Available online at www.sciencedirect.com
ScienceDirect
European Journal of Medicinal Chemistry 43 (2008) 364–372
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://www.elsevier.com/locate/ejmech

Original article

A chemometric study of the 5-HT_{1A} receptor affinities presented by arylpiperazine compounds

Karen C. Weber, Albérico B.F. da Silva*

Instituto de Química de São Carlos, Universidade de São Paulo, P.O. Box 780, 13566-590 São Carlos, SP, Brazil

Received 5 December 2006; received in revised form 28 March 2007; accepted 29 March 2007
Available online 24 April 2007

Abstract

Arylpiperazine compounds are promising 5-HT_{1A} receptor ligands that can contribute for accelerating the onset of therapeutic effect of selective serotonin reuptake inhibitors. In the present work, the chemometric methods HCA, PCA, KNN, SIMCA and PLS were employed in order to obtain SAR and QSAR models relating the structures of arylpiperazine compounds to their 5-HT_{1A} receptor affinities. A training set of 52 compounds was used to construct the models and the best ones were obtained with nine topological descriptors. The classification and regression models were externally validated by means of predictions for a test set of 14 compounds and have presented good quality, as verified by the correctness of classifications, in the case of pattern recognition studies, and by the high correlation coefficients ($q^2 = 0.76$, $r^2 = 0.83$) and small prediction errors for the PLS regression. Since the results are in good agreement with previous SAR studies, we can suggest that these findings can help in the search for 5-HT_{1A} receptor ligands that are able to improve antidepressant treatment.

Keywords: Arylpiperazines; 5-HT_{1A} receptor; QSAR; Chemometrics

1. Introduction

Selective serotonin reuptake inhibitors (SSRIs) are the most effective class of antidepressants in current clinical use. However, they present the serious drawback of a delay of two to six weeks in the onset of therapeutic effect. This delay can be attributed to the need of SSRI to overcome the inhibitory influence of 5-HT_{1A} receptor, which reduces neuronal firing rate and neurotransmitter release. With constant administration these receptors are desensitized and the SSRI is able to show its clinical effects [1–3].

Clinical studies have shown that when a 5-HT_{1A} receptor antagonist, such as pindolol or WAY 100635, is administrated along with different SSRIs, an increase of extra-cellular serotonin concentration in terminal areas is observed [4–6]. So, the combination of a 5-HT_{1A} receptor antagonist and a SSRI may accelerate the onset of antidepressant action improving the efficacy of the pharmacological treatment of depression [7].

The most important class of 5-HT_{1A} receptor ligands are arylpiperazine compounds (see general structure in Fig. 1). Some QSAR studies were performed on different series of arylpiperazines, and the main results have indicated certain structural features required for high 5-HT_{1A} receptor affinity [8–11]. In a CoMFA study performed on a series of 48 arylpiperazines [8], and afterwards in a study employing Hansch and ANN analyses on 32 compounds [9], Lopez-Rodriguez et al. suggested that voluminous substituents at *ortho* and *meta* positions of the aromatic ring Ar_1 may contribute for high 5-HT_{1A} receptor affinities, as well as a chain of three to four CH_2s between Ar_2 and N1 (see Fig. 1). Another CoMFA study [10] had previously indicated the important factors for optimal 5-HT_{1A} binding as the following: (i) a favorable steric region close to Ar_1; (ii) a sterically forbidden region near the basic nitrogen; (iii) a favorable negative charge close to the *ortho* position of Ar_1 and (iv) opposite lipophilic and electrostatic effects on the nitrogen substituent [10].

* Corresponding author. Tel./fax: +55 16 33739975.
E-mail address: alberico@iqsc.usp.br (A.B.F. da Silva).

0223-5234/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2007.03.036

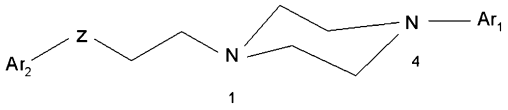


Fig. 1. General structure of arylpiperazine compounds.

Furthermore, in a Multiple Linear Regression study [11], a good correlation ($q^2 = 0.83$) between theoretical descriptors and affinity constant values was obtained, where the electrostatic interaction between the protonated amine function and a primary nucleophilic site of the receptor necessary for recognition were described by the molecular orbital (MO) indexes localized on the NH^+ group, while short-range attractive and repulsive intermolecular interactions were described by MO indexes computed on the whole molecules (polar and dispersive forces), and by *ad hoc* defined size and shape descriptors (dispersive and steric forces) [11].

In the search for a new class of antidepressant compounds with dual activity as 5-HT_{1A} receptor antagonists and as serotonin reuptake inhibitors, Martinez-Esparza et al. synthesized a series of arylpiperazine compounds presenting structural moieties related to serotonin reuptake and also other moieties derived from typical 5-HT_{1A} receptor ligands [12]. In the present work our aim is to construct reliable, qualitative and quantitative structure–activity relationship models for arylpiperazine compounds concerning their affinities to the 5-HT_{1A} receptor. In order to identify theoretical descriptors that can be related to these affinities, we have calculated a large amount of electronic, structural, and topological descriptors to be employed in analyses performed with the chemometric methods PCA (Principal Component Analysis), HCA (Hierarchical Cluster Analysis), KNN (K-nearest Neighbor), SIMCA (Soft Independent Modeling of Class Analogy), and PLS (Partial Least Squares), by using a protocol described previously [13–16].

2. Methodology

2.1. Data set

From the arylpiperazine compounds synthesized by Martinez-Esparza et al. [12] we selected 52 compounds to constitute the training set (shown in Table 1) and 14 compounds for a test set to be used in the external validation of the final chemometric models (see Table 2). The choice of compounds was done in a manner that the two sets presented structural diversity and a good distribution of the biological property (pK_i values ranging from 5.30 to 8.30). The set of compounds was split into two classes: compounds with pK_i values greater than 6.70 were assumed as *higher 5-HT_{1A} affinity compounds* (Class 1) and those with pK_i <6.70 were considered as *lower 5-HT_{1A} affinity compounds* (Class 2).

2.2. Geometry optimization and descriptor calculation

The structures of all compounds were initially optimized with the molecular mechanics method MM+ [17,18]. A final geometry optimization was then performed by using the semi-empirical method AM1 [19,20]. From these structures, 10 electronic descriptors were calculated with the AM1 method, and additionally 567 topological descriptors were calculated with the Dragon 2.1 software [21]. All these descriptors were assumed to represent electronic, structural and topological properties of the compounds that can be correlated to the 5-HT_{1A} receptor affinities experimentally observed. The chemometric analyses were performed with the Pirouette software [22], after autoscaling the variable values in order to give the same importance for all of them.

2.3. Variable selection

In order to reduce the high number of calculated descriptors we have computed the Fisher's weight, W_{1-2}, of each descriptor. In this way, we were able to select the ones presenting good discriminating power to distinguish the higher and the lower affinity compounds [23]. The W_{1-2} for the i descriptor and for samples belonging to the given Classes 1 and 2 is calculated through Eq. (1):

$$W_{1-2}(i) = \frac{[\overline{X}_i(1) - \overline{X}_i(2)]^2}{S_i^2(1) + S_i^2(2)} \quad (1)$$

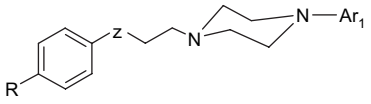
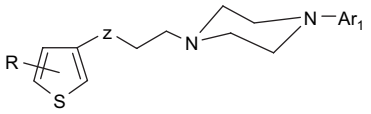
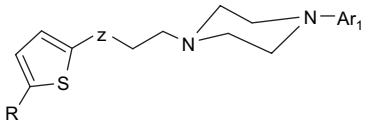
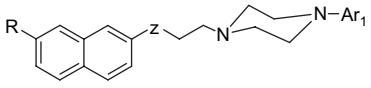
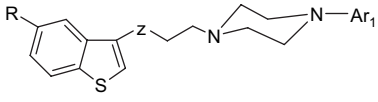
where X_i are the mean values of variables for samples from each class and S_i^2 are the variance values for each class.

The 20 descriptors presenting significant Fisher's weight values, i.e., $W_{1-2} > 1.00$, were selected as those that possess a better ability in the discrimination between the higher and the lower 5-HT_{1A} affinity compounds. After this procedure, we tested different combinations of these descriptors until good HCA and PCA separations were found without any sample being located in an incorrect group. The best HCA and PCA models were obtained with the following variables: AECC, ICR, HVcpx, MATS5v, GATS5e, GGI5, JGI5, JGI8 and H3m. The types and definitions of these descriptors are listed in Table 3. AECC, ICR and HVcpx are topological indices obtained from molecular graphs [24,25]; MATS5v and GATS5e are 2D autocorrelation descriptors, also obtained from molecular graphs, by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag) [26,27]; GGI5, JGI5 and JGI8 are Galvez topological charge indices, which evaluate the charge transfers between pairs of atoms and the global charge transfers in the molecule [28]; and, finally, H3m is a GETAWAY descriptor, calculated from the elements of the leverage matrix obtained by centered atomic coordinates [29,30].

2.4. Chemometric analyses

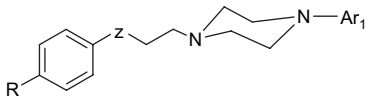
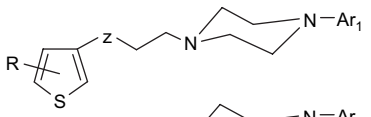
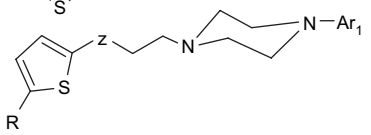
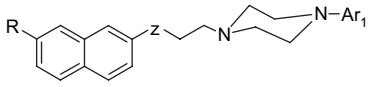
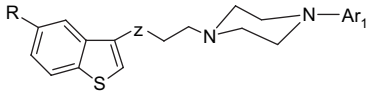
The chemometric methods employed in this work can be classified into three categories: Unsupervised Pattern Recognition (HCA and PCA), Supervised Pattern Recognition (KNN and SIMCA) and Multivariate Calibration (PLS regression).

Table 1
Molecular structures and pK_i values of the training set compounds

Compound	General structure	R	Z	Ar_1	pK_i
1		H	CHOH	2-Methoxyphenyl	7.32
2		H	CHO-4-$CF_3C_6H_4$	2-Methoxyphenyl	6.35
3		H	CNOH	2-Methoxyphenyl	7.76
4		H	CO	4-Chlorophenyl	6.10
5		H	CHO-4-$CH_3C_6H_4$	4-Chlorophenyl	5.84
6		H	CHO-3,4-$OCH_2OC_6H_3$	4-Chlorophenyl	6.26
7		H	CO	4-Methoxyphenyl	5.30
8		H	CHOH	4-Methoxyphenyl	5.30
9		H	CO	2-Chlorophenyl	6.74
10		H	CHOH	2-Chlorophenyl	6.94
11		H	CHO-4-$CF_3C_6H_4$	2-Chlorophenyl	5.30
12		H	CO	4-Fluorophenyl	6.10
13		H	CHO-4-$CF_3C_6H_4$	4-Fluorophenyl	5.30
14		H	CO	2-Pyridyl	7.30
15		H	CHOH	2-Pyridyl	6.81
16		H	CO	4-Nitrophenyl	5.30
17		H	CHOH	4-Nitrophenyl	5.30
18		H	CHO-4-$CF_3C_6H_4$	4-Nitrophenyl	5.30
19		Phenyl	CO	2-Methoxyphenyl	5.44
20		Phenyl	CHO-4-$CF_3C_6H_4$	2-Methoxyphenyl	5.30
21		Methoxy	CO	2-Methoxyphenyl	5.76
22		Methoxy	CHOH	2-Methoxyphenyl	6.49
23		Methoxy	CHO-4-$CF_3C_6H_4$	2-Methoxyphenyl	6.00
24		H	CHOH	2-Methoxyphenyl	7.30
25		H	CHO-4-$CF_3C_6H_4$	2-Methoxyphenyl	6.59
26		H	CNOH	2-Methoxyphenyl	8.19
27		H	CO	4-Chlorophenyl	6.15
28		H	CO	2-Chlorophenyl	6.70
29		H	CHOH	2-Chlorophenyl	6.70
30		H	CO	1-Naphthyl	7.46
31		2,5-Dimethyl	CO	2-Methoxyphenyl	8.30
32		2,5-Dimethyl	CO	2-Hydroxyphenyl	8.12
33		2,5-Dimethyl	CHOH	2-Hydroxyphenyl	7.04
34		2,5-Dimethyl	CO	1-Naphthyl	7.00
35		H	CO	2-Methoxyphenyl	8.00
36		H	CHOH	2-Methoxyphenyl	7.72
37		H	CO	4-Chlorophenyl	5.30
38		H	CHOH	4-Chlorophenyl	5.30
39		5-Methyl	CO	2-Methoxyphenyl	7.76
40		H	CHOH	2-Methoxyphenyl	6.38
41		H	CO	4-Chlorophenyl	5.30
42		H	CHOH	4-Chlorophenyl	5.30
43		H	CO	2-Methoxyphenyl	7.36
44		H	CHOH	2-Methoxyphenyl	7.70
45		H	CO	4-Chlorophenyl	5.30
46		H	CHOH	4-Chlorophenyl	5.30
47		H	CO	2-Hydroxyphenyl	6.96
48		H	CHOH	2-Hydroxyphenyl	7.74
49		H	CO	4-Chloro-2-methoxyphenyl	6.30
50		H	CHOH	4-Chloro-2-methoxyphenyl	6.44
51		H	CHOH	4-Fluoro-2-methoxyphenyl	6.30
52		H	CO	1-Naphthyl	7.00

HCA helped us to define the classes to which the compounds belong and PCA provided an initial knowledge of the basic structure of the data set. KNN and SIMCA, two methods based on the assumption that closer the samples lie in measurement space, the more likely they belong to the same class, were employed to build classification models for the arylpiperazine compounds. The PLS regression was performed to obtain a quantitative structure–activity relationship model of the 5-HT_{1A} receptor affinities presented by the compounds under study. With the exception of HCA, each of these methods was employed here in two steps. First, a model was built and refined based on the training set compounds and then it was used for making predictions of unknown samples (test set compounds).

Table 2
Molecular structures and pK_i values of the test set compounds

Compound	General structure	R	Z	Ar_1	pK_i
53		H	CO	2-Methoxyphenyl	7.30
54		H	CHOH	4-Chlorophenyl	6.10
55		H	CHO-4-$CF_3C_6H_4$	4-Methoxyphenyl	5.30
56		H	CO	2-Pyrimidyl	6.92
57		H	CHO-4-$CF_3C_6H_4$	2-Pyrimidyl	5.80
58		H	CHO-4-$CF_3C_6H_4$	2-Pyridyl	5.80
59		Phenyl	CHOH	2-Methoxyphenyl	6.07
60		H	CO	2-Methoxyphenyl	7.80
61		H	CHOH	4-Chlorophenyl	5.56
62		2,5-Dimethyl	CHOH	2-Methoxyphenyl	7.92
63		5-Methyl	CHOH	2-Methoxyphenyl	7.47
64		5-Nitro	CO	2-Methoxyphenyl	6.47
65		H	CO	2-Methoxyphenyl	6.60
66		H	CO	4-Fluoro-2-methoxyphenyl	6.30

3. Results and discussion

3.1. Unsupervised pattern recognition

The goal of unsupervised pattern recognition analyses is to find realistic densities or clusters of samples in the space determined by the measurements, which reflect the possible existence of meaningful interrelationships. The existence of clustering in a data set is evaluated without using class membership information [31].

Table 3
Selected descriptors and their definitions

Descriptor	Type	Definition
AECC	Topological	Average eccentricity [24]
ICR		Radial centric information index [25]
HVcpx		Graph vertex complexity index [25]
MATS5v	2D autocorrelations	Moran autocorrelation − lag 5/weighted by atomic van der Waals volumes [26]
GATS5e		Geary autocorrelation − lag 5/weighted by atomic Sanderson electronegativities [27]
GGI5	Galvez topological charge index	Topological charge index of order 5 [28]
JGI5		Mean topological charge index of order 5 [28]
JGI8		Mean topological charge index of order 8 [28]
H3m	GETAWAY	H autocorrelation of lag 3/weighted by atomic masses [29,30]

3.1.1. Hierarchical cluster analysis

In the Hierarchical Cluster Analysis (HCA) methodology distances between pairs of samples are calculated and compared. Small distances between samples imply that they are similar. On the other hand, dissimilar samples will be separated by relatively large distances. HCA starts with each sample defined as its own cluster, then groups together similar samples to form new clusters until all samples are part of a single cluster. The main purpose of HCA is to represent data in a manner that emphasizes natural groupings assigning, thus, categories to which samples belong. The visualization of the groups corresponding to different classes is achieved in the form of dendrograms where these classes can be easily identified. Different dendrograms can be obtained according to the techniques used to link the clusters [31].

Fig. 2 shows the dendrogram of the samples obtained with the incremental linkage. The branches on the left of the dendrogram represent single samples. The length of the branches linking two clusters is related to their similarity. The longer the branch, the less the similarity; the shorter the branch, the greater the similarity and, therefore, the smaller the intercluster distance. Similarity is plotted along the top of the graphic with 1.0 corresponding to an exact duplicate and 0.0 indicating maximum distance and dissimilarity [32].

In Fig. 2, our training set compounds appear clustered into two groups: Group 1, characterized by the higher 5-HT_{1A}

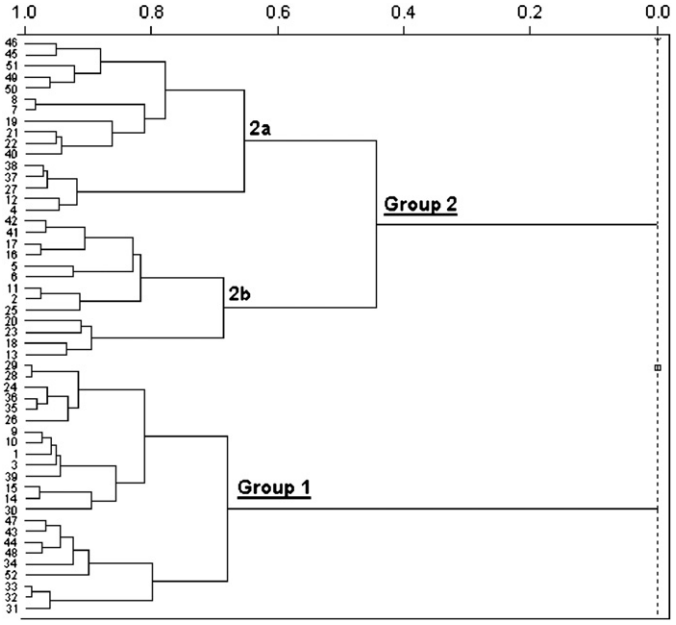


Fig. 2. Dendrogram for the training set compounds obtained with incremental linkage.

affinity compounds (**1**, **3**, **9**, **10**, **14**, **15**, **24**, **26**, **28**–**36**, **39**, **43**, **44**, **47**, **48**, **52**), and Group 2, containing the lower 5-HT_{1A} affinity compounds (**2**, **4**–**8**, **11**–**13**, **16**–**23**, **25**, **27**, **37**, **38**, **40**–**42**, **45**, **46**, **49**–**51**). Group 2 splits into two sub-groups: sub-group 2a, in which Z substituents are CO or CHOH (see structures in Table 1) and sub-group 2b in which most of the compounds present the common feature of sizeable Z substituents. Other common features in these groups can be observed, as follows: in Group 1, almost all compounds present a 2-methoxyphenyl or 2-hydroxyphenyl substituent as Ar_1, while in Group 2 most compounds present 4-chloro or 4-fluorophenyl as Ar_1 substituents. Additionally, most compounds present a thiophene ring with substitutions in different positions or attached to a benzene ring. These findings are in good agreement with previous SAR studies, which indicate that *ortho* substitution in Ar_1 as well as the presence of a thiophene ring in the molecule [8,9,12] is favorable for high 5-HT_{1A} affinities.

Since the compounds are grouped according to their pK_i values (Group 1, $pK_i > 6.70$ and Group 2, $pK_i < 6.70$), the classes of the compounds in all further analyses were attributed following this criterion, thus Class 1 corresponds to Group 1 and Class 2 corresponds to Group 2.

3.1.2. PCA results

Principal Component Analysis (PCA) is a mathematical manipulation of the data matrix where the aim is to represent the variation present in many variables using a small number of principal components (PCs). PCA finds linear combinations of the original independent variables that account for maximum amounts of variation. So, the plotting of samples in the new space formed by the first two or three PCs (placed on the axes instead of original variables) guarantees the display of intersample relationships to be optimal from the point of view of the represented variance. It is important to mention that in multivariate data, none of the original variables describes completely the variation in the data set. However, the first principal component is calculated in such a way that it describes the variation in the data set more than any original variable. Thus, PCA provides the best possible view of variability in the independent variables, which reveals whether there is natural clustering in the data set and whether there are outlier samples. It may also be possible to ascribe chemical (or biological or physical) meaning to the data patterns that emerge from PCA and to estimate which portion of the measurement space is noise. Finally, a PCA model can be constructed which can serve as a benchmark for comparisons with unknown samples [33,34].

Fig. 3 illustrates the scores plot of the first two PCs obtained with the combination of the nine variables cited above

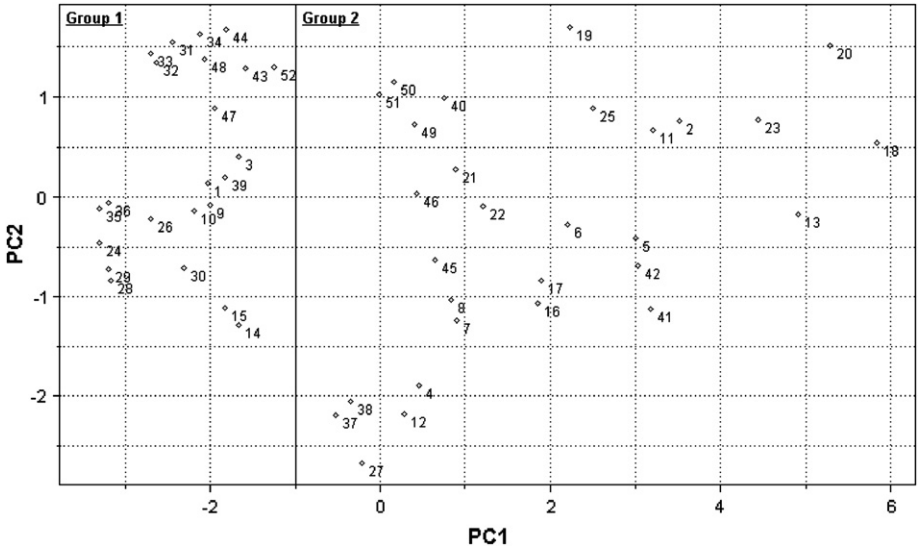


Fig. 3. Scores plot of PC2 against PC1 for the 52 training set compounds.

(see Table 3). Together, these components contain 83% of the total variance from the original data set providing, then, a reliable representation of it. In Fig. 3 it is possible to see that PC1 separates the data set in the same two groups as HCA: Group 1 (higher 5-HT_{1A} affinity compounds) and Group 2 (lower 5-HT_{1A} affinity compounds).

Table 4 shows the loadings of each variable in PC1 and PC2. There we can notice that all variables have similar importance for PC1, with major contributions of GGI5, HVcpx, ICR and AECC. On the other hand, the 2D autocorrelation descriptor GATS5e presents the higher contribution to PC2.

The PCA model obtained with these variables was submitted to two validation procedures: cross-validation and external validation with the test set from Table 2. The leave-one-out cross-validation resulted in 100% of correct information, i.e., no compound was incorrectly allocated. The projection of the test set compounds in the PC space is shown in Fig. 4. The purpose of prediction in PCA is to decide if the unknown sample differs significantly from the training set or not. This decision is based mainly on the magnitude of the residuals when the unknown sample is projected into the model PC space, i.e., samples significantly different will have large residuals. For PCA prediction, a box around the scores on each PC based on the training set values is constructed. For each PC, the lower limit for the scores is:

$$t_{\mathrm{lim(lower)}} = t_{\min} - ct^{*}s_t \quad (2)$$

and the upper limit is:

$$t_{\mathrm{lim(upper)}} = t_{\max} + ct * s_t \quad (3)$$

where $t_{\min}$ and $t_{\max}$ are the minimum and maximum values of the scores, s_t is the scores standard deviation, t^* is a t value with degrees of freedom based on the number of samples in the training set and c is a constant called standard deviation multiplier that can be changed in order to contract or expand the boundaries around $t_{\min}$ and $t_{\max}$ by choosing a negative and positive value, respectively [32]. In Fig. 4, the hyperbox formed by the upper and lower score limits is represented by the rectangle surrounding the samples. The five test set compounds presenting higher affinities are in the same region of the corresponding Group 1 of the training set, as well as the lower 5-HT_{1A} affinity test compounds are in the Group 2 region. Thus the external validation, as well as cross-validation, indicates that our PCA model presents good predictive ability. As the test set compounds are shown to be similar to the training set ones, they have proved to be suitable for being used in the external validation of all models obtained in this work.

Table 4
Variable loadings in each PC

	PC1	PC2
AECC	0.36	0.30
ICR	0.37	0.04
HVcpx	0.37	0.27
MATS5v	0.25	−0.60
GATS5e	−0.24	0.59
GGI5	0.37	0.21
JGI5	0.29	−0.26
JGI8	0.35	0.08
H3m	0.36	0.10

3.2. Supervised pattern recognition

In supervised pattern recognition methods, the training set samples in the measurement space are previously ''tagged'' with a known classification. The primary objective is to develop a rule which classifies these samples correctly and then apply the same rule for classification of unknown samples [31].

3.2.1. KNN results

In the KNN method, the Euclidean distance separating each pair of samples in the training set is calculated and stored in a distance table. The class to which most of the nearest neighbors belong is assigned for any particular sample. Thus, the classification model obtained can be used in class prediction of unknown samples [33].

The descriptors selected with the help of HCA and PCA were employed in our KNN analysis. When considering until nine nearest neighbors, all compounds from the training set were correctly classified, which shows that the classes are quite distinct and the selected variables have good discriminating ability.

An external validation of the KNN classification was achieved for this model. As mentioned before, the criterion used to split the training set into two classes of compounds was also employed to assign the classes for the test set compounds, i.e., compounds with $pK_i > 6.70$ belong to Class 1 and compounds with $pK_i < 6.70$ belong to Class 2. No prediction errors were found when computing the distances to 1, 3, 5, 7 and 9 neighbors. The five higher 5-HT_{1A} affinity compounds from the test set were assigned to Class 1 and the nine lower 5-HT_{1A} affinity compounds were located in Class 2. This indicates that a reliable and predictive model is obtained.

3.2.2. SIMCA results

SIMCA develops principal component models for each class of the training set. Then, when the values of the independent variables of a new sample are projected into the PC space of each class, the new sample is assigned to the class it best fits. Besides building a classification model to be used for predictions of unknown samples, SIMCA also provides diagnostic tools related to other interesting aspects of classification such as discriminating and modeling power of variables, class distances and outlier detection. Moreover, the variance structure of each class yields clues about category complexity and may even reveal the phenomena causing one category to differ from another. An additional attractive feature of SIMCA is its realistic prediction options compared to KNN. The later assigns every sample to exactly one training set class (the class of the nearest neighbors) while SIMCA is able to identify if the sample does not belong to any class or if it can be member of both classes since the outcomes are expressed probabilistically [33].

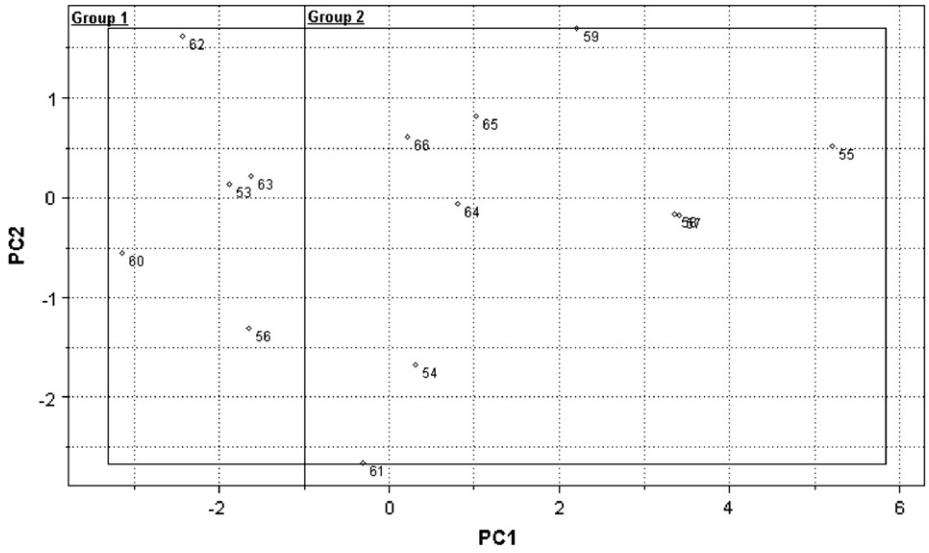


Fig. 4. Scores plot for PCA prediction performed for the 14 test set compounds.

The best SIMCA model was built with the same descriptors as HCA, PCA and KNN. Fig. 5 presents the class distances calculated according to the residuals of the samples when they are adjusted to the classes. This plot is divided by two lines that represent critical residual variances. Compounds lying in the northwest quadrant (NW) belong only to the x-axis class, because they are at distances small enough to be considered members of this class. Similarly, compounds in the southeast quadrant (SE) are members of the y-axis class only. Compounds in the southwest quadrant (SW) may belong to both classes, while the ones in the northeast quadrant (NE) belong to none. From Fig. 5 we can see the 29 lower 5-HT_{1A} affinity compounds in the SE quadrant, which means that they are assigned to Class 2 only. On the other hand, there are 12 higher 5-HT_{1A} affinity compounds lying in the NW quadrant (Class 1 only), while SIMCA predicts that 11 compounds might belong not only to Class 1 but also to Class 2 (since they are positioned in the SW quadrant).

This model was also employed for making predictions on the test set compounds. For this set, no compound was predicted as not belonging to one of the classes or to both. As the KNN model presented before, SIMCA resulted in a predictive model.

Other data trends can be analyzed from the SIMCA classification. The interclass distance (a measure of class separation) was computed as 6.22, showing that the classes are quite distinct in our SIMCA model (as a rule of thumb, classes are considered separable when the interclass distance is greater than 3 [34]). Considering the measures of variable importance, the most relevant descriptor for separation of the classes is GATS5e, which have the higher discriminating power. The descriptor presenting the highest modeling power is AECC,

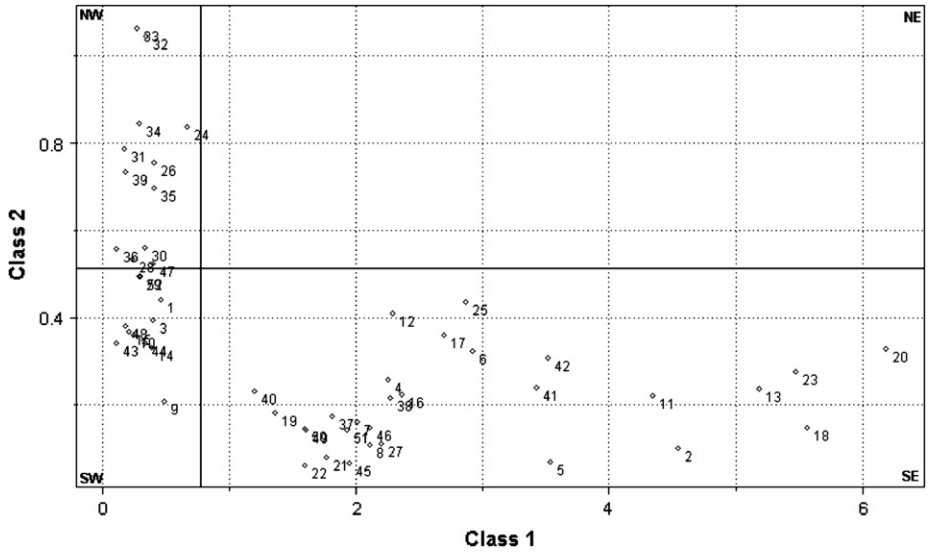


Fig. 5. Class distances obtained for the 52 training set compounds.

what means that this is the variable that better describes information for the two classes of the training set.

3.3. PLS regression

The chemometric regression method PLS was employed in this work to model the dependent variable (pK_i values), Y, using a set of independent variables (X_i), representing the chemical structures of the investigated compounds. In the same manner as PCA, the PLS method finds linear combinations of the original independent variables that account for maximum amounts of variation. But in PLS, the loading matrix is defined in such a way that not only the variance is maximized but also the product of the variance times the correlation to Y is optimized. Thus, in the matrix form we have the regression equation given by:

$$Y = X\beta + F \quad (4)$$

where β is the regression vector and F represents the residual errors in estimating Y.

The best model is chosen on the basis of the cross-validation predicted residual error sum of squares (PRESS). The optimal number of PLS components is the one that minimizes PRESS. If the model presents good quality, which is verified mainly by the correlation coefficients r^2 and q^2 and also by the prediction residuals, it can be then used for making predictions of the biological properties of unknown compounds structurally similar to the ones from the training set [35].

The best PLS model found in this work was built with the same descriptors as the pattern recognition models presented before. The optimal number of PLS components was chosen by evaluating PRESS, for the residuals of prediction from the cross-validation of the model, leaving one sample out (see Table 5). The optimal estimated number of PLS components was six, which is the one corresponding to the lowest PRESS and highest q^2. This resulted in a model with $r^2 = 0.83$ and $q^2 = 0.76$, which indicates that a predictive model is obtained. Fig. 6 shows the regression plot for the predicted pK_i values against the experimental ones for the training set compounds. Most of the prediction residuals are smaller than 0.60, with only two compounds presenting high residuals (0.92 and 0.99, for compounds **19** and **33**, respectively).

Fig. 6 also shows the plot of test set compounds used to perform an external validation of the PLS model. The results of predictions for the test set are shown in Table 6. There we can see that the prediction errors are not higher than 0.34 (except compound **64**), indicating that a model with good predictive power is obtained.

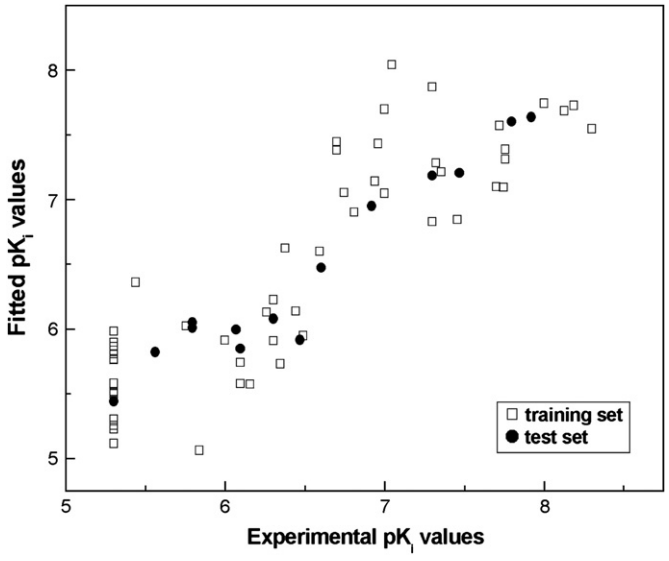


Fig. 6. Plot of fitted *versus* experimental pK_i values for the PLS regression obtained with six PLS components, showing the training and test set compounds.

When we compare the results obtained with our chemometric analyses, we can notice that the qualitative and quantitative structure–activity relationship models presenting reliability and predictive power were obtained with the same set of variables, which is a strong indication that the descriptors selected in our work are important for 5-HT_{1A} receptor affinity and also that our correlations are not obtained by chance. The models are able to separate the compounds according to their pK_i values and the following requirements to receptor affinity become evident in the two groups of compounds (Class 1 and Class 2): *ortho* substitution in Ar_1 ring and a thiophene ring instead of phenyl or naphthyl rings. This is in agreement with previous SAR studies with arylpiperazine compounds [8,9,12]. Additionally, the principal components (in PCA and SIMCA analyses) and the latent variables (PLS components) built with these

Table 5
Variance percentage, PRESS, q^2 and r^2

PLS component	% Cumulative variance	PRESS	q^2	r^2
1	66	17.757	0.621	0.654
2	81	12.645	0.730	0.776
3	87	11.911	0.746	0.811
4	92	11.597	0.754	0.825
5	95	11.342	0.759	0.831
6	98	11.305	0.760	0.832
7	100	11.539	0.755	0.833
8	100	12.091	0.744	0.834
9	100	12.185	0.743	0.835

Table 6
Prediction residuals for the 14 test set compounds

Compound	Experimental pK_i	Predicted pK_i	Residual
53	7.30	7.22	0.08
54	6.10	5.79	0.30
55	5.30	5.54	−0.24
56	6.92	7.03	−0.11
57	5.80	6.14	−0.34
58	5.80	6.09	−0.29
59	6.07	6.05	0.02
60	7.80	7.69	0.11
61	5.56	5.80	−0.24
62	7.92	7.59	0.33
63	7.47	7.20	0.27
64	6.47	5.78	0.69
65	6.60	6.51	0.10
66	6.30	6.06	0.24

descriptors account for significant percentages of the whole variance in the data matrix, showing that the data structure was well captured by the variables used in the models. Through all these considerations, we feel very confident in assuring that the PLS model obtained here, as well as the pattern recognition models, can be helpful in the design of new arylpiperazine compounds similar to our training set compounds.

4. Conclusions

In this work, reliable and predictive SAR and QSAR models were obtained by means of chemometric approaches using a set of fast and easily calculated descriptors. All models have presented internal consistency and are externally validated with a test set of compounds. The selected descriptors, which encode useful information about several aspects of molecular structure, have shown to be closely related to 5-HT_{1A} receptor affinities presented by the compounds under study, since they were successfully employed in all chemometric analyses. Sterical properties are well represented by the topological indices. The selection of Galvez charge indices indicates the importance of electronic features of the compounds that may be involved in ligand binding to the receptor. Moreover, the characteristics of the compounds in each studied group (Class 1 and Class 2) are in good agreement with previous SAR analysis on arylpiperazine compounds, and confirm the accepted requirements for 5-HT_{1A} receptor recognition.

Acknowledgements

The authors would like to thank the Brazilian agency FAPESP for the financial support.

References

[1] J. Annanth, Psychother. Psychosom. 67 (1998) 61–70.
[2] D.S. Kreiss, I. Lucki, J. Pharmacol. Exp. Ther. 274 (1995) 866–876.
[3] P.R. Albert, S. Lemonde, Neuroscientist 10 (2004) 575–593.
[4] S.E. Gartside, V. Umbers, M. Hajos, T. Sharp, Br. J. Pharmacol. 115 (1995) 1064–1070.
[5] L.J. Dreshfield, D.T. Wong, K.W. Perry, E.A. Engleman, Neurochem. Res. 21 (1996) 557–562.
[6] T. Sharp, V. Umbers, S.E. Gartside, Br. J. Pharmacol. 121 (1997) 941–946.
[7] V. Perez, J. Soler, D. Puigdemont, E. Alvarez, F. Artigas, Arch. Gen. Psychiatry 56 (1999) 375–379.
[8] M.L. López-Rodríguez, M.L. Rosado, B. Benhamu, M.J. Morcillo, E. Fernández, K.-J. Schaper, J. Med. Chem 40 (1997) 1648–1656.
[9] M.L. López-Rodríguez, M.J. Morcillo, E. Fernández, M.L. Rosado, L. Pardo, K.-J. Schaper, J. Med. Chem. 44 (2001) 198–207.
[10] P. Gaillard, P.-A. Carrupt, B. Testa, P. Schambel, J. Med. Chem. 39 (1996) 126–134.
[11] M.C. Menziani, P.G. de Benedetti, M. Karelson, Bioorg. Med. Chem. 6 (1998) 535.
[12] J. Martínez-Esparza, A.M. Oficialdegui, S. Pérez-Silanes, B. Heras, L. Orús, J.A. Palop, B. Lasheras, J. Roca, M. Mourelle, A. Bosch, J.C. Del Castillo, T.R. Ordera, J. Del Rio, A. Monge, J. Med. Chem. 44 (2001) 418–428.
[13] J. Souza Jr., R.H.A. Santos, M.M.C. Ferreira, F.A. Molfetta, A.J. Camargo, K.M. Honório, A.B.F. da Silva, Eur. J. Med. Chem. 38 (2003) 929–938.
[14] S.L. da Silva, S. Marangoni, K.C. Weber, P. Homem-de-Mello, K.M. Honório, A.B.F. da Silva, Internet Electron. J. Mol. Des. 4 (2005) 515–526.<http://www.biochempress.com>.
[15] K.C. Weber, K.M. Honório, A.T. Bruni, A.B.F. da Silva, J. Mol. Model. 12 (2006) 915–920.
[16] K.C. Weber, K.M. Honório, A.T. Bruni, A.B.F. da Silva, Struct. Chem. 17 (2006) 307–313.
[17] N.L. Allinger, Y.H. Yuh, J.H. Lin, J. Am. Chem. Soc. 111 (1989) 8551–8566.
[18] N.S. Ostlund, HyperChem: Program for Molecular Visualization and Simulation, University of Waterloo, Canada, 1995.
[19] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, J. Am. Chem. Soc. 13 (1985) 3902–3909.
[20] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery Jr., T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, J.A. Pople, Gaussian 03, Revision C.02, Gaussian, Inc., Wallingford, CT, 2004.
[21] R. Todeschini, V. Consonni, M. Pavan, Dragon 2.1, Milan, 2002.
[22] Pirouette 3.11, Infometrix Inc, Woodinville, 2002.
[23] R.E. Bruns, J.F.G. Faigle, Quim. Nova (1985) 84–99.
[24] E.V. Kostantinova, J. Chem. Inf. Comput. Sci. 36 (1997) 54–57.
[25] C. Raychaudhury, S.K. Ray, J.J. Gosh, A.B. Roy, S.C. Basak, J. Comput. Chem. 5 (1984) 581–588.
[26] P.A.P. Moran, Biometrika 37 (1950) 17–23.
[27] R.C. Geary, Incorp. Statist. 5 (1954) 115–145.
[28] J. Galvez, R. Garcia, M.T. Salabert, R. Soler, J. Chem. Inf. Comput. Sci. 34 (1994) 520–525.
[29] V. Consonni, R. Todeschini, M. Pavan, J. Chem. Inf. Comput. Sci. 42 (2002) 682–692.
[30] V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, J. Chem. Inf. Comput. Sci. 42 (2002) 693–705.
[31] M.A. Sharaf, D.L. Illman, B.R. Kowalski, Chemometrics, Wiley & Sons, New York, 1986, 357p.
[32] Pirouette: User's Guide, Infometrix Inc., Woodinville, 2002, 430p.
[33] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, John Wiley & Sons, New York, 1998, 459p.
[34] R.G. Brereton, Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies, Elsevier, Amsterdam, 1992.
[35] S. Wold, M. Sjostrom, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.