Available online at www.sciencedirect.com
ScienceDirect
European Journal of Medicinal Chemistry 43 (2008) 2861–2869
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://www.elsevier.com/locate/ejmech

Original article

# A novel quantitative structure–activity relationship method to predict the affinities of MT3 melatonin binding site

Hongying Du [a], Jie Wang [a,b], Xiaoyun Zhang [a], Zhide Hu [a,*]

[a] *Department of Chemistry, Lanzhou University, Gansu, China*
[b] *Department of Biomedical Engineering, Yale University, New Haven, Connecticut, United States*

Received 24 November 2007; received in revised form 21 January 2008; accepted 7 February 2008
Available online 29 February 2008

**Abstract**

The linear regression (LR) and non-linear regression methods – grid search-support vector machine (GS-SVM) and projection pursuit regression (PPR) were used to develop quantitative structure–activity relationship (QSAR) models for a series of derivatives of naphthalene, benzofurane and indole with respect to their affinities to MT3/quinone reductase 2 (QR2) melatonin binding site. Five molecular descriptors selected by genetic algorithm (GA) were used as the input variables for the LR model and two non-linear regression approaches. Comparison of the results of the three methods indicated that PPR was the most accurate approach in predicting the affinities of the MT3/QR2 melatonin binding site. This confirmed the capability of PPR for the prediction of the binding affinities of compounds. Moreover, it should facilitate the design and development of new selective MT3/QR2 ligands.


*Keywords:* Melatonin; Quantitative structure–activity relationship; Genetic algorithm; Grid search-support vector machine; Projection pursuit regression

## 1. Introduction

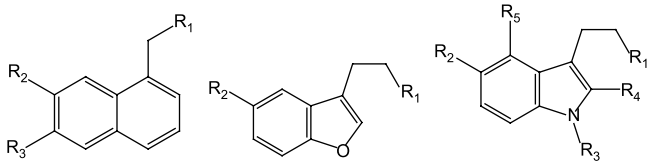
Melatonin (*N*-acetyl-5-methoxytryptamine) (compound no. **C29** in Table 1) is an indole-derived neurohormone of long-standing interest which is derived from serotonin, and is produced by the pineal gland during any dark period, whatever be the species considered, including humans [1–3]. Melatonin has been detected in numerous central and peripheral tissues using the specific radioligand 2-[$^{125}$I]-iodomelatonin [4,5]. As a consequence, melatonin is suspected to relay the circadian rhythm and the information on the photoperiod to the peripheral organs for daily and seasonal physiological regulations. Furthermore, melatonin has a proven role in the sleep/wake cycle [6], and is involved in numerous physiological functions depending on the circadian rhythm, such as the immune [7] and the cardiovascular systems [8]. These effects are mediated through activation of binding sites [9–11]. There are two high affinity melatonin receptors and a binding site, which have been identified to date. Among them, the MT1 [9] and MT2 [10] receptors have been cloned from human tissues. The pharmacology of these two receptors is well documented, and several compounds, including melatonin, are ligands with picomolar binding affinity [12]. Another putative melatonin binding site was identified on pharmacological grounds, with lower melatonin affinity (nanomolar range), very rapid ligand association/dissociation kinetics, and an original pharmacological profile [13–15]. In line with MT1 and MT2 receptors, this putative binding site was named MT3, according to the nomenclature recommendations of the IUPHAR [11], which was recently identified as the quinone reductase 2 (QR2) [3], an enzyme closely related to the detoxifying enzyme, quinine reductase 1. However, the physiological importance of the MT3/QR2 site is still unknown and it is particularly interesting to design and synthesize new selective ligands, which will provide pharmacological tools to assess and better characterize the role of this melatonin binding site.

Therefore, a major challenge to pharmaceutical scientists in drug discovery is to find an efficient way to get the affinity of

* Corresponding author. Tel.: +86 931 891 2540; fax: +86 931 891 2582.
*E-mail address:* hu_zhide@yahoo.com.cn (Z. Hu).

0223-5234/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2008.02.012

Table 1
Structures, MT3 affinities (pIC50) and the predicted values of naphthalene (Family A), benzofurane (Family B), and indole (Family C)




Family A Family B Family C

| No. | R1 | R2 | R3 | R4 | R5 | pIC50 | LR | SVM | PPR |
|---|---|---|---|---|---|---|---|---|---|
| **A1**[a] | $CH_2NHCO-i-C_3H_7$ | $OCH_3$ | H | | | 7.36 | 7.31 | 7.00 | 7.27 |
| **A2** | $CH_2NHCOCH_2$–*N*-pyrrolidin-2-one | $OCH_3$ | H | | | 7.21 | 7.15 | 7.26 | 7.24 |
| **A3** | $CH_2NHCOPh$ | $OCH_3$ | H | | | 7.54 | 7.51 | 7.35 | 7.51 |
| **A4** | $CH_2NHCO-c-C_4H_9$ | H | $OCH_3$ | | | 7.15 | 7.65 | 7.25 | 7.15 |
| **A5** | $CH_2NHSO_2CH_3$ | H | H | | | 7.55 | 4.26 | 7.48 | 7.70 |
| **A6** | $CH_2NHCOCH_2CH=CH_2$ | OH | H | | | 8.11 | 8.00 | 8.18 | 8.05 |
| **A7** | 5-Imidazolidine-2,4-dione | $OCH_3$ | H | | | 7.21 | 6.84 | 7.14 | 7.20 |
| **A8** | $CH_2$–*N*-oxazolidin-2-one | $OCH_3$ | H | | | 7.85 | 7.58 | 7.78 | 7.79 |
| **A9** | $CH_2NHCO-$2-furyl | $OCH_3$ | H | | | 7.27 | 7.40 | 7.34 | 7.26 |
| **A10** | $CH_2NHCO-$2-furyl | OH | H | | | 7.96 | 7.94 | 7.89 | 7.98 |
| **A11** | $CH_2NHCOCH_3$ | $SO_2NH_2$ | H | | | 7.49 | 7.78 | 7.56 | 7.49 |
| **A12** | $CH_2NHCOO-t-C_4H_9$ | OH | H | | | 7.29 | 7.95 | 7.42 | 7.34 |
| **A13** | $CH_2NHCO-$2-furyl | $SO_2NHCH_3$ | H | | | 8.04 | 8.14 | 7.78 | 8.05 |
| **A14** | $CH_2NHCOCH_3$ | $SCH_3$ | H | | | 7.74 | 7.85 | 7.67 | 7.69 |
| **A15** | $CH_2NHCOCH_3$ | $SO_2CH_3$ | H | | | 7.68 | 7.51 | 7.75 | 7.68 |
| **A16** | $CH_2NHCOCH_3$ | $SOCH_3$ | H | | | 7.43 | 7.38 | 7.42 | 7.44 |
| **A17**[a] | $NHCO-c-C_3H_7$ | $NHCOOCH_3$ | | | | 7.62 | 7.75 | 7.49 | 7.69 |
| **A18**[a] | $NHCOCH_3$ | $CONHCH_3$ | | | | 7.92 | 7.90 | 7.64 | 7.67 |
| **A19** | $NHCO-$2-furyl | $COOCH_3$ | | | | 7.28 | 7.26 | 7.31 | 7.19 |
| **B20** | $NHCO-$2-furyl | $CONH_2$ | | | | 7.17 | 2.58 | 7.10 | 7.23 |
| **B21** | $NHCO-c-C_5H_9$ | $COOCH_3$ | | | | 7.74 | 7.46 | 7.63 | 7.73 |
| **B22** | $NHCOPh$ | $NHCOOCH_3$ | | | | 7.17 | 7.20 | 7.24 | 7.14 |
| **B23** | $NHCO-i-C_3H_7$ | $NHCOOCH_3$ | | | | 7.64 | 7.86 | 7.57 | 7.61 |
| **B24** | $NHCOCH_2CH=CH_2$ | $NHCOOCH_3$ | | | | 7.96 | 7.83 | 8.03 | 7.99 |
| **B25** | $NHCO-$2-furyl | $NHCOOCH_3$ | | | | 7.25 | 7.36 | 7.32 | 7.30 |
| **B26**[a] | $NHCOCH_3$ | $OCH_3$ | | | | 7.19 | 8.00 | 7.62 | 7.55 |
| **B27** | $NHCOCH_3$ | $COOCH_3$ | | | | 7.85 | 7.56 | 7.52 | 7.79 |
| **B28** | $NHCOCH_3$ | $NHCOOCH_3$ | | | | 7.80 | 7.98 | 7.63 | 7.73 |
| **C29**[a] | $NHCOCH_3$ | $OCH_3$ | H | H | H | 7.19 | 7.74 | 7.38 | 7.27 |
| **C30** | $NHCOCH_3$ | $OCH_3$ | $CH_3$ | I | $NO_2$ | 9.89 | 10.07 | 9.96 | 9.91 |
| **C31**[a] | $NHCOCH_3$ | $OCH_3$ | H | I | $NO_2$ | 9.70 | 9.95 | 9.84 | 9.78 |
| **C32** | $NHCOCH_3$ | $OCH_3$ | H | $COOC_2H_5$ | H | 8.52 | 8.57 | 8.59 | 8.59 |
| **C33**[a] | $NHCOCH_3$ | $OCH_3$ | $CH_3$ | I | H | 10.05 | 9.74 | 9.57 | 9.63 |
| **C34** | $NHCOCH_3$ | $NHCOOCH_3$ | H | I | H | 9.52 | 9.40 | 9.59 | 9.51 |
| **C35** | $NHCOCH_3$ | $NHCOOCH_3$ | H | H | H | 7.24 | 7.45 | 7.56 | 7.25 |
| **C36** | $NHCOCH_3$ | $OCH_3$ | H | H | $NO_2$ | 8.92 | 8.51 | 8.44 | 8.85 |
| **C37** | $NHCOCH_3$ | $NO_2$ | H | H | H | 7.38 | 7.29 | 7.49 | 7.47 |
| **C38**[a] | $NHCOCH_3$ | $OCH_3$ | $CH_3$ | H | $NO_2$ | 9.51 | 8.72 | 8.84 | 9.11 |
| **C39**[a] | $NHCOCH_3$ | $OCH_3$ | $CH_3$ | H | H | 8.55 | 8.01 | 7.85 | 7.94 |

[a] Test set.

the new compounds for melatoninergic binding sites MT3 in early ligand discovery. The traditional methods are always time-consuming and costly, however, quantitative structure–activity relationship (QSAR) method provides a promising approach for the estimation of the affinity based on the descriptors solely derived from the molecular structures. The advantage of this method over the other approaches lies on the fact that it mainly requires the information of the chemical structure and is slightly dependent on the experimental data [16]. This way can develop a method for the prediction of the property of new compounds that have not been synthesized or found. It can also identify and describe the major structural features of the molecules that are relevant to molecular property variations. Once QSAR models are tested as efficient and creditable approaches, they can be used to estimate the activities of drugs, and guide to find a new ligand with high affinity. These methods have been widely used to predict the property and activity of drugs and compounds [11,16–19]. Due to the above reasons, it is necessary to develop these methods which will greatly improve work efficiency.

In this study, the descriptors based on the CODESSA software [20] calculated from structure alone were used to predict the affinities of the MT3/QR2 melatonin binding site. The genetic algorithm (GA) was used to select the most important

molecular descriptors, and build a linear regression (LR) model. Two non-linear models were constructed by using grid search-support vector machine (GS-SVM) and projection pursuit regression (PPR). The aim of this investigation was to explore the most major structural factors affecting the affinity of the MT3/QR2 melatonin binding site. The predicted results were very satisfactory for both training set and test set compounds. Furthermore, the information obtained from this work can be very helpful in the design and development of new selective MT3/QR2 ligands.

## 2. Experimental

### 2.1. Data set

In this study, the data set of 39 ligands was collected, whose affinities of MT3/QR2 melatonin binding site were reported in Ref. [11,21]. The affinity values were expressed as pIC50 (Table 1), which ranges from 7.15 (low affinity) to 10.05 (high affinity). The selected compounds belong to three structurally different families in terms of the different cyclic tensors. These tensors were naphthalene (Family A: 16 compounds), benzofurane (Family B: 12 compounds), and indole (Family C: 11 compounds). The data set was randomly divided into two subsets: the training set contained 30 compounds (76.9%) and the test set contained 9 compounds (23.1%). The training set was used to build a regression model, and the test set was used to evaluate the predictive ability of the obtained model.

### 2.2. Molecular descriptor generation

To obtain a QSAR model, the selected drugs were represented by the molecular descriptors. The calculation process of the molecular descriptors was shown as the following: two-dimensional structures of the compounds were drawn with the ISIS DRAW program. All of the structures were transferred into HyperChem 7.0 and pre-optimized using the MM+ molecular mechanics force field. A more precise optimization was done with the semi-empirical AM1 method in MOPAC, and then the structures of minimum energy were obtained. The resulting structures were transferred into the CODESSA software to calculate the descriptors. There were several kinds of descriptors obtained, including constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors. These descriptors could represent a variety of aspects in the compounds, and had been successfully used in various QSAR and QSPR researches [22–24].

### 2.3. Principal component analysis (PCA) of the data set

The diversity of the training set and the test set was analyzed using the principal component analysis (PCA) method. Using all the descriptors generated by the CODESSA software, PCA was used to deduce the dimensions of the descriptors by dropping the unnecessary data information. In order to do PCA, the constant descriptors and some descriptors with missing values must be excluded. After this step, the PCA method was used for analysis, for which PC1, PC2, and PC3 made 19.90%, 16.02%, and 13.13% contribution to the total PCs, respectively. In all these three PCs made a total of 49.05% of the variation in the data, and played major roles. It should be noted that all loading plots showed similar trends, therefore, only the PC1, PC2, and PC3 loading plots were shown for the compounds. Fig. 1 illustrates the distribution of compounds over the first three principal component space. Inspecting this figure, it could be concluded that samples in both the training and the test sets seemed to be evenly scattered in the 3D space. So it confirmed that it was feasible for the splitting of the data set. Moreover, the compounds in the training set were representative of the whole data.

### 2.4. QSAR model development and evaluation

After analyzing splitting of the data set into the diversity of the training set and test set, the next step was to select the main factors which were the most important for the affinity toward the MT3 binding site. As a powerful tool in searching the most suitable parameters [25], GA was used to select the most important popular molecular descriptors. In the present work, five molecular descriptors (see Table 2) were selected. Based on the selected descriptors, the LR and two non-linear models (GS-SVM and PPR) were constructed.

After the regression model was constructed, the root mean square error (RMSE) and the absolute average relative deviation (AARD) were used to evaluate the model's predictive performance; they were calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{\mathrm{exp}} - y_{\mathrm{pred}}\right)^2}{n}}$$

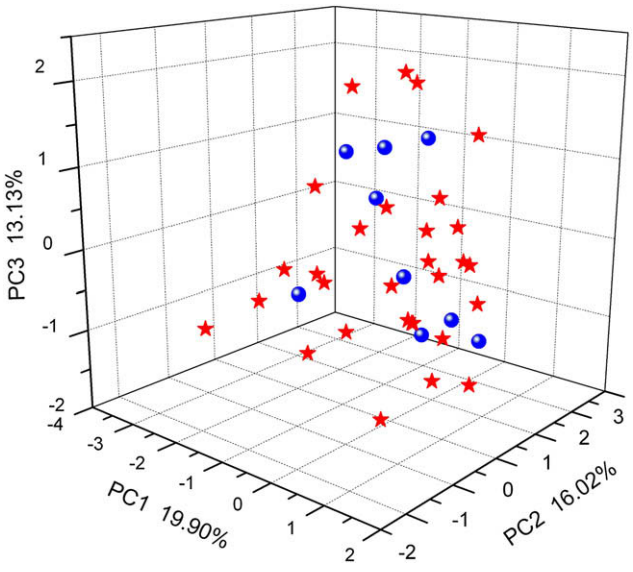



Fig. 1. The principal component analysis of the training set and the test set.

Table 2
The involved parameters and the statistical parameters of the linear regression (LR) model

| Descriptors | Meaning | Coefficient | *t*-test | $t_p$ |
|---|---|---|---|---|
| Constant | | 108.15 | | |
| MW | Molecular weight | 0.01 | 7.16 | 2.13E-07 |
| BI | Balaban index | 2.40 | 10.49 | 1.90E-10 |
| PPSA-1/TMSA | FPSA-1 fractional PPSA (PPSA-1/TMSA) [Zefirov's PC] | 6.30 | 5.46 | 1.30E-05 |
| MERIO | Min electroph. react. index for an O atom | −2237.72 | −2.41 | 2.37E-02 |
| MBOH | Max bond order of an H atom | −114.15 | −6.52 | 9.60E-07 |
| $n = 30$, $R^2 = 0.8777$, adjusted $R^2 = 0.8522$, $R^2_{cv} = 0.7686$, $F = 34.45$ (95% confidence level) | | | | |

$$\mathrm{AARD} = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_{\mathrm{exp}} - y_{\mathrm{pred}}|}{y_{\mathrm{exp}}}$$

where $y_{\mathrm{exp}}$ and $y_{\mathrm{pred}}$ are the experimental and predicted data for the selected compounds. It can be seen that the lower the RMSE and AARD, the more accurate is the model obtained.

## 3. The theory of the modeling methods

### 3.1. Genetic algorithm (GA)

GA is a stochastic optimization technique [26,27], which derives from the concepts of biological process of inheritance: natural selection, evolution, mutation, and the genetic crossover. It has been widely used to solve the variable selection problems [28,29]. The basic theory of GA could be found in Refs. [26,27]; here, we only briefly summarize the main ideas of GA. In the research, the variables are represented as genes on a chromosome, and they are generally coded as binary strings. Through a simulated natural selection and the action of the genetic operators mutation and recombination, chromosomes that satisfy at best to a predefined fitness function are found. The fitness function is deduced from the gene composition of a chromosome. So it contains many procedures: parameters and fitness, representation, populations and generation, selection, crossover and mutation, replacement, termination, *etc*. These procedures have been described in detail in Ref. [29]. When adding another descriptor does not improve the fitness significantly, the best variable selection is obtained.

### 3.2. Support vector machine (SVM)

SVM was developed by Vapnik, and gained popularity due to its many attractive features and promising empirical performance [30,31]. SVM has been used for classification, regression, and function approximation works. A thorough discussion of the theory of SVM was provided by Cristianini and Shawe-Taylor [32]. So only a brief introduction to SVM will be given here. The excellent properties of SVM embody the structural risk minimization (SRM) principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle. SRM minimizes an upper bound on VC dimension (''generalization error''), as opposed to ERM that minimizes the error on the training data. It is the difference that equips SVM with good generalization performance, which is the goal in statistical learning. Originally, SVM was developed for pattern recognition problems [33] and now, with the introduction of $\varepsilon$-insensitive loss function, SVM has been widely used to solve non-linear regression estimation. The estimated function is a linear expansion in terms of functions defined on a certain subset of the data (support vectors), and the final number of coefficients in such an expansion does not depend on the dimensionality of the space of input variables. These two properties make SVM an especially useful technique for dealing with very large data sets in a high-dimensional space.

### 3.3. Projection pursuit regression (PPR)

PPR developed by Friedman and Stuetzle [34] is a powerful tool for seeking interesting projections of high-dimensional data into lower-dimensional space and, therefore, can overcome the curse of dimensionality. At present, it has been successfully applied to tackle some chemical problems [35,36]. Friedman and Stuetzle's concept of PPR avoided many difficulties experienced with other existing non-parametric regression procedures. It does not split the predictor space into two regions thereby allowing, when necessary, more complex models. In addition, interactions of predictor variables are directly considered since linear combinations of the predictors are modeled with general smooth functions. The basic theory of PPR can be found in Refs. [37,38]. Here, we only give a brief description. Given the $(k \times n)$ data matrix $X$, where $k$ is the number of observed variables and $n$ is the number of units, and an $m$-dimensional orthonormal matrix $A$ $(m \times k)$, the $(m \times n)$ matrix $Y = AX$ represents the coordinates of the projected data onto the $m$-dimensional $(m < k)$ space spanned by the rows of $A$. As such projections are infinite, it is important to have a technique to pursue a finite sequence of projections that can reveal the most interesting structures of the data. Projection pursuit (PP) is such a powerful tool that combines both ideas of projection and pursuit [37,38]. In a typical regression problem, PPR aims to approximate the regression pursuit function $f(x)$ by a finite sum of ridge functions with suitable choices of $\alpha_i$ and $g_i$.

$$g^{(p)}(x) = \sum_{i=1}^{p} g_i\left(\alpha_i^T x\right) \quad (1)$$

where $\alpha_i$ are $m \times n$ orthonormal matrices, $p$ is the number of ridge functions.

All calculation programs implementing SVM and PPR were written in R-file under R2.3.1 [39] environment running operating system on a Pentium IV with 512M RAM.

## 4. Results and discussion

### 4.1. Results of genetic algorithm–linear regression model

In this QSAR study, GA was used to select the main descriptors and build the linear model. A variety of subset sizes of descriptors were investigated to determine the optimum number of descriptors in the regression model. If adding another descriptor did not significantly improve the statistics of the model, it was determined that the optimum subset size had been achieved. The influences of the number of the descriptors on the coefficients of determination ($R^2$) and the RMSE to the training set and test set are shown in Figs. 2 and 3, respectively. The higher the values of $R^2$ for the training and test sets and the lower the RMSE, the better the results. From Figs. 2 and 3, it was clearly seen that five descriptors were the best selection. The involved parameters and the statistical parameters of this model are summarized in Table 2, and the correlation matrix of these selected descriptors is shown in Table 3. Table 4 shows the statistical results of the LR model for the training and test sets, and the predicted affinity values are listed in Table 1, and Fig. 4 shows the predicted pIC50 *vs.* experimental values for all of the 39 compounds.

The selected molecular descriptors contained one constitutional descriptor – molecular weight (MW), one topological descriptor – Balaban index (BI), one electrostatic descriptor – FPSA-1 fractional PPSA (PPSA-1/TMSA) [Zefirov's PC] (PPSA-1/TMSA), two quantum chemical descriptors – min electroph. react. index for an O atom (MERIO) and max bond order of an H atom (MBOH). By interpreting the meaning of these descriptors, we could get important structural information, which was related with the affinities of the ligands with MT3/QR2. The constitutional descriptor MW reflects only the molecular composition of the compound without using the geometry or electronic structure of the molecules, and it is calculated from the atomic masses and the number of the corresponding atoms. The topological descriptor BI [40] is defined by the following formula:

$$J = \left(\frac{q}{\mu + 1}\right) \sum_{i,j}^{q} (S_i S_j)^{-1/2}$$

where $q$ is the number of edges in the molecular graph, $\mu = q - n + 1$ is the cyclometric number, $n$ is the number of vertices in the graph and $S_i$, $S_j$ – the distance sums (or distance degrees), obtained by summation on the row $i$ and column $i$ (or row $j$ and column $j$, respectively) of the distance matrix between atoms in the molecule. It describes the atomic connectivity and branching information in the molecule and has some correlation with the hydrophobic interaction of the molecules. The electrostatic descriptor PPSA-1/TMSA belongs to the charged partial surface area (CPSA) descriptor, which was invented by Jurs et al. [41,42] in terms of the whole surface area of the molecule and in terms of functional group portions. It encodes the features responsible for polar interactions between molecules. The quantum chemical descriptors – MERIO encodes the polarity of the molecules, and it is related with the

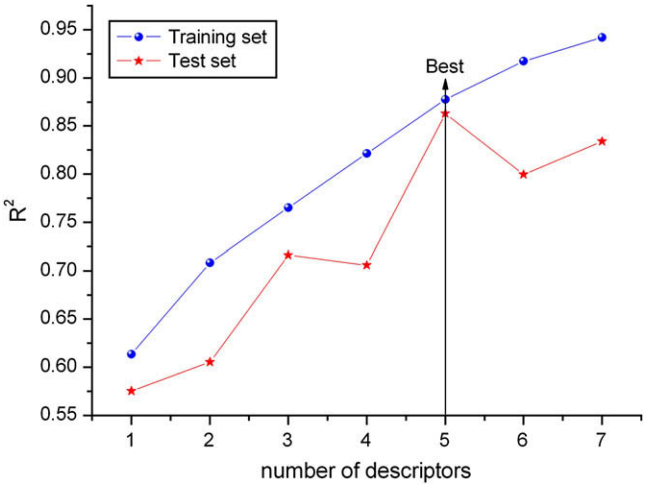



Fig. 2. Influence of the number of descriptors on the coefficients of determination ($R^2$) of the training set and the test set.

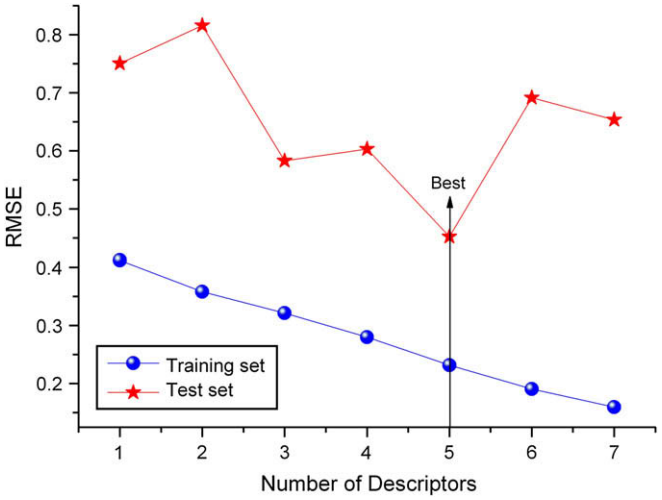



Fig. 3. Influence of the number of descriptors on the root mean square error (RMSE) of the training set and the test set.

Table 3
The correlation matrix of the selected molecular descriptors

| | MW | BI | PPSA-1/TMSA | MERIO | MBOH |
|---|---|---|---|---|---|
| MW | 1 | | | | |
| BI | −0.0960 | 1 | | | |
| PPSA-1/TMSA | 0.3654 | −0.3061 | 1 | | |
| MERIO | −0.2073 | −0.1259 | −0.1926 | 1 | |
| MBOH | 0.2128 | 0.4465 | 0.3532 | −0.4147 | 1 |

Table 4
The comparison of different regression models (LR, GS-SVM, PPR)

| Model | Data set | $R^2$ | RMSE | AARD (%) |
|---|---|---|---|---|
| Training set | Linear regression | 0.8777 | 0.232 | 2.31 |
| | GS-SVM | 0.9475 | 0.153 | 1.46 |
| | PPR | 0.9938 | 0.052 | 0.52 |
| Test set | Linear regression | 0.8630 | 0.452 | 4.53 |
| | GS-SVM | 0.8827 | 0.427 | 4.47 |
| | PPR | 0.9344 | 0.320 | 3.07 |

ability of forming the hydrogen bond. The other MBOH also is related to the ability of forming a hydrogen bond.

From the above discussion, the selected descriptors all had explicit physical meaning, and could reflect different aspects of the molecule. Based on the interpretation of these descriptors, it could be clearly seen that the affinity between the ligand and MT3/QR2 was a complicated problem. They were related to several properties of the molecular structures, such as the compositions, steric factors, polarity, the ability to form hydrogen bond of the ligands, and the hydrophobic interaction between the ligands and MT3/QR2. These properties are consistent with the result in Ref. [11].

### 4.2. Results of GS-SVM model

From Table 1 and Fig. 4, it can be seen that the linear model was not sufficiently accurate, and the factors influencing the affinity values of drugs were complicated and not all of them were in linear correlation with the affinities. Therefore, the non-linear model was built by GS-SVM based on the same subset of descriptors. Similar to other statistical methods, the performance of SVM was influenced by several parameters. These parameters included the type of kernel function, $\varepsilon$ of $\varepsilon$-insensitive loss function and the capacity parameter $C$. There were several kernel functions used in training and predicting, such as linear, polynomial, sigmoid, radial basis function, *etc*. However, radial basis function was commonly used, for its good general performance and few parameters. The form of the radial basis function in $R$ is

$$K(x_i, x) = \exp\{-\gamma|x - x_i|\}^2$$

where $\gamma$ is a constant, the parameter of the kernel; $x$ and $x_i$ are two independent variables; $\gamma$ controls the amplitude of the Gaussian function, therefore, it controls the generalization ability of SVM. So it is very important to find an optimum value for $\gamma$.

The optimal value for $\varepsilon$ depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for $\varepsilon$, there is the practical consideration of the number of resulting support vectors. $\varepsilon$-Insensitivity prevents the entire training set meeting boundary conditions and so allows for the possibility of sparsity in the dual formulation's solution. So, choosing the appropriate value of $\varepsilon$ is critical from theory. The last parameter $C$ was a regularization parameter that controlled the tradeoff between maximizing the margin and minimizing the training error.

In the previous research, traditional SVM was used to find the optimal values for these parameters to solve the regression problems. It usually used a single fact analysis method to find the best model. The results of this method would be local optimization. In fact, the factors of SVM influenced each other, when it was used to accomplish the regression problem, so it was not the best choice. In our work, GS-SVM was used to obtain the global optimization. This method used multifact correlation analysis to find the best model, so it was more effective than the traditional SVM. Using the grid search method, it was concluded that the parameter '$\varepsilon$' influenced the results slightly. So we only show the relationship between $C$, $\gamma$ and the important statistical parameters of the regression models (Fig. 5 (A–D)). The higher the values of $R^2$ for the training set and test set, the better the results, and the lower the RMSE and AARD, the better the results. So the best values were selected as 140, 0.02, and 0.05 for $C$, $\gamma$, and $\varepsilon$, respectively. The predicted results are shown in Table 1 and Fig. 6. The statistical parameters for the best regression model are shown in Table 4. On comparing, the non-linear regression model SVM was better than the results of linear regression, but not very satisfying.

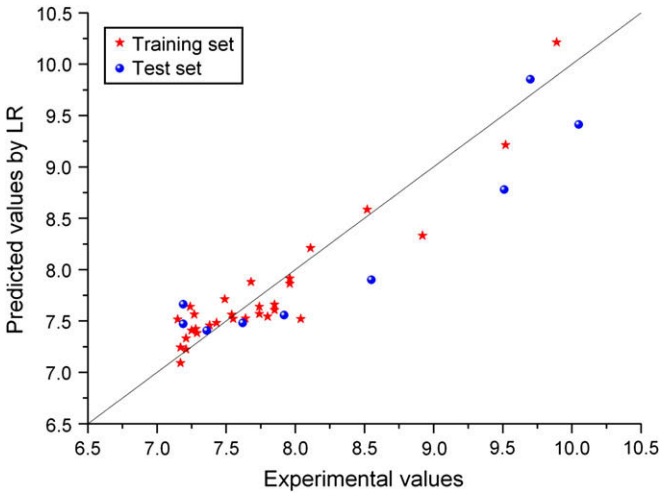



Fig. 4. Experimental values *vs*. predicted values for the training set and the test set by linear regression (LR).

### 4.3. Results of PPR model

In order to find a more accurate model, we also tried to use another non-linear regression method – PPR, which is a simple but powerful tool for seeking interesting projections of high-dimensional data into lower-dimensional space and, therefore, can overcome the curse of dimensionality. In this investigation, the PPR algorithm also had several parameters, which needed to be adjusted, such as 'nterms', 'max. terms', 'optlevel', and 'span'. The parameter 'nterms' controls the number of variables to be entered in the model, 'max. terms' is the maximum number of terms to choose from when building the model, 'optlevel' means the levels of optimization which

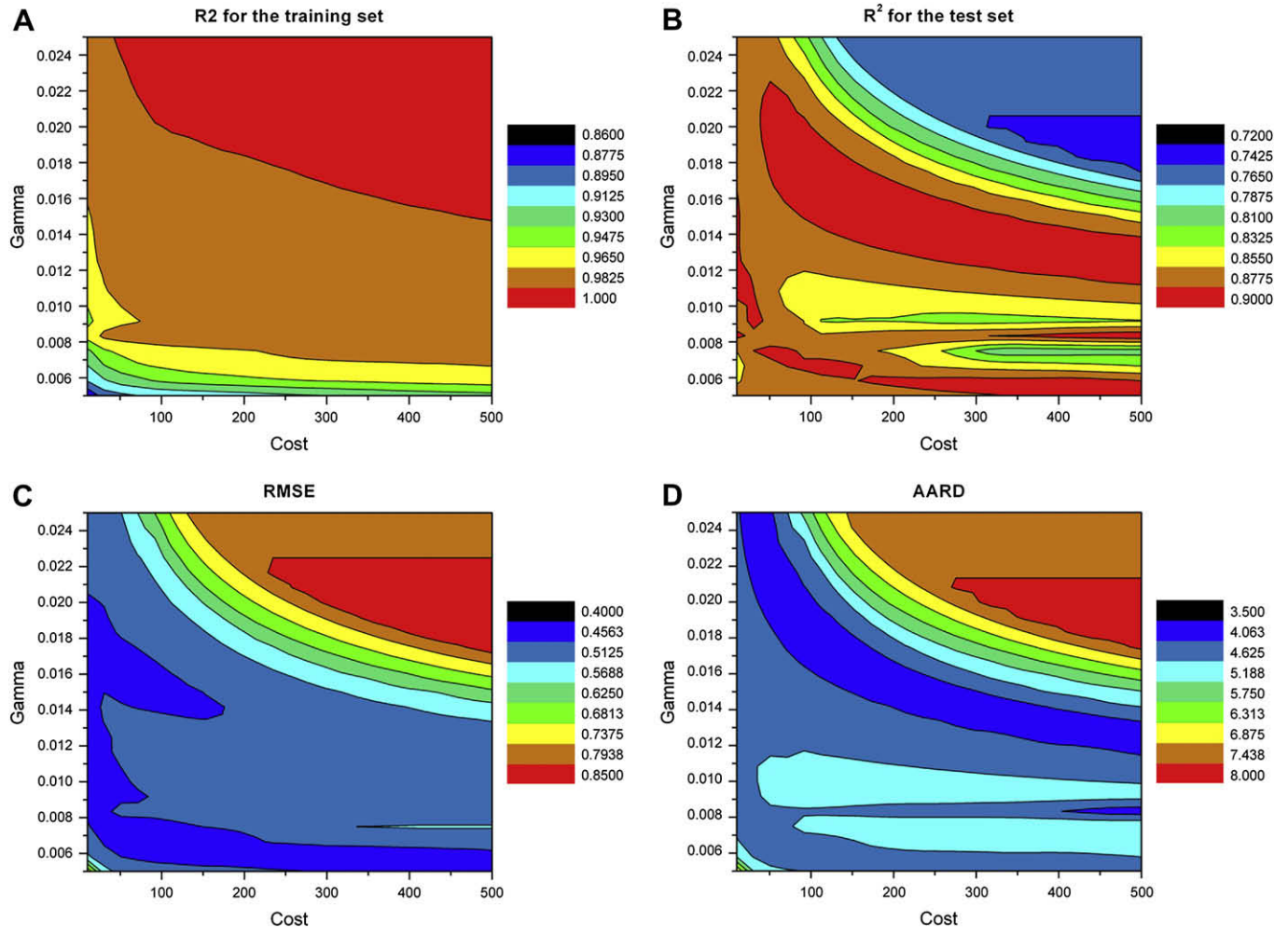



Fig. 5. Using the grid search method to find the best parameters of SVM in predicting the training set and the test set. (A) the coefficients of determination ($R^2$) for the training set; (B) the coefficients of determination ($R^2$) for the test set; (C) the root mean square error (RMSE) for the test set; (D) the absolute average relative deviation (AARD) for the test set.

differ in how thoroughly the models are refitted during this process, and 'span' defines the fraction of the observations in the span of running the lines smoother. Here, we also used the grid search method to find the best values for these parameters. The optimal values for 'nterms', 'max. terms', 'optlevel', and 'span' were determined as 5, 3, 9 and 0, respectively. The results of this method are shown in Table 1 and Fig. 7; the statistical parameters of this regression model are collected in Table 4. It could be clearly seen that this method was the most efficient way to search the affinities for these drugs.

### *4.4. Comparison of the results obtained by different approaches*

The prediction results by the three methods, LR, GS-SVM and PPR are collected in Table 1, and the statistical parameters for these three methods are listed in Table 4. From the comparison, it could be clearly seen that the non-linear regression methods gave promising results. Although we improved the traditional SVM method and tried to get the best results, the results were not far improved. However, the other simple non-linear regression method PPR gave very good prediction results compared to the other two methods. In Ref. [11], the 3D-comparative molecular field analysis (CoMFA) method was employed to predict inhibitory activity of the same compounds. The coefficients of determination were 0.897 and 0.875 for the training set and test set, respectively. By comparing the results of CoMFA and PPR methods, it also concluded that the PPR method was a very promising tool to predict the affinities between the MT3/QR2 and ligands.

## 5. Conclusions

In our research, the classical feature selection method GA was used to select the main relevant descriptors and build a linear model. The GS-SVM and PPR methods were used to construct the non-linear QSAR model based on the same selected parameters. Both the linear and non-linear models provided good results, at the same time, the non-linear PPR models

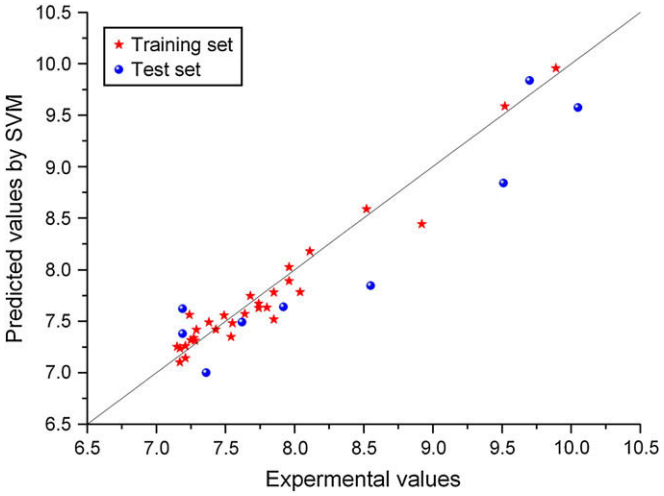



Fig. 6. Experimental values *vs*. predicted values for the training set and the test set by support vector machine (SVM).

produced the best results with good predictive capability, so the following conclusions could be obtained: (a) the classical GA method was an effective method for variable selection, and the selected parameters could account for the fact that the structural features of the compounds were related to the affinities between MT3/QR2 and ligands; (b) non-linear relationship could accurately describe the relationship between the structural parameter and the affinities of the MT3/QR2 ligands; (c) PPR was proved to be a very useful tool in the prediction of the affinities of the ligands, and it was a very promising machine learning method and would gain more extensive applications. In short, our study has found an efficient way to research the affinities between the MT3/QR2 melatonin binding site and the ligands, and should facilitate the design and development of new selective MT3/QR2 ligands. Furthermore, the proposed approach can also be extended to other similar QSAR investigations.

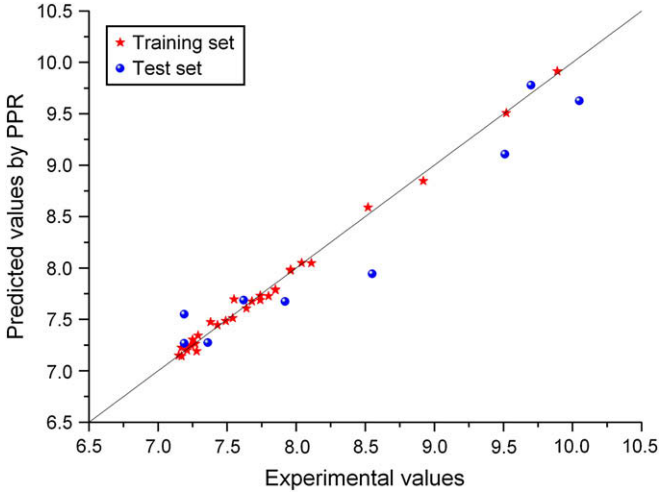



Fig. 7. Experimental values *vs*. predicted values for the training set and the test set by projection pursuit regression (PPR).

## Acknowledgements

The authors would like to thank the R Development Core Team for the R2.3.1 software, and also would like to express their gratitude to Lanzhou University foreign teacher Allan Grey who thoroughly corrected the English in the paper. Special thanks to the anonymous reviewers, and to the Editor for their professional, extensive and useful comments.

## References

[1] O. Nosjean, J.P. Nicolas, F. Klupsch, P. Delagrange, E. Canet, J.A. Boutin, Biochem. Pharmacol. 61 (2001) 1369–1379.
[2] R.J. Reiter, Endocr. Rev. 12 (1991) 151–180.
[3] O. Nosjean, M. Ferro, F. Cogé, P. Beauverger, J.M. Henlin, F. Lefoulon, J.L. Fauchère, P. Delagrange, E. Canet, J.A. Boutin, J. Biol. Chem. 275 (2000) 31311–31317.
[4] P.J. Morgan, P. Barrett, H.E. Howell, R. Helliwell, Neurochem. Int. 24 (1994) 101–146.
[5] P. Delagrange, B. Guardiola-Lemaitre, Clin. Neuropharmacol. 20 (1997) 482–510.
[6] J.R. Redman, S.M. Armstrong, K.T. Hg, Science 219 (1983) 1089–1091.
[7] G.J. Maestroni, J. Pineal. Res. 14 (1993) 1–10.
[8] E. Scalbert, B. Guardiola-Lemaitre, P. Delagrange, Therapie 53 (1998) 459–465.
[9] S.M. Reppert, D.R. Weaver, T. Ebisawa, Neuron 13 (1994) 1177–1185.
[10] S.M. Reppert, C. Godson, C.D. Mahle, D.R. Weaver, S.A. Slaugenhaupt, J.F. Gusella, Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 8734–8738.
[11] A. Farce, S. Dilly, A. Sabaouni, S. Yous, P. Berthelot, J.A. Boutin, P. Delagrange, P. Renard, P. Chavatte, QSAR Comb. Sci. 26 (2007) 820–827.
[12] M.L. Dubocovich, Trends Pharmacol. Sci. 16 (1995) 50–56.
[13] M.J. Duncan, J.S. Takahashi, M.L. Dubocovich, Endocrinology 122 (1988) 1825–1833.
[14] E.J. Molinari, P.C. North, M.L. Dubocovich, Eur. J. Pharmacol. 301 (1996) 159–168.
[15] P. Paul, C. Lahaye, P. Delagrange, J.P. Nicolas, E. Canet, J.A. Boutin, J. Pharmacol. Exp. Ther. 290 (1999) 334–340.
[16] A.R. Katritzky, U. Maran, V.S. Lobanov, M. Karelson, J. Chem. Inf. Comput. Sci. 40 (2000) 1–18.
[17] B. Hemmateenejad, K. Javadnia, M. Elyasi, Anal. Chim. Acta 592 (2007) 72–81.
[18] J. Wang, H.Y. Du, H.X. Liu, X.J. Yao, Z.D. Hu, B.T. Fan, Talanta 73 (2007) 147–156.
[19] M. Daszykowski, I. Stanimirova, B. Walczak, F. Daeyaert, M.R. de Jonge, J. Heeres, L.M.H. Koymans, P.J. Lewi, H.M. Vinkers, P.A. Janssen, D.L. Massart, Talanta 68 (2005) 54–60.
[20] http://www.codessa-pro.com/.
[21] V. Leclerc, S. Yous, P. Delagrange, J.A. Boutin, P. Renard, D. Lesieur, J. Med. Chem. 45 (2002) 1853–1859.
[22] M.C. García-Álvarez-Coque, J.R. Torres-Lapasió, J.J. Baeza-Baeza, Anal. Chim. Acta 579 (2006) 125–145.
[23] J. Kolar, A. Štolfa, M. Strlič, M. Pompe, B. Pihlar, M. Budnar, J. Simčič, B. Reissland, Anal. Chim. Acta 555 (2006) 167–174.
[24] F. Luan, C.X. Xue, R.S. Zhang, C.Y. Zhao, M.C. Liu, Z.D. Hu, B.T. Fan, Anal. Chim. Acta 537 (2005) 101–110.
[25] D. Rogers, A.J.J. Hopfinger, J. Chem. Inf. Comput. Sci. 34 (1994) 854–866.
[26] C.B. Lucasius, G. Kateman, Trends Anal. Chem. 10 (1991) 254–261.
[27] R. Leardi, J. Chemometr. 15 (2001) 559–569.
[28] N.A.M. Barakat, J.H. Jiang, Y.Z. Liang, R.Q. Yu, Chemometr. Intell. Lab. 72 (2004) 73–82.
[29] M. Maeder, Y.-M. Neuhold, G. Puxty, Chemometr. Intell. Lab. 70 (2004) 193–203.
[30] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, Berlin, 1982.
[31] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[32] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000.
[33] C.J.C. Burges, Data Min. Knowl. Disc. 2 (1998) 1–47.
[34] J.H. Friedman, W. Stuetzle, J. Am. Stat. Assoc. 76 (1981) 817–823.
[35] P.J. Huber, Ann. Stat. 13 (1985) 435–475.
[36] P. Diaconis, M. Shahshahani, SIAM J. Sci. Stat. Comput. 5 (1984) 175–191.
[37] Y.P. Du, Y.Z. Liang, D. Yun, J. Chem. Inf. Comput. Sci. 42 (2002) 1283–1292.
[38] H.X. Liu, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Talanta 71 (2007) 258–263.
[39] http://www.r-project.org/.
[40] A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA: Reference Manual, University of Florida, Gainesville, Florida, 1994.
[41] D.T. Stanton, P.C. Jurs, Anal. Chem. 62 (1990) 2323–2329.
[42] D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci. 32 (1992) 306–316.