Journal of Computational Science & Engineering 10 (2014) 391-405

Journal of Computational Science & Engineering

Available at http://www.asocse.org ISSN 1710-4068

2D-QSPR modeling to predict the maximum absorption of 1,4-naphthoquinone dyes

Rongchao Li, Xuanxuan Chen, Feng Luan*

Department of Applied Chemistry, Yantai University, Yantai 264005, Shandong China.

Article Information

Article history:

Received 10 December 2013

Revised 15 January 2014

Accepted 26 January y 2014

Available online 29 January 2014

Keywords:

Dyes
Quantitative Structure-Activity/Property Relationships (QSA/PR)
Maximum absorption
1, 4-naphthoquinone dyes
Multiple Linear Regression (MLR)
Radial Basis Function Neural Networks (RBFNN)

Abstract

Dyes have been applied and are playing increasingly important role in many industries including textile, printing, medical, and energy industry *etc*. The wide applications of them imply that specific dyes possessing given properties should be effectively designed. In present study, a QSPR modeling work performed on the maximum absorption of 1, 4-naphthoquinone derivatives was investigated using the classical 2D-QSPR method by linear multiple linear regression (MLR) model and nonlinear radial basis function neural network (RBFNN). Both of the models were verified by internal validation and external validation. The RBFNN model gives better results: N=40,R^2=0.9524, LOOq^2=0.8815, F=578.758, RMS=17.4988 for training set, and R^2=0.9231, q^2_{ext} =0.9202, F=84.010, RMS=25.3406 for the external test set. While the MLR model identify the most appropriate descriptors and gives a simple and intuitionistic model whose statistic parameter are: R^2=0.8887, LOOq^2=0.8627, F=231.552, and RMS=28.6358 for training set; R^2=0.8622, q^2_{ext}=0.8308, F=423.809, and RMS=33.8613 for the external test set. The appropriate results proved the models to be meaningful and useful to predict the maximum absorption of 1, 4-naphthoquinone derivatives.

1. Introduction

The history of dyes can be traced back to 2600 BC when there was the earliest written

* Corresponding author: fluan@sina.com

2014 © The American Computational Science Society. All rights reserved.

record of the use of dyes in China [1]. Colour, which has ever been called natural dyes, was only obtained from the natural sources such as plant and animals. Until 1856, an English chemist named William Henry Perkin synthesized the first dye-aniline purple indicating synthetic dyes entered into industrial production period [2, 3]. From then on, synthetic dyes are playing a increasing important role in many fields of our life. By now, more than 8600 kinds of dyes have been synthesized and used. Among so many dyes, most of them are synthetic dyes and they quickly replaced the traditional natural dyes since their superior properties such as less cost, a vast range of new colors, imparting better properties to the dyed materials *etc*.

Now, all of these dyes are not only used limited in dyeing and printing for textiles but also widely used in every other fields of life, such as dye ink for printing industry [4], dye laser for medical field [5, 6] dye-sensitized solar cells for energy field [7], thermal transfer systems, molecular switches, media storages or photovoltaic devices and food industry *etc* [8-10]. As we know, both of the researchers and the industrialists focused their attention mostly on the special properties of different dyes in the process of designing or using different kinds of dyes. In general speaking, the properties of dyes include colour, brightness, solubility, affinity, fastness, absorption maxima, mutagenicity, diffusion constant, lipophilicity, bioelimination, antimalarial activity and so on [11]. All of these properties or activities of dyes can be tested by experimental method. However, the experimental method requires pure reagent, expensive equipment, and much time spending on it. In addition, the method also has to face the problems that some of the dyes are toxic, which may be carcinogenic or mutagenic, or harmful to our health in some other way. As the point view of chemistry, properties are determined by the structures of the compounds. For instance, the colour of azo benzene compounds is dependent on the number of –N=N- bond and the length of conjugation chain. The longer the conjugated chain, the deeper the color, the more likely the electronic excitation which means the longer wavelengths molecules absorb. Also to the solubility of dyes, on one hand it is decided by the nature of compound (mostly depended on the polarity), and on the other hand is related to the extrinsic condition such as temperature, pressure and solvent *etc*. Thus, it gives us inspiration to investigate the relationship between the structure and properties or activities of the dyes, either qualitatively or quantitatively.

QSAR/QSPR, as a calculated method, has been used several decades. Up to now, it was widely applied in physical chemistry, pharmaceutical chemistry, environmental chemistry, toxicology, and other research fields [12-15]. This method requires no reagent and no equipment. It only needs the researchers compute the molecular descriptors by computer chemical software, and then a relationship model between the activity/property and the molecular descriptors will be established. If the model is robust and has good prediction ability, it can be used to predict the property/activity of the dyes even the compounds have not been synthesized or found yet. It is also an effective tool to make up for the shortage of the experimental method, to aid the design of new dyes.

In the past decades, many researchers paid their attention to investigate the activity /property of dyes by QSAR/QSPR method. Some excellent results have been obtained, however, no review on this topic is found until now. We would survey the information related to this area published from 1995 to 2011 in the following. Affinity is one of the major properties of dyes since dyes are colored substances that have affinity to the substrate to which they are being applied. To this property, Timofei *et al.* have performed several QSAR studies on some kind of dyes including anthraquinone vat dyes [16], symmetrical bisazo dyes [17], disazo anionic dyes [18], disperse azo dye [19], heterocyclic monoazo dye [20, 21],anionic azo dyes [22], anthraquinone vat mono and disazo and disperses dyes [23], direct azo dyes [24]. Besides this, there are also some other research groups focused their attention on this property such as azo and anthraquinone dyes [25], the heterocyclic monoazo dyes [26], anionic sulfur-

2014 © The American Computational Science Society. All rights reserved.

containing azo dyes and anthraquinone dyes [27].

Fastness property is another important property of dyes. For instance, there are many test items for textile color fastness property, such as color fastness to perspiration; color fastness to light (usually use Xenon meter to perform this test), color fastness to rubbing (use Crock meter to perform this test) and color fastness to washing (use SW-12M to perform this test) [28]. This property has also been studied by this approach [29-31].

Spectroscopic property is related to the important color property, which determined mainly by the structure of the dyes. It is usually represented by the maximum absorption wavelengths (λ_{max}). Once the λ_{max} of the compound is known, the color will be unequivocally determined. Nowadays, the λ_{max} of the dyes is usually obtained experimentally [32]. There are some reports about applications of QSPR methods to investigate the relationship between the λ_{max} and the structure of dyes in recent years, such as azobenzenes (AB) [33,34], 9,10-anthraquinones and azobenzenes [35], 1, 4-naphthoquinone derivatives [36], and sensitizing dyes[37,38].

Solubility is often said to be one of the "characteristic properties of a substance," which means that solubility is commonly used to describe the substance, to indicate a substance's polarity, to help to distinguish it from other substances, and as a guide to applications of the substance. According to an IUPAC definition [39], solubility is the analytical composition of a saturated solution expressed as a proportion of a designated solute in a designated solvent. This property is also important to dyes and been studied by QSA/PR approaches [40-43]. Other activities (such as mutagenicity) have also been taken care of using QSAR studies since some dyes being toxic, volatile, explosive or radioactive substances [44-46].

Besides the above-mentioned properties, there are also other properties such as diffusion constants [47], bioelimination [48], biodegrade-ability [49, 50], growth inhibitory effect [51], tinctorial properties [52], sensitization potential [53], water tracer properties [54], fish bioconcentration [55], adsorption mechanisms [56], toxicity [57], brightening properties [58], and half-wave potential (E1/2) [59] being studied by QSA/PR approaches.

In summary, many properties or activities of differen dyes have been studied by QSAR or QSPR modeling in the past decades, and these studies even thrown some light on the road to design of new dyes, to evaluate the different properties and to explain the mechanism of dyeing process. So many research groups, now and even in the future, would keep on focusing their attention on different chemical families to explore the relationship between the molecular structure of the dyes and their properties using a variety of QSA/PR methods. To the best of our knowledge, there has been no general QSPR study on the kind of 1,4-naphthoquinone dyes using the general descriptors and nonlinear methods. In current paper, we will present the development of classical 2D-QSPR models correlating λ_{max} with 43 1, 2-naphthoquinone derivatives.

2. Materials and Methods

2.1 Date Set

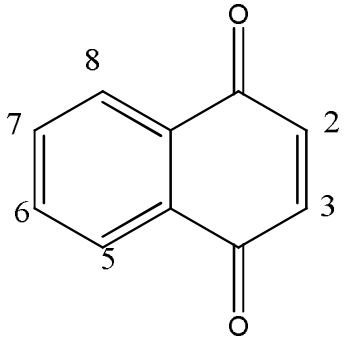

Fig. 1 General structure of 1, 4-naphthoquinone derivatives

In this study, we used 43 1,4-naphthoquinone derivatives (See Figure 1) the same as Atabati *et al* investigated [36]. In this group of compounds, the atom "H"s in position 2,3,5,8 of 1,4-naphthoquinone were replaced by the substituents such as methyl radical, chlorine, amino group, hydroxyl, cyan, or not. All of the studies compounds and their corresponding experimental λ_{max} values were listed in Table 1.

2014 © The American Computational Science Society. All rights reserved.

Table1. Chemical's number, substituents and their corresponding experimental and calculated λ_{max}

NO.	Substituents	λ_{max}^{exp}	λ_{max}^{cal}		Data Set
			MLR	RBFNN	
1	-	330	306	329	A
2	2,3-Cl,5-NH_2	531	515	537	C
3	2,3-Cl,5-OH,8-NH_2	576	568	576	C
4	2,3-Cl,5,8-OH	495	477	463	T
5	2,3-Cl,5,8-OMe,8-NH_2	582	522	558	D
6	2,3-Cl,5,8-OMe,8-NHMe	612	527	590	T
7	2,3-Cl,5,8-OMe	476	481	492	T
8	2-Cl,3-NHMe,5-NO_2	490	477	501	C
9	2-Cl,3-OMe	338	409	364	E
10	2-Me	333	311	312	T
11	2,3-Me	330	324	325	C
12	2,3-Me, 5-NH_2	500	506	518	E
13	2,3-Me, 6-NH_2	479	474	460	D
14	2,3-Me, 5-NHMe	530	523	533	B
15	2,3-Me, 5-NHC_6H_{11}	538	522	522	T
16	2-Me,3,5-OH	409	417	415	B
17	5-NH_2	516	528	505	A
18	5-NH_2,8-OH	572	584	579	B
19	5-NH_2,8-OMe	572	587	603	T
20	2-NHMe	452	457	454	D
21	2-NHMe,3-Cl	477	490	485	E
22	2-NHMe,3-Cl,5-NO_2	490	-	-	outlier
23	2-NHMe,3-CN	482	439	426	A
24	2,3-NHMe	544	539	568	D
25	2-NHMe,3-NHPh	552	532	547	E
26	2-NHMe,3-OMe	492	466	464	D
27	5-NHMe	552	547	527	A
28	5-NHMe,8-OH	597	594	590	A
29	5-NHMe,8-OMe	595	598	601	E
30	2-NMe_2	471	494	493	C
31	2-NHPh,3-Cl	482	515	526	T
32	2-NHPh,3-CN	440	475	465	C
33	2,3-OH	440	417	422	T
34	2,5-OH	420	-	-	outlier
35	2,5,8-OH	480	489	480	B
36	2,6-OH	390	376	383	T
37	5,6-OH	461	438	453	A
38	5,8-OH	486	469	483	B
39	2-OMe	333	301	332	B

2014 © The American Computational Science Society. All rights reserved.

40	2,3-OMe	335	387	357	D
41	2,3-Me,5-NO_2	386	-	-	outlier
42	5-OMe	400	417	394	A
43	5,8-OMe	456	502	470	E

As the commonly QSPR or QSAR approach did, the 43 compounds were divided into 34 training set and 9 external test set randomly. The wavelength of training set is ranging from 330nm to 597nm. The wavelength of external test is ranging from 333nm to 612nm，which covers almost the same scale as the training set. The training set was used to select proper descriptors and construct models (both of linear multiple linear regression (MLR) model and nonlinear radial basis function neural network (RBFNN) and the external test set was used to evaluate the constructed models.

2.2 Descriptor Generation

Structures of all the compounds were drawn in ISIS Draw 2.3 [60]. The preliminary molecular geometry optimization was performed using molecular mechanics MM+ force fields by HyperChem software [61]. The further optimization of the molecular structures was performed by semi-empirical PM3 method using the polak-ribiere algorithm until the root mean square gradient was 0.01 kcal/mol [62]. Then the last optimization was done using PM3 method in MOPAC software package [63]. Doing so, the files exported from HyperChem and Mopac were transferred into CODESSA [64] for generating descriptors.

Finally, we obtained 487 descriptors of each molecular including 38 constitutional descriptors, 38 topological descriptors, 12 geometrical descriptors, 77 electrostatic descriptors and 322 quantum-chemical descriptors. In addition, 7 other descriptors, which cannot obtain by CODESSA were calculated by HyperChem software. These descriptors (494 in all) are also used for further statistic analysis.

2.3 Selection of Molecular Descriptors

The task of building the best QSAR or QSPR model for a particular property needs a preliminary search for gathering the best subset of descriptors that can be used to represent the property properly. In the present study, this preliminary feature selection entailed pruning the descriptors pool through the heuristic method (HM) available in the framework of the CODESSA program [58, 64]. This method, as an effective method, has been used in several papers [65, 66].

The heuristic method (HM) can quickly provide either a good estimation about what quality of correlation to expect from the data or allow deriving several best regression models. Besides, it shows which descriptors have bad or missing values, which are insignificant from the standpoint view of a single-parameter correlation, and which are highly inter-correlated. HM proceeds with a pre-selection of the descriptors to eliminate: (1) those that are not available for each structure, (2) those having a small variation in magnitude for all structures, (3) those that provide a *F*-test value below 1.0 in the one-parameter correlations and (4) those whose *t*-values are less than the user-specified values, *etc*.

This is followed by listing the remaining descriptors in descending order of correlation coefficient when used in one-parameter correlations. Subsequently, the given property is correlated with (1) the top descriptor in that list with each of the remaining descriptors, (2) the next one with each of the remaining descriptors, and so on. The best pairs as evidenced by the highest *F*-values in two-parameter correlations are then chosen and used for further inclusion of descriptors in a similar manner.

2.4 Multiple Linear Regressions (MLR)

As we know, the MLR model is a very simple, strong, descriptive model. In this model, molecular descriptors (X) are correlated with one or more response variable (property values-Y). The equation is:

2014 © The American Computational Science Society. All rights reserved.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n \qquad (1)$$

Where b_0 is the constant, b_1 to b_n is the coefficient of the descriptor X_1 to X_n. In MLR analysis, the severity of multicollinearity is often analyzed by the variance inflation factor (VIF) [67]. VIF is calculated as $1/(1- R_i^2)$, where R_i^2 is the multiple correlation coefficient of the regression equation in step one. Previous researches have recommended that the VIF value should be smaller than 5, otherwise the information retrieved from the descriptors can be hidden by multicollinearity [68].

Another parameter that is usually applied to examine the relative importance as well as the contribution of each descriptor in the model is the value of the mean effect, MF, defined as follows [69]:

$$\mathrm{MF}_j = \frac{b_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m b_j \sum_{i=1}^{i=n} d_{ij}} \qquad (2)$$

Where MF_j represents the mean effect for the considered descriptor j, b_j is the coefficient of the descriptor j, d_{ij} stands for the value of the target descriptor j for molecule i, n is number of compounds in the training set, and m is the descriptors number for the model. The MF value indicates the relative importance of a descriptor compared with the other descriptors in the model. Its sign indicates the variation direction in the values of the property as a result of the increase (or reduction) of the descriptor values.

2.5 Radial Basis Function Neural Networks (RBFNN)

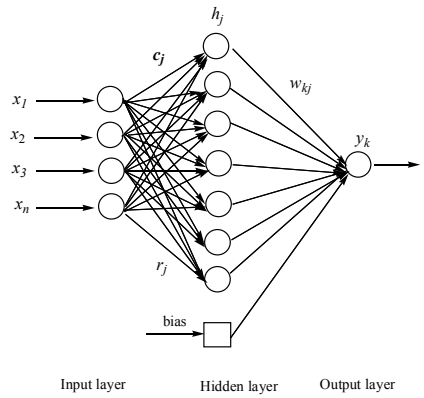

Fig. 2 The architecture of RBFNN.

RBFNN [70, 71], as the nonlinear methods usually giving better results, have been employed to establish the non-linear model in present study. The basic form of the RBFNN architecture involves three different layers. Figure 2 shows the basic network architecture. The input layer is made of n source nodes while the second layer is a hidden layer of high enough dimension which senses a different purpose from that in a multilayer perception. The output layer supplies the response of the network to the activation patterns applied to the input layer. The transformation from the input layer to the hidden units is nonlinear whereas the transformation from the hidden units to the output layer is linear.

As to the hidden layer of RBFNN, it consists of a number of RBF units (n_h) and bias (b_k). Each neuron on the hidden layer employs a radial basis function as a nonlinear transfer function to operate on the input data. The more often used RBF is a Gaussian function that is characterized by a center (c_j) and width (r_j). The RBF functions measures the Euclidean distance between 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:

$$h_j(x_i) = \exp\left(-\left\|x_i - c_j\right\|^2 / r_j^2\right) \qquad (3)$$

denoting h_j the output of the j_{th} RBF unit.

The linear operation carried out by the output layer is as follows:

$$y_k(x_i) = \sum_{j=1}^{n_k} w_{kj} h_j(x_i) + b_k \qquad (4)$$

where y_k is the k_{th} output unit for the input vector x_i, 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.

Therefore, designing a RBFNN involves selecting the number of hidden layer units, the RBF centers and widths, and the connection weight. Here, we have used Gaussian functions

2014 © The American Computational Science Society. All rights reserved.

with a constant width, being that the same for all units. The adjustment of the connection weight between the hidden and output layers was performed by a least-squares solution after number of trial experiments taking into account the LOO-CV error, *i.e.* the one that gave the minimum error was chosen as the optimal value. The overall performance of RBFNN is evaluated in terms of root mean squared error (*RMS*).

$$RMS = \sqrt{\frac{\sum_{i=1}^{n_s}(y_k - \hat{y}_k)^2}{n_s}} \quad (5)$$

Where y_k is the desired output, $\hat{y}_k$ is the actual output of the network, and n_s is the number of compounds in the analyzed set.

All programs implementing RBFNN were written in M-file based on a MATLAB script for RBFNN. The scripts were run on a personal computer. The goodness and robustness of the model was estimated by the same statistic parameters used in the linear model.

2.6 Model Applicability Domain (AD)

Ideally, the AD should define the structural, physicochemical and response space of the model. It is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSAR model should make reliable predictions. This region is defined by the nature of the chemicals in training set and can be characterized in various ways. To visualize the outliers in a model, outliers in both the descriptor space and the response space, a plot of standardized residuals (R) *vs.* leverages (or hat values, h), called the Williams plot is sometimes used [74]. In this study, a simple measure of a chemical being too far from the applicability domain of the model is its leverage, h_i [75], defined as:

$$\mathrm{h_i} = x_i^T (X^T X)^{-1} x_i \quad (\mathrm{i}=1,2,.....n) \quad (6)$$

Where x_i is the descriptor row vector of the query compound, X is the n*k-1 matrix of k model descriptor values for n training set compounds. The superscript 'T' refers to the transpose of the matrix/vector.

3. Results and Discussion

3.1 Results of the MLR

Using the heuristic method, firstly, 306 descriptors were removed from the origin pool of 494 descriptors according to the criterion of deleting bad values, missing values or insignificant values. Secondly, 37 highly correlated descriptors were removed subsequently. Finally, we used the left 151 descriptors for deriving MLR models. In order to compare the result with Atabati's [36], the preset number of descriptors is set to 3. Thus, ten best models were developed. The model having the best statistical results was used for further study. The selected molecular descriptors are HOMO energy (ε HOMO), atom charge weighted PNSA (PNSA-3), and min nucleoph react index for a C atom (*N'C*). The relationship between the λ max and the three descriptors is list below:

$$Y = 1694 + 137.3 \times \varepsilon_{HOMO} - 0.602 \times (PNSA - 3) + 921490 \times N'_C \quad (7)$$

The statistic results of the model are: N=43, R^2=0.8850, LOOq^2=0.8570, F=246.197 and RMS=28.3820 for training set. For the external test set, the values are: R^2=0.8468, q^2_{ext}=0.8053, F=38.690, and RMS=35.2053. Here, R^2 is the correlation coefficient, LOOq^2 is the leave one out correlation coefficient, RMS is the root mean square error of the fit, F is the value of the Fisher test, and q^2_{ext} is the an external q^2 [76, 77]. In this equation, $\varepsilon_{\mathrm{HOMO}}$ and N'_C have positive effect on $\lambda_{\max}$, which means that the value will augment with the increase of the two descriptor. While the other descriptor (PNSA-3) has the negative effect on the $\lambda_{\max}$, which means that the value will reduce with the increase of the value of the descriptor.

2014 © The American Computational Science Society. All rights reserved.

3.2 Model Applicability Domain Analysis and Improved MLR Model

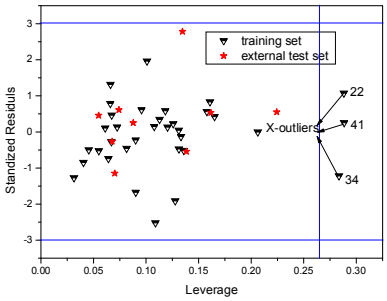

Fig. 3 The William plot of the training set and external test set.

Once the model built, we should firstly consider its applicability domain. In this study, William plot was used to eliminate the outliers from the model. In Figure 3, the X-coordinate stands for the leverage, h_i, and Y-coordinate represents the standard residual. The horizontal and vertical straight lines represent the limits of the normal values of Y outliers and X outliers, respectively. The limit of Y-coordinate is ± 3 times of the standardized residuals. The limit of X-coordinate is h^*, which is 3m/n, where m is the number of selected descriptors, and n is the number of training set. In present study, h^* is 0.2647. We can see from Figure 3, there are three X-outliers in this model, and all of them belong to the training set. So the three outliers were eliminated in the model, and an improved model was built using the left compound with the same descriptors. The relationship in is list below:

$$Y = 1663 + 133.1 \times \varepsilon_{HOMO} - 0.369 \times (PNSA - 3) + 1097158 \times N'_C \quad (8)$$

The statistical results of the new model changed to: N=40, R^2=0.8887, LOOq^2=0.8627, F=231.552, and RMS=28.6358 for training set. Using the built model, the prediction results were obtained with the external test set: R^2=0.8622, q^2_{ext}=0.8308, F=423.809, and RMS=33.8613. Compared with the results of the two models, the new one has a better statistical result, especially for the external test set. The experimental λ_{max} and calculated λ_{max} of every compound are shown in Table 1. The good quality of this model can be further judged by inspection of Figure 4 (A), where a plot of the predicted λ_{max} values versus the experimental values is shown for the whole dataset compounds as well as Figure 5 (A), which shows the scatter plot of residuals of the whole data set.

Another aspect deserving special attention is the degree of collinearity among the variables of a MLR model, it may easily be diagnosed by analyzing the cross-correlation matrix. The correlation matrix of the three descriptors is shown in Table 2. It can be seen that the linear correlation coefficient value of each of the two descriptors is <0.80, which means that each descriptor is independent [78].

The results of variance inflation factor (VIF) and the value of the mean effect (MF) were also shown in Table 3. VIF represents the multicolinearity of the selected descriptors [67]. When VIF>5, the information of descriptors can be hidden by correlation of descriptors [68]. In Table 3, all the VIF values are less than 5, indicating that the obtained model has statistical significance.

When the model built, we want to know the relative importance as well as the contribution of each descriptor in the model. For this purpose, the MF value was calculated and list in Table 3. As can be seen in the table, the value of ε_{HOMO} is much bigger than the other two. Thus, it can be concluded that ε_{HOMO} is the most important descriptor in the model.

3.3 Results of the RBFNN

In the research fields of QSAR or QSPR, nonlinear algorithms, which can represent the complicate relationships between the activity

2014 © The American Computational Science Society. All rights reserved.

or property of the molecules and the structures, are usually used. And the nonlinear models usually give better results than the linear ones. In order to see if there were better results, we have therefore decided to resort to the nonlinear RBFNN method to develop a model based on the same subset of descriptors. In so doing, the parameters that most influence the performance of RBFNNs were optimised in order to build a suitable model. For instance, the optimal width for the RBFNN was carefully selected by systemically changing its value in the training step. In this way, the value that gives the best leave one out (LOO) cross-validation result was selected in the model. The value of the optimal width (2.3) and the corresponding number of centers (14 hidden layer nodes) is defined based on the best leave one out (LOO) cross-validation result. At last, a 3-14-1 model was built.

The training quality of R^2=0.9524, LOOq^2=0.8815, F=578.758, RMS=17.4988; and a predicting quality of R^2=0.9231, q^2_{ext} =0.9202, F=84.010, RMS=25.3406 were obtained for the models. The predicted λ_{max} is listed in Table 1 and the plot of the predicted λ_{max} values versus the experimental values and the plot of the distribution of the residuals, both attained with such model for

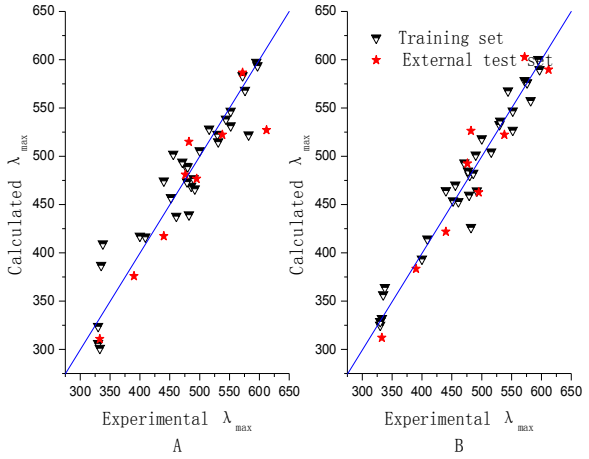

Fig. 4 Plot of training set and external test set by MLR(A) and RBFNN(B)

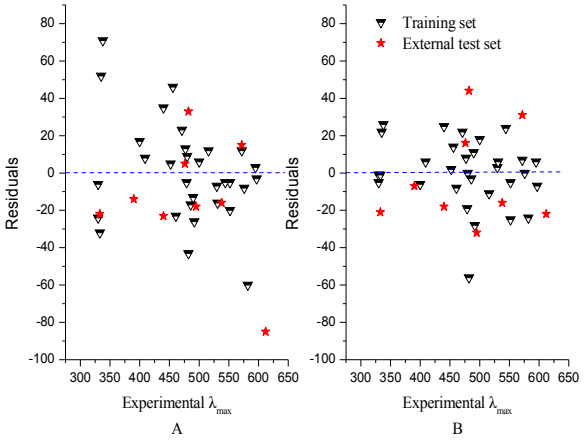

Fig. 5 Residuals of training set and test set by MLR(A) and RBFNN(B)

Table 2 Correlation matrix of the 3 descriptors used in the model

	ε_{HOMO}	PNSA-3	N'_C
ε_{HOMO}	1		
PNSA-3	0.640	1	
N'_C	0.013	0.114	1

Table 3 The VIF and MF values of each descriptor

Descriptor	Definition	VIF	MF
ε_{HOMO}	HOMO energy	1.70	1.029
PNSA-3	Atomic charge weighted PNSA	1.72	-0.011
N'_C	Min nucleoph react index for a C atom	1.02	-0.018

the whole dataset of compounds, are displayed in Figure 4 and 5(B), respectively.

Comparing the results obtained by the MLR with those of the RBFNN, it can be concluded that the performance of RBFNN is better than that of MLR, as usually do. Also from Figure 4 and 5, we can see that both of the two methods

2014 © The American Computational Science Society. All rights reserved.

showing satisfactory results.

3.4 Validation and Comparison of the two models

A good QSPR model should have a significantly statistical result for both internal validation and external validation. In this paper, the two models were evaluated by internal validation (LOOq^2, leave many out, Y-randomly) and external validation (R^2, R_0^2, q^2_{ext}, k, *etc*.), respectively.

Table 4 Validation of the model of MLR

Training set	R^2	F	RMS	Test set	R^2	F	RMS
A+B+C+D	0.9153	248.552	15.3205	E	0.9410	63.842	40.2138
B+C+D+E	0.8781	158.520	30.1881	A	0.9501	95.135	25.8030
A+C+D+E	0.8753	161.388	30.9747	B	0.9763	165.118	20.3183
A+B+C+D	0.8784	166.150	30.7213	C	0.9345	57.048	20.0266
A+B+C+E	0.9110	235.324	26.2377	D	0.9221	47.339	37.7106
Average	0.8916	193.9868	26.6884		0.9448	85.6964	28.8144

Table 5 Validation of the model of RBFNN

Training set	R^2	F	RMS	Test set	R^2	F	RMS
A+B+C+D	0.9309	341.265	20.0383	E	0.9088	39.862	36.5390
B+C+D+E	0.9600	528.559	15.7843	A	0.9314	67.922	36.2922
A+C+D+E	0.9030	214.068	25.0511	B	0.9082	39.556	26.7838
A+B+C+D	0.9137	243.429	23.7290	C	0.8632	25.239	30.7786
A+B+C+E	0.9257	286.737	21.9739	D	0.8781	28.825	29.7643
Average	0.9266	322.8116	21.3153		0.8979	40.281	32.0316

Considering the internal validation, LOOq^2 is 0.8627 for MLR model, and 0.8815 for RBFNN model, which are greater than 0.5 (as an indicator by some authors) [76, 77]. The internal validation was further checked by five-fold-out cross validation (FFO). In present study, the training set was split into five subsets (A, B, C, D, and E), and the procedure was run five times. In each run, one of these subsets was kept apart, while the other four were used to construct the MLR models. The statistical results of FFOCV obtained on each of the five training-test set splits are shown in Table 4 and 5. As can be seen, the results disclose an average training quality of R^2 = 0.8916; F=193.9868, RMS = 26.6884, and an average predicting quality of R^2= 0.9448; F=85.6994, RMSFFO = 28.8144 for the model, showing that it has a satisfactory statistical stability and validity.

As we know, some models have good leave one out and leave many out results, however, when the dependent-variable vector, Y-vector, is randomly shuffled and the new models may change to bad statistical results. This maybe due to a chance correlation or structural redundancy of the training set [79]. So, Y-randomly validation test is necessary for the QSPR model. *Y-randomization* test is a widely used technique to ensure the robustness of a QSPR model [79, 80]. In this test, the dependent-variable vector is randomly shuffled and a new QSPR model is developed using the original independent-variable matrix. The process is repeated several times. If the model has low R^2 and LOOq^2, it implies that an acceptable QSPR can be obtained for the given data set by the current modeling method. In present study, a ten times test were developed for both of the linear and nonlinear models, the R^2 and LOOq^2 of the two models were shown in Table 5. From Table 5, we can see that the models have low R^2 and LOOq^2**.** The R^2 and LOOq^2 results of Y-randomly for our models

2014 © The American Computational Science Society. All rights reserved.

are listed in Table 6. The low R^2 and LOOq^2 indicate that both models are robust and have no chance correlation or structural redundancy.

Table 6 The R^2 and LOOq^2 values of ten Y-randomly validation tests

NO.	MLR		RBFNN	
	R^2	LOOq^2	R^2	LOOq^2
1	0.2201	0.0654	0.1992	<0.0001
2	0.2261	0.0847	0.3277	0.0968
3	0.2975	0.1277	0.2495	0.0109
4	0.1166	0.0001	0.265	0.0019
5	0.0932	0.0532	0.2121	0.0113
6	0.1569	0.0013	0.1807	0.0279
7	0.0752	0.0679	0.0765	0.0025
8	0.0598	0.0837	0.0991	0.0029
9	0.0633	0.0342	0.2773	0.0133
10	0.1253	0.0048	0.1361	0.0001
Average	0.1434	0.0523	0.2023	0.0186

In the QSPR studies, besides training set compounds were used to construct the two models, the left compounds as test set were used to assess the predictive power of the constructed models. In this paper, about one-fifth of (9 compounds) the whole compounds (40 compounds) were chosen as the external test set. As recommendation in the reference [81], the following parameters were used to estimate the predictive power of a QSAR model using the external test set: (i) correlation coefficient R^2 between the predicted and observed activities; (ii) coefficients of determination (predicted versus observed activities, R_0^2, and observed versus predicted activities, $R_0'^2$); (iii) slopes k and k' of regression lines (predicted versus observed activities, and observed versus predicted activities) through the origin. The authors concluded that a QSAR model had an acceptable predictive capability if the following conditions were satisfied [83]:

$$q^2>0.5; \quad (9)$$

$$R^2>0.6; \quad (10)$$

$$\frac{(R^2-R_0^2)}{R^2} < 0.1 \text{ or } \frac{(R^2-R_0'^2)}{R^2} < 0.1; \quad (11)$$

$$0.85\leq k \leq 1.15 \text{ or } 0.85\leq k' \leq 1.15; \quad (12)$$

Table 7 The statistical results of the external test set for the two models

	MLR	RBFNN
R^2	0.8622	0.9231
q^2_{ext}	0.8308	0.9202
R_0^2	0.8589	0.9207
$(R^2- R_0^2)/ R^2$	0.004	0.003
k	0.9693	0.9958
F	43.809	84.010
RMS	33.8613	25.3406

The statistical results of our models for external test set are listed in Table 7. From the table, we can see that all the statistical results meet the requirements list above, which proves the good predictive ability of our models. In a

2014 © The American Computational Science Society. All rights reserved.

word, the MLR model and RBFNN model that we built are robustness and have a great predictive power.

In the end, we gave a comparison of our work to the results of Atabati's [36]. The statistic parameters of previous work were: R^2=0.8305, RMS=31.61 for the training set, and R^2=0.8793, RMS= 0.9377 for the external test set. We can see our model is somewhat better than the model built by Atabati *et al*.

3.5 Discussion of the input parameters

By interpreting the descriptors in the model, it is possible to gain some insight into factors affecting the λ_{max} value and find out which structural factor plays an important role. In the model, there are three descriptors (HOMO energy (ε_{HOMO}), atom charge weighted PNSA (PNSA-3), and min nucleoph react index for a C atom (N'_C)). ε_{HOMO} and N'_C belong to quantum-chemical descriptors and PNSA-3 is an electrostatic descriptor. Among them, ε_{HOMO} and N'_C bear positive coefficients, and PNSA-3 has negative one. As usually, the positive sign of the descriptor means that it has a positive effect on the activity, and vice versa.

As mentioned above, we used the absolute MF value to estimate the mean effect for the each descriptor. The order of the three descriptors used in our model is: $\varepsilon_{HOMO}(|1.029|)$ >N'_C ($|-0.018|$)>PNSA-3($|-0.011|$), which means quantum-chemical descriptors (ε_{HOMO}) play a major positive role on λ_{max}, while N'_C and PNSA-3 plays the similar influence on this property. In the following, we will explain the three descriptors in detail.

ε_{HOMO} is the energy of the highest occupied molecular orbital. It is rather simple and common to contract QSPR models. As we know, the λ_{max} is related to the excited process of the electron. The larger value of this descriptor is, the smaller is the energy of excitation, thus λ_{max} will be larger, which is consistent with the positive sign of the descriptor and can be easily understood. That is, ε_{HOMO} has a positive influence on λ_{max}, indicating that molecules with higher energy of the highest occupied molecular orbital tend to increase their λ_{max}.

The second descriptor is N'_C which stands for the minimum value of a C nucleophilic. It defined as:

$N'_C = \sum_{i \in C} C^2_{iHOMO}$ where C_{iHOMO} denote the i-th AO coefficients on the highest occupied molecular orbital. This descriptor is also closely related to the ε_{HOMO}. Here molecules with more occupied molecular orbitals and higher C atom AO coefficient tend to increase λ_{max}.

PNSA-3 is atom charge weighted partial negative surface ares (PNSA). This descriptor reflects the characteristics of the charge distribution of the molecules. As it bears a negative coefficient in the MLR model, that indicates that increasing the atomic charge weighted partial negative surface area of the molecular can lead to a small λ_{max} values. The presence of this descriptor in our model underlines the importance of the partial negative charges for the excitation process.

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 λ_{max}. Moreover, analysis of the results obtained indicated that the factors that influence the exited energy such as energy related quantum descriptors (ε_{HOMO} and N'_C) play a leading role in the λ_{max}, and the electric interactions has a secondary influence factor.

4. Conclusions and perspective

In the present study, QSAR/QSPR studies developments to find the relationship between the molecular structure and some properties of different dyes have been summarized firstly. Then, we built 2D-QSAR models by 43 1,4-naphthoquinone derivatives to investigate the relationship between maximum absorption

2014 © The American Computational Science Society. All rights reserved.

of 1,4-naphthoquinone derivatives and their structures. Both of the two models were validated strictly by internal validation and external validation methods. The prefect statistical results of the two models proved their well predictive power. The general 2D-QSAR methods provided a rapid, simple and accurate method to predict the maximum absorption wavelength. We hope this paper can give some useful information in the field of dyes.

Acknowledgement

This work was financially supported by the Natural Science Foundation of Shandong Province (grant No. ZR2012BM016)

Reference

[1] P.J.T. Morris, A.S. Travis, 81 (1992) 59-100.
[2] S. Garfield, WW Norton & Company, (2002) 242.
[3] M. Karkmaz, E. Puzenat, C. Guillard, J.M. Herrmann, Applied Catalysis B: Environmental, 51 (2004) 183-194.
[4] C.G. Carumpalos, H.E. Pansing, Water base flexographic dye ink. In: Google Patents, 1979.
[5] B.D. Zelickson, S.L. Kilmer, E. Bernstein, V.A. Chotzen, J. Dock, D. Mehregan, C. Coles, Laser Surg. Med., 25 (1999) 229-236.
[6] J.M. Garden, L.L. Polla, O.T. Tan, Arch. Dermatol, 124 (1988) 889-896.
[7] M. Grätzel, J. Photoch Photobio C, 4 (2003) 145-153.
[8] A. Natansohn, P. Rochon, Chem. Rev., 102 (2002) 4139-4176.
[9] H. Zollinger, Wiley–VCH, Weinheim, 2003.
[10] V. Balzani, A. Credi, M. Venturi, Molecular Devices and Machines, Wiley–VCH, Weinheim, 2004.
[11] http://textiletodaybd.com/index.php?pid=magazine&id=212.
[12] W. Karcher, J. Devillers, Netherlands, 1990.
[13] G.R. Famini, C.A. Penski, L. Wilson, J phys. Org. Chem., 5 (1992) 395-408.
[14] T.W. Schultz, M.T.D. Cronin, T.I.J. Netzeva, Mol. Struc-THEOCHEM, 622 (2003) 23-38.
[15] A. R. Katritzky, D.C. Fara, R.O. Petrukhin, D.B. Tatham, U. Maran, A. Lomaka, M. Karelson, Curr. Top. Med. Chem., 2 (2002) 1333-1356.
[16] W.M.F. Fabian, S. Timofei, L. Kurunczi, J. Mol. Struc-THEOCHEM, 340 (1995) 73-81.
[17] W.M.F. Fabian, S. Timofei, J. Mol. Struc-THEOCHEM, 362 (1996) 155-162.
[18] T.I. Oprea, L. Kurunczi, S. Timofei, Dyes Pigments, 33 (1997) 41-64.
[19] S. Timofei, L. Kurunczi, T. Suzuki, W.M.F. Fabian, S. Muresan, Dyes Pigments, 34 (1997) 181-193.
[20] S. Timofei, M.F. Walte, J. Fabian, Chem. Inf. Comput. Sci., 38 (1998) 1218-1222.
[21] L. Kurunczi, F.S. Timofei, A. Bora, E.Seclaman, Int. J. Quantum Chem., 107 (2007) 2057-2065.
[22] S. Funar-Timofei, J. Chem. Inf. Comput. Sci., 42 (2002) 788-795.
[23] S. Timofei, L. Kurunczi, W. Schmidt, Z. Simon, SAR. QSAR Environ. Res., 13 (2002) 219-226.
[24] G. Schüürmann, S. Funar-Timofei, J. Chem. Inf. Comput. Sci., 43 (2003) 1502-1512.
[25] J. Polanski, R. Gieleciak, M. Wyszomirski, J. Chem. Inf. Comput. Sci., 43 (2003) 1754-1762.
[26] J. Polanski, R. Gieleciak, M. Wyszomirski, Dyes Pigments, 62 (2004) 61-76.
[27] N. Zhokhova, I. Baskin, V. Palyulin, A. Zefirov, N. Zefirov, Russ. J. Appl. Chem., 78 (2005) 1013-1017.
[28] J. Park, Rev. Prog. Coloration, 10 (1979) 20.
[29] M.R. De Giorgi, R. Carpignano, Dyes Pigments, 30 (1996) 79-88.
[30] M.R. De Giorgi, R. Carpignano, G. Crisponi, Dyes Pigments, 34 (1997) 1-12.
[31] M.R. De Giorgi, R. Carpignano, A. Cerniani, Dyes Pigments, 37 (1998) 187-196.
[32] J. Liu, Y. Ikushima, Z. Shervani, J. Supercrit Fluid, 32 (2004) 97-103.
[33] B. Buttingsrud, B.K. Alsberg, P.O. Strand, Phys. Chem. Chem. Phys., 9 (2007) 2226-2233.
[34] X. Xu, F. Luan, H.T. Liu, J.B. Cheng, X.Y. Zhang, Spectrochim Acta, A83 (2011)353.
[35] G. Fayet, D. Jacquemin, V. Wathelet, E.A. Perpte, P. Rotureau, C. Adamo, J. Mol. Graph. Model, 28 (2010) 465-471.
[36] M. Atabati, K. Zarei, M. Mohsennia, Anal. Chim.

2014 © The American Computational Science Society. All rights reserved.

Acta., 663 (2010) 7-10.
[37] J. Xu, H. Zhang, L.X. Wang, G.J. Liang, X.L. Shen, W. Xu, Spectrochim Acta, A76 (2010) 239.
[38] J. Xu, H. Zhang, L.X. Wang, G.J. Liang, X.L. Shen, Mol Simulat, 37 (2011) 1.
[39] A.D. McNaught, A. Wilkinson, IUPAC Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Oxford: Blackwell Scientific Publications, 1997.
[40] D.T.H.M. Sijm, G. Schüürmann, P.J. de Vries, A. Opperhuizen, Environ. Toxicol. Chem., 18 (1999) 1109-1117.
[41] R. Tabaraki, T. Khayamian, A.A. Ensafi, J. Ensafi, Mol. Graph. Model, 25 (2006) 46-54.
[42] R. Tabaraki, T. Khayamian, A. Ensafi, Dyes Pigments, 73 (2007) 230-238.
[43] A. Tarasova, F. Burden, J. Gasteiger, D.A. Winkler, J. Mol. Graph. Model, 28 (2010) 593-597.
[44] A. Garg, K.L. Bhat, C.W. Bock, Dyes Pigments, 55 (2002) 35-52.
[45] K.L. Bhat, S. Hayik, L. Sztandera, C.W. Bock, QSAR Comb. Sci, 24 (2005) 831-843.
[46] F. Pasha, M. Muddassar, H.W. Chung, S.J. Cho, H. Cho, J. Mol. Model, 14 (2008) 293-302.
[47] G.W. Byers, Macromolecules, 26 (1993) 4242-4248.
[48] A. Greaves, J. Churchley, M. Hutchings, D. Phillips, J. Taylor, Water Res, 35 (2001) 1225-1239.
[49] T. Suzuki, S. Timofei, L. Kurunczi, U. Dietze, G. Schüürmann, Chemosphere, 45 (2001) 1-9.
[50] Y. Lin, X.D. Li, J. Environ. Sci., 19 (2007) 800.
[51] G. Oros, T. Cserháti, E. Forgács, Chemosphere, 52 (2003) 185-193.
[52] A. Hopfinger, S. Wang, J. S. Tokarski, B.Q. Jin, M. Albuquerque, P.J. Madhav, C. Duraiswami, J. Am. Chem. Soc., 119 (1997) 10509-10524.
[53] H. Søsted, D. Basketter, E. Estrada, J. Johansen, G. Patlewicz, Contact Dermatitis, 51 (2004) 241-254.
[54] J. Mon, M. Flury, J.B. Harsh, J. Hydrol, 316 (2006) 84-97.
[55] E. Papa, J. Dearden, P. Gramatica, Chemosphere, 67 (2007) 351-358.
[56] H. Metivier-Pignon, C. Faur, P.L. Cloirec, Chemosphere, 66 (2007) 887-893.
[57] S. Funar-Timofei, L. Kurunczi, V. Vlaia, T. Olariu, D. Ciubotariu, 2008: World Scientific and Engineering Academy and Society (WSEAS), (2008) 108.
[58] E. Wilkowska, L. Konopski QSAR Comb. Sci., 27 (2008) 357-364.
[59] H.T. Liu, P. Han, Y.Y. Wen, F. Luan, Y. Gao, X.Y. Li, Dyes Pigments, 84 (2010) 148.
[60] ISIS Draw2.3, MDL Information Systems, Inc, 1990-2000.
[61] HyperChem 6.01. Hypercube. Inc, 2000.
[62] M.J.S. Dewar, D.M. Storch, J. Am. Chem. Soc.,107 (1985) 3898-3902.
[63] J.P.P. Stewart, MOPAC 6.0, Quantum chemistry program exchange, QCPE, No. 455. Bloomington, IN: Indiana University; (1989).
[64] A.R. Katritzky, V.S. Lobanov, M. Karelson CODESSA: training manual. Gainesville, FL: University of Florida, 1995.
[65] F. Luan, H. Liu, Y. Gao, Q.Z. Li, X. Zhang, Y. J. Guo, Colloid. Interface Sci., 336 (2009) 773.
[66] F. Luan, H.T. Liu, Y.Y. Wen, X.Y. Zhang, Flavour. Frag. J., 23 (2008) 232-238.
[67] J. Neter, W. Wasserman, M.H. Kutner, Applied Linear Statistical Models, 2nd ed., Irwin: Homewood, IL, (1985) 391.
[68] D.H. Cho, S.K. Lee, B.T. Kim, K.T. No, Bull Korean Chem Soc, 22 (2001)388.
[69] S. Riahi, E. Pourbasheer, R. Dinarvand, M.R. Ganjali, P. Norouzi, Chem. Biol. Drug. Des., 72 (2008) 575-584.
[70] X.J. Yao, A. Panaye, J.P. Doucet, R.S. Zhang, H.F. Chen, M.C. Liu, Z.D. Hu, B.T. Fan, J. Chem. Inf. Comput. Sci., 44 (2004) 1257-1266.
[71] Y.H. Xiang, M.C. Liu, X.Y. Zhang, R.S. Zhang, Z.D. Hu, J. Chem. Inf. Comput. Sci., 42 (2002) 592-597.
[72] M.J.L. Orr, Introduction to radial basis function networks. University of Edinburg 1996.
[73] M.J.L. Orr, MATLAB routines for subset selection and ridge regression in linear neural networks. Centre for Cognitive Science, Edinburgh University, 1996.
[74] M. Pavan, A.P. Worth, T.I. Netzeva, Preliminary analysis of an aquatic toxicity dataset and assessment of QSAR models for narcosis. JRC report EUR 21479 EN, 2005.

2014 © The American Computational Science Society. All rights reserved.

[75] A.C. Atkinson, Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis Clarendon Press Oxford, 1985.
[76] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci., 22 (2003) 69-77.
[77] P. Gramatica, QSAR Comb. Sci., 26 (2007) 694-701.
[78] J.G. Topliss, R.P. Edwards, J. Med. Chem., 22 (1979) 1238-1244.
[79] S. Wold, L. Eriksson, S. Clementi, Statistical validation of QSAR results. Chemometric methods in molecular design, VCH, (1995) 309.
[80] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2003) 69-77.
[81] A. Golbraikh, A. Tropsha J. Mol. Graph .Model, 20 (2002) 269-276.

2014 © The American Computational Science Society. All rights reserved.