran-3(2*H*)-one	-1.52	0.8823	-0.0241	0.8530
8	1,3,3-Trimethyl-2-oxa-bicyclo[2.2.2]octane	1.08	4.2629	2.8822	3.1557
9	2-Methylfuran	3.54	3.5674	2.9453	3.5502
10	1-(Furan-2-yl)ethanone	5.04	4.9649	6.3127	4.3729

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n \quad (1)$$

The theory of RBFNN has been extensively presented in some works[27,28]. Here, only a brief description of the RBFNN principle was given. Fig. 1 shows the basic network architecture. It consists of an input, a hidden 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 often used RBF is a Gaussian function that is characterized by a center (c_j) and a width (r_j). The RBF function performs the nonlinear transformation by measuring the Euclidean distance between the input vector (x) and the radial basis function center (c_j). The RBF in the hidden layer as given below:

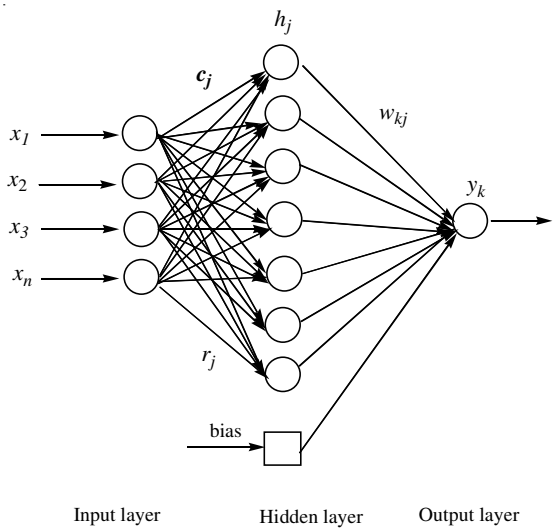

Fig. 1. Architecture of RBFNN

$$h_j(X) = \exp\left(-\frac{\| X - c_j \|^2}{r_j^2}\right) \quad (2)$$

In this equation, h_j is the notation for the output of the j^{th} RBF unit. For the j^{th} RBF, c_j and r_j are the center and the spread, respectively. The operation of the output layer is linear, which is given below

$$y_k(X) = \sum_{j=1}^{n_k} w_{kj} h_j(X) + b_k \quad (3)$$

where y_k is the k^{th} output unit for the input vector x, w_{kj} is the weight connection between the k^{th} output unit and the j^{th} hidden layer unit and b_k is the bias. It can be seen from eqns. 2 and 3, designing a RBFNN involves the selection of centers, number of hidden layer units, width and weights. There are various ways for selecting the centers, such as random subset selection, K-means clustering, orthogonal least squares learning algorithm, RBF-PLS, *etc*. The widths of the radial basis function networks can either be chosen the same for all the units or can be chosen differently for each unit. In this paper, considerations were limited to the Gaussian functions with a constant width, which was the same for all units. 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 RBFNN is evaluated in terms of a root-mean-squared error (RMS) according to the equation below:

$$RMS = \sqrt{\frac{\sum_{i=1}^{n_k} (y_k - \hat{y}_k)^2}{n_k}} \quad (4)$$

where y_k is the desired output and $\hat{y}_k$ is the actual output of the network; n_k is the number of compounds in analyzed set. The performance of RBFNN is determined by the values of 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 j^{th} hidden layer unit and the k^{th} output unit. The centers of RBFNN are determined with the forward subset selection method proposed by Orr[29,30]. 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.

Theory of support vector machines: Support vector machines (SVM) are gaining popularity due to many attractive features and promising empirical performance[31]. It can solve high-dimension problems and therefore avoid the "curse of dimensionality". A detailed description of the theory of SVM can be referred in several excellent books and tutorials[32,33]. The basic idea and its performance are simply introduced here. SVM are generated by a two-step procedure: first, the sample data vectors are mapped to a very high-dimensional space. The dimension of this space is significantly larger than that of the original data space. Then, the SVM algorithm finds a hyperplane in this space with the largest margin separating classes of data. SVM can also be applied to regression by the introduction of an alternative loss function. The decision function of regression is as follows:

$$f(x) = \left(\sum_{i=1}^{l} y_i \alpha_i k(x, x_i) + b\right) \quad (5)$$

The overall performances of SVM were also evaluated in terms of root mean square error (RMS), which was showed above, eqn. 4.

For each model, the goodness of the fit was assessed by examining the determination coefficient (R^2), the adjusted determination coefficient (R^2_{adj}), Fisher's statistics (F) as well as the standard deviation (s^2)[34]. The robustness of the models was evaluated by means of internal cross-validation (CV), specifically by the leave-one-out (LOO) and the leave-n-out (LNO) techniques[35]. The estimated measure of the predictive ability of the model was determined also by the R^2 and s^2 and F values. This procedure was implemented in the MATLAB software. In addition, the ratio between the number of compounds in the training set and the number of adjustable parameters in the model[36], known as the ρ statistics was added.

RESULTS AND DISCUSSION

Results of MLR: The MLR was used to develop the linear model for the prediction of odor threshold using all the descriptors calculated. Firstly, the heuristic method was used to reduce the pool of descriptors. The descriptors were reduced from 480-180. Secondly, various subset sizes were investigated to determine the optimum number of descriptors. To determine the optimum number of descriptors, the heuristic correlations provided the optimal equations for different numbers of descriptors in the range of 1-9. Plot of R^2, R^2_{cv} and S^2 values against the number of descriptors (Fig. 2) gave guidance regarding the number of descriptors to retain in the model.

It can be seen from Fig. 2 that R^2 and R^2_{cv} rise steeply with the number of parameters increasing from 1-9, while S^2 decreases steeply. In the present study, the best correlation equation with six descriptors was used for the analysis. A detailed description of the linear model was summarized in Table-2.

TABLE-2
DESCRIPTORS, COEFFICIENTS, STANDARD ERROR AND *t*-TEST VALUES OF THE MULTIPLE LINEAR MODEL

	Coefficients	Standard error	*t*-test	Descriptors
0	5.56	1.49	3.72	Intercept
1	-1.91	0.34	-5.66	Tot hybridization component of the molecular dipole
2	-0.29	0.07	-4.37	Count of H-donors sites [quantum-chemical PC]
3	-7.76	2.29	-3.40	FNSA-2 fractional PNSA (PNSA-2/TMSA) [qantum-chemical PC]
4	-34.70	8.65	-4.01	RPCG Relative positive charge (QMPOS/QTPLUS)[Zefirov's PC]
5	15.60	4.02	3.87	Principal moment of inertia A
6	59.70	27.40	2.18	HACA-1/TMSA [Zefirov's PC]

N = 40; R^2 = 0.8012; F = 22.17; RMS = 1.0011.

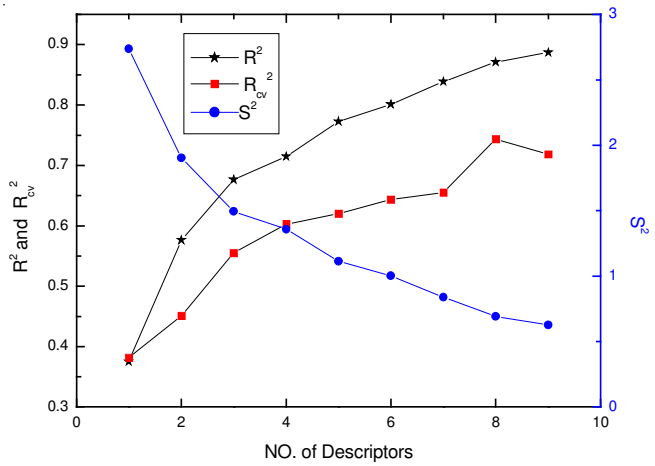

Fig. 2. Influences of the number of the descriptors on the correlation coefficient (R^2), the standard deviation (S^2) and cross validation (R^2_{cv})

Thus, ρ statistics of the new linear models is 6.8, which is much higher than the reference value 4, indicating that the model is proper[36].

The correlation matrix of the six selected descriptors was shown in Table-3. From Table-3, it can be seen that the linear correlation coefficient value of each of the two descriptors is < 0.80, which means the descriptors are not collinear[24,37]. With the external test set, the prediction results were obtained. The statistical parameters were R^2 = 0.648, F = 14.708, RMS = 1.7165. The predicted *versus* observed log T was shown in Table-1. Fig. 3 shows the predicted *versus* observed log T values for all of the compounds studied.

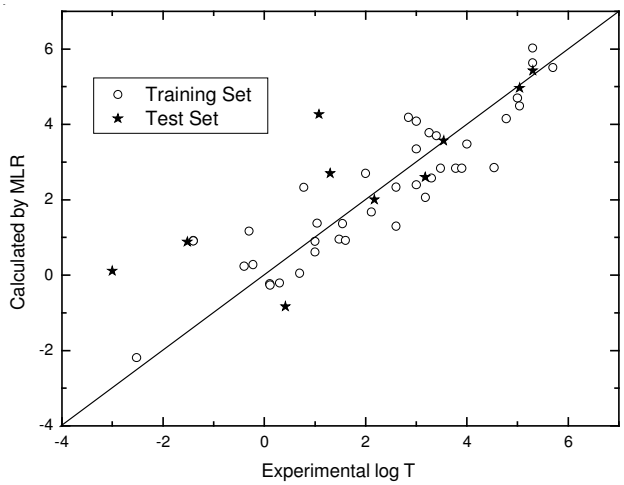

Fig. 3. Calculated *versus* experimental log T by MLR

TABLE-3
CORRELATIONS MATRIX OF THE SIX DESCRIPTORS USED IN THE MODEL

	I_A	T_d	H_{z2}	H_C	F_{N2}	R_+
I_A	1.000	–	–	–	–	–
T_d	0.326	1.000	–	–	–	–
H_{z2}	0.456	0.132	1.000	–	–	–
H_C	0.330	0.189	0.112	1.000	–	–
F_{N2}	0.102	-0.003	0.566	0.095	1.000	–
R_+	-0.534	-0.306	-0.328	0.259	0.254	1.000

Note: R+: RPCG Relative positive charge (QMPOS/QTPLUS) [Zefirov's PC]; F_{N2} : FNSA-2 Fractional PNSA (PNSA-2/TMSA) [Quantum-Chemical PC]; H_C : Count of H-donors sites [Quantum-Chemical PC]; H_{z2} : HACA-1/TMSA [Zefirov's PC];T_d: Tot hybridization comp. of the molecular dipole; I_A: Principal moment of inertia A.

Results of RBFNN: From the above result of MLR, we can see that the result is not so satisfied, especially for the external test set. So the nonlinear statistic method, RBFNN was used to develop a non-linear model based on the same subset of descriptors to see whether the results could be improved. The parameter that influences the performance of RBFNN was optimized. The selection of the optimal width value for RBFNN was performed by systemically changing its value in the training step. The value that gives the best leave-one-out (LOO) cross-validation result was used in the model. For this data set, the optimal value was determined as 4.00. The corresponding number of centers (hidden layer nodes) of RBFNN is 13. The predicted results of the nonlinear models were shown in Table-1 and Fig. 4. The obtained model had a square correlation coefficient R^2 = 0.8767, F = 269.98, with an RMS of 0.7152 for the training set. The statistical parameters of the test set were R^2 = 0.7746; F = 27.481 and RMS = 1.3570.

Results of SVM: The same as the RBFNN, the selection of the parameters for SVM was performed by systemically changing their value using the training step. The robustness of the models and their internal predictive ability were evaluated based on leave-one-out (LOO) cross-validation. The value, which gives the best LOO cross-validation result, was used in the model. The overall performances of SVM were evaluated in terms of RMS. The γ, ε and C for this data set were finally fixed to 0.003, 0.4 and 100, respectively. The predicted results of the nonlinear models were shown in Table-1 and Fig. 5. The SVM model gave similar results to MLR, that is, R^2 = 0.8023, RMS = 0.9271 for the training set and R^2 = 0.7033 and RMS = 1.5888 for the test set.

Comparison and validation of the MLR, RBFNN and SVM models: Comparison of the correlation models obtained

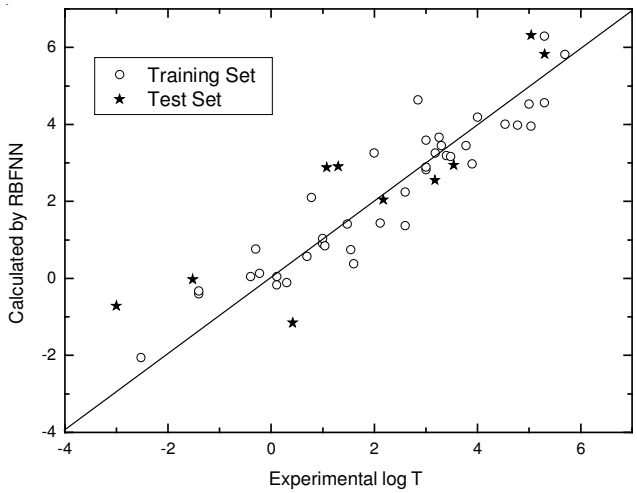

Fig. 4. Calculated *versus* experimental log T by RBFNN

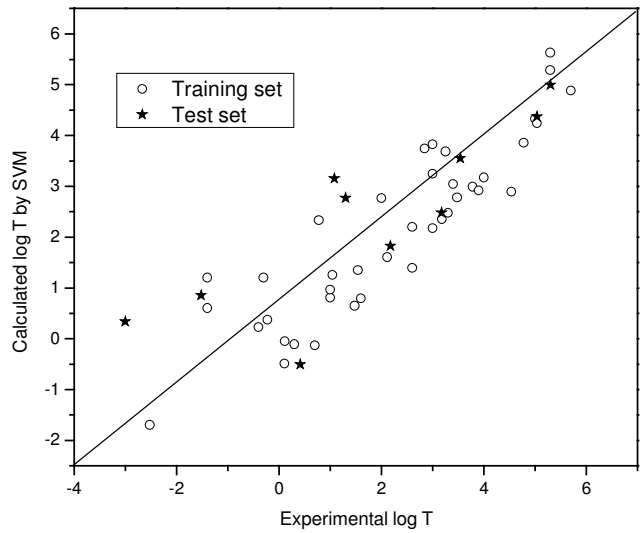

Fig. 5. Calculated *versus* experimental log T by SVM

by MLR, RBFNN and SVM, it is clear that the whole performance of RBFNN is better than that obtained by MLR and SVM. To further validate the models built, a fivefold cross-validation algorithm was applied for validation of the prediction results. In this process, the training set was then split into four parts: A (1, 5, 9, 13, ...), B (2, 6, 10, ...), C (3, 7, 11, ...), D (4, 8, 12, ...). The test set was defined as part E and each part contains 10 compounds. The remainder of the procedure was repeated five times. In each run, a different one of the five parts was kept apart, while the other four parts were used to construct all of the MLR, RBFNN and SVM models. The part that was kept separate was then used to verify the model. The reported RMS and R^2 for the training and test set for all of the models and for each of the five training-test set splits was also shown in Table-4. The results shown in Table-4 disclose an average training quality of $R^2 = 0.7721$, RMS = 1.1336 and an average predicting quality of $R^2 = 0.6776$, RMS = 1.3127 for MLR model. The results of the RBFNN model were $R^2 = 0.7839$, RMS = 0.9970 for training set; and an average predicting quality of $R^2 = 0.7268$, RMS = 1.2281, which proves that the proposed model has a relatively satisfactory statistical stability and validity. While to the SVM model, the results were: $R^2 = 0.7777$, RMS = 1.1034 for training set and an average predicting quality of $R^2 = 0.6615$, RMS = 1.3347. The results were similar to those of MLR. From the average results of each model, we can see that the models have a relatively satisfactory statistical stability and validity.

Discussions of the input parameters: As is well known, the factors influencing odors property is complex. Fragrance molecules enter through our nostrils then interact with receptors in the olfactory epithelium. The pungent sensations arisen from the activation of receptors are present within the free endings of the trigeminal nerve[38]. The process is not only dependent on the characteristic of physiological factors, but also on the physicochemical properties and the molecular structure. From the viewpoint of chemistry, property is determined by structure if the experimental condition is same. In this study, we try to seek the structure factors that influence the odor threshold of oxygen and nitrogen containing heterocycles. The six descriptors, which encode different structure feature of each compound, involved in the model can be classified as follows: (i) four as electrostatic descriptors; (ii) one as quantum chemical descriptor; (iii) one as geometrical descriptor.

RPCG relative positive charge (QMPOS/QTPLUS) [Zefirov's PC] (RPCG), an electrostatic descriptor, is a charged partial surface area descriptor. It was defined as the partial charge of the most positive atom divided by the total positive charge of the molecule and it represents the effect of the polar intermolecular interactions. The coefficient of the descriptor is negative. HACA-1/TMSA [Zefirov's PC] (HACA-1), another electrostatic descriptor, describes the ability of the compound to act as a hydrogen bond acceptor. HACA is defined as the sum of solvent accessible surface area of hydrogen bonding acceptor atoms in the molecule[38,39]. The minimum value of HACA is 0 and the maximum value is 2. 1 is for our dataset. The count of H-donors sites [ZefiroV's PC] (HD) distinguishes the molecules according to the number of hydrogen donor sites that are capable of donating a hydrogen to the surrounding media. Thus it indicates noncovalent hydrogen bonds action. As expected, hydrogen bond descriptors

TABLE-4
VALIDATION OF CORRELATIONS FOR THE MLR,RBFNN AND SVM MODELS

Training subset	R^2			RMS			Test subset	R^2			RMS		
	MLR	RBFNN	SVM	MLR	RBFNN	SVM		MLR	RBFNN	SVM	MLR	RBFNN	SVM
A+B+C+D	0.8012	0.8767	0.8023	1.0011	0.7165	0.9271	E	0.6476	0.7746	0.7033	1.7165	1.3570	1.5888
A+B+C+E	0.7664	0.7396	0.7572	1.1620	1.1142	1.0925	D	0.6563	0.6406	0.6063	1.2138	1.2441	1.2762
A+B+D+E	0.7937	0.7790	0.7963	1.0928	1.0274	0.9994	C	0.5107	0.6940	0.4670	1.4504	1.3258	1.5050
A+C+D+E	0.7660	0.7807	0.7843	1.1804	1.0380	1.0565	B	0.7154	0.7096	0.6869	1.3146	1.3533	1.3685
B+C+D+E	0.7323	0.7436	0.7486	1.2250	1.0890	1.1114	A	0.8589	0.8181	0.8433	0.8681	0.9285	0.9351
Average	0.7721	0.7839	0.7777	1.1336	0.9970	1.0374	–	0.6776	0.7268	0.6615	1.3127	1.2281	1.3347

are of major importance in modeling the transfer action. The hydrogen bond descriptors include simple integer examples such as the counts of hydrogen acceptor or donor sites together with the ratio of the maximal number of hydrogen bond donor or acceptor sites in a molecule to the corresponding minimal value, (HA, HD)min/max and advanced hydrogen bond descriptors expressed in the form of partial surface area. The latter include the hydrogen acceptor charged surface area (HACA) and the hydrogen donor charged surface area (HDCA). In our model, they have contrary coefficients. HACA-1 has a positive coefficient, while HD has a negative one. Fractional negative charge weighted surface area (FNSA-2) means total charge weighted negative surface area dived by total molecular surface area. Because of its negative coefficient in the linear model, increasing this descriptor also decreases the log T values.

The quantum chemical descriptor used most frequently is the total hybridization component of the molecular dipole (D_{thc}). The descriptor contributes negatively to the odor threshold.

The geometrical descriptors describe the size of the molecules and are derived from the three-dimensional coordinates of the atomic nuclei, the atomic masses and the atomic radii in the molecule. The descriptor contained in the model that belongs to this group is principal moment of inertia A (IA). The moments of inertia characterize the mass distribution in the molecule. It brings a positive contribution to the odor threshold. This observation implies that, all things being equal, increasing the value of this descriptor can lead to the larger values of odor threshold.

Conclusion

QSPR approach is used to investigate the relationship between the structures of oxygen and nitrogen containing heterocyclic compounds and their odor threshold values. The relatively high R^2, low RMS obtained from the models suggest that the models possesses well predictive ability, which allows us to estimate the log T of these compounds in cases where these values are not readily available or not tested easily. Of these models, MLR is more simple and interpretable and easy practical to use for the experimental scientists. Also, this paper provided a simple and straightforward way to predict log T of a diverse set of compounds from their structures alone and gave some insight into structural features related to this property of the compounds.

ACKNOWLEDGEMENTS

The authors thank the National Natural Science Foundation of China (NSFC) Fund (NO. 51073132) for financial support.

REFERENCES

1. M. Zviely, *P&F Magazine*, **5**, 4 (2006).
2. P. Laffort and F. Patte, *J. Chromatogr. A*, **406**, 51 (1987).
3. S. Mihara and H.S. Masuda, *J. Agric. Food Chem*., **36**, 1242 (1988).
4. J.I. Seemai, D.M. Ennis, H.V. Secor, L. Clawson and J. Palen, *Chem. Senses*, **14**, 395 (1989).
5. B. Winter, In ed.: J.L. Fauchere, In QSAR: Quantitative Structure-Activity Relationships in Drug Design, Alan R. Liss, New York, pp. 401-405 (1989).
6. P.A. Edwards and P.C. Jurs, *Chem. Senses*, **14**, 281 (1989).
7. P.A. Edwards, L.S. Ankerl and P.C. Jurs, *Chem. Senses*, **16**, 447 (1991).
8. M. Chastrette, *SAR QSAR Environ. Res*., **6**, 215 (1997).
9. T. Yamanaka, *Chem. Senses*, **20**, 471 (1995).
10. M.H. Abraham, J.M.R. Gola, J.E. Cometto-Muñiz and W.S. Cain, *Chem. Senses*, **27**, 95 (2002).
11. O. Ivanciuc, *J. Mol. Des*., **1**, 269 (2002).
12. K.M. Hau and D.W. Connell, *Indoor Air-Int. J. Indoor Air Quality Climate*, **8**, 23 (1998).
13. Y.X. Tan and K.J. Siebert, *J. Agric. Food Chem*., **52**, 3057 (2004).
14. B. Wailzer, J. Klocker, G. Buchbauer, G. Ecker and P. Wolschann, *J. Med. Chem*., **44**, 2805 (2001).
15. D. Zakarya, L. Farhaoui, M. Hamidi and M. Bouachrine, *J. Mol. Mode.*, **12**, 985 (2006).
16. F. Luan, X.H. Wang, H.T. Liu, Y. Gao, Y. Guo, Z.Y. Xie and X.Y. Zhang, *Flavour Frag. J*., **24**, 62 (2009).
17. F. Luan, H.T. Liu, Y.Y. Wen and X.Y. Zhang, *Flavour Frag. J*., **23**, 232 (2008).
18. P.W. Li, Z.H. Yu and H. Li, Food Flavor and Fragrance Chemistry: Heterocyclic aroma Chemicals, China Light Industry Press, Peking (1992).
19. ISIS Draw 2.3, MDL Information Systems, Inc., 1990-2000.
20. HyperChem 4.0, Hypercube, Inc. (1994).
21. M.J.S. Dewar, E.G. Zoebisch, E.F. Healy and P.J.J. Stewart, *J. Am. Chem. Soc*., **107**, 3898 (1985).
22. J.P.P. Stewart, MOPAC 6.0, Quantum Chemistry Program Exchange, QCPE, No. 455, Indiana University, Bloomington, IN (1989).
23. A.R. Katritzky, V.S. Lobanov and M. Karelson, CODESSA: Training Manual; University of Florida, Gainesville, FL (1995).
24. A.R. Katritzky, V.S. Lobanov and M. Karelson, Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual, Version 2.0 (1994).
25. E. Deconinck, D. Coomans and Y.Y. Heyden, *J. Pharm. Biomed. Anal.*, **43**, 119 (2007).
26. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi and J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics-Part A, Elsevier Science, Amsterdam (1997).
27. X.J. Yao, A. Panaye, P. Doucet, R.S. Zhang, H.F. Chen, M.C. Liu, Z.D. Hu and B.T. Fan, *J. Chem. Inf. Comput. Sci*., **44**, 1257 (2004).
28. Y.H. Xiang, M.C. Liu, X.Y. Zhang, R.S. Zhang, Z.D. Hu, B.T. Fan, J.P. Doucet and A. Panaye, *J. Chem. Inf. Comput. Sci*., **42**, 592 (2002).
29. M.J.L. Orr, Introduction to Radial Basis Function Networks, Centre for Cognitive Science, Edinburgh University (1996).
30. M.J.L. Orr, MATLAB Routines for Subset Selection and Ridge Regression in Linear Neural Networks, Centre for Cognitive Science, Edinburgh University (1996).
31. V.N. Vapnik, The Nature of Statistical Learning Theory, Springer (1995).
32. V.N. Vapnik, Statistical Learning Theory, Wiley, New York (1998).
33. N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, UK (2000).
34. B.B. Xia, W.P. Ma, B. Zheng, X.Y. Zhang and B.T. Fan, *Eur. J. Med. Chem*., **43**, 1489 (2008).
35. R. Garcia-Domenech and D.J.V. Julian-Ortiz, *J. Chem. Inf. Comp. Sci*., **38**, 445 (1998).
36. L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell and P. Gramatica, *Environ. Health. Perspect*., **111**, 1361 (2003).
37. J.G. Topliss and R.P. Edwards, *J. Med. Chem*., **22**, 1238 (1979).
38. W.L. Silver and T.E. Finger, In eds.: T.V. Getchell, R.L. Doty, L.M. Bartoshuk and J.B. Snow Jr., The Trigeminal System in Smell and Taste in Health and Disease, Raven Press, New York, pp. 97-108 (1991).
39. M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley-Interscience, New York (2000).