MIA-QSAR coupled to principal component analysis-adaptive neuro-fuzzy inference systems (PCA-ANFIS) for the modeling of the anti-HIV reverse transcriptase activities of TIBO deriva...

ARTICLE *in* EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY · APRIL 2010

Impact Factor: 3.45 · DOI: 10.1016/j.ejmech.2009.12.028 · Source: PubMed

Mohammad Goodarzi
Vrije Universiteit Brussel
107 PUBLICATIONS **922** CITATIONS

SEE PROFILE

Matheus Freitas
Universidade Federal de Lavras (UFLA)
109 PUBLICATIONS **1,111** CITATIONS

SEE PROFILE

Vol. 45 • No 4 • April 2010 • ISSN 0223-5234
ELSEVIER
EUROPEAN JOURNAL OF
MEDICINAL
CHEMISTRY
Available online at www.sciencedirect.com
ScienceDirect

European Journal of Medicinal Chemistry 45 (2010) 1352–1358

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

MIA–QSAR coupled to principal component analysis-adaptive neuro-fuzzy inference systems (PCA–ANFIS) for the modeling of the anti-HIV reverse transcriptase activities of TIBO derivatives

Mohammad Goodarzi [a], Matheus P. Freitas [b,*]

[a] Department of Chemistry, Faculty of Sciences, and Young Researchers Club – Islamic Azad University, Arak Branch, Arak, Markazi, Iran
[b] Departamento de Química, Universidade Federal de Lavras – UFLA, Caixa Postal 3037, 37200-000, Lavras, MG, Brazil

ARTICLE INFO

Article history:
Received 24 March 2009
Received in revised form
9 December 2009
Accepted 14 December 2009
Available online 4 January 2010

Keywords:
TIBO derivatives
Anti-HIV reverse transcriptase activities
MIA–QSAR
PCA–ANFIS

ABSTRACT

The activities of a series of HIV reverse transcriptase inhibitor TIBO derivatives were recently modeled by using genetic function approximation (GFA) and artificial neural networks (ANN) on topological, structural, electronic, spatial and physicochemical descriptors. The prediction results were found to be superior to those previously established. In the present work, the multivariate image analysis applied to quantitative structure–activity relationship (MIA–QSAR) method coupled to principal component analysis-adaptive neuro-fuzzy inference systems (PCA–ANFIS), which accounts for non-linearities, was applied on the same set of compounds previously reported. Additionally, partial least squares (PLS) and multilinear partial least squares (N-PLS) regressions were used for comparison with the MIA–QSAR/PCA–ANFIS model. The ANFIS procedure was capable of accurately correlating the inputs (PCA scores) with the bioactivities. The predictive performance of the MIA–QSAR/PCA–ANFIS model was significantly better than the MIA–QSAR/PLS and N-PLS models, as well as than the reported models based on CoMFA, CoMSIA, OCWLGI and classical descriptors, suggesting that the present methodology may be useful to solve other QSAR problems, specially those involving non-linearities.

1. Introduction

The acquired immuno deficiency syndrome (AIDS) epidemic has claimed about two million lives in 2007, and nearly 33 million people globally live with the virus [1]. A class of anti-HIV drugs is the HIV reverse transcriptase inhibitors. HIV records RNA into DNA using a key enzyme, reverse transcriptase. Blocking this step has been used to prevent the virus replication [2,3]. Tetrahydroimidazo[4,5,1-*jk*][1,4]benzodiazepine (TIBO) derivatives have been successfully used as HIV reverse transcriptase inhibitors [4], and are useful models for deriving novel potent compounds. The first step of rational drug design often consists in performing QSAR modeling based on congeneric structures of ligands; this procedure has been carried out by several researchers for a class of TIBO derivatives [4–9].

Multivariate image analysis (MIA) applied to QSAR has provided predictive models for several compound classes [10–21], and was proved to be a valuable tool in proposing new active entities. In the MIA–QSAR method, descriptors are pixels (binaries) of 2D images – chemical structures. In addition to the widely used PLS regression, some mathematical approaches have been coupled to MIA–QSAR in order to achieve progressively more predictive models, such as N-way and non-linear methods [18–22]. In this work, principal component analysis-adaptive neuro-fuzzy inference systems (PCA–ANFIS) technique was applied together with MIA–QSAR to verify the enhancement in predictive ability of the bioactivities of TIBO derivatives when compared to previous works [4–9] and to MIA–QSAR models based on well established regression methods.

The application of chemometrics, particularly principal component analysis (PCA) to multivariate chemical data is becoming widespread, especially owing to the availability of digitized spectroscopic data and commercial software for laboratory computers. PCA [23] creates *p* latent variables (*Y*) as linear combinations of the original *p* variables (*X*), in such a way that new orthogonal axes are built to explain the maximum variance possible in just a few dimensions.

$$Y_i = Xe_i \qquad (1)$$

where for the data matrix *X* of dimensionality $(m \times n)$, *e* denotes the *i*th loading vector of dimensionality $(1 \times n)$ and y_i represents the *i*th score vector of dimensionality $(m \times 1)$.

* Corresponding author. Tel.: +55 35 3829 1891; fax: +55 35 3829 1271.
E-mail address: matheus@ufla.br (M.P. Freitas).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.12.028

The purpose of a neuro-fuzzy system is to apply neural learning techniques to identify the parameters and/or structure of neuro-fuzzy systems. These neuro-fuzzy systems can combine the benefits of the two powerful paradigms (neural networks and fuzzy systems) into a single capsule. They have several features, which make them suitable for a wide range of scientific applications. These strengths include fast and accurate learning, good generalization capabilities, excellent explanation facilities in the form of meaningful fuzzy rules, and the ability to accommodate both data and existing expert knowledge about the problem under consideration. In this work, the goal of ANFIS was to find a model or mapping that will correctly associate the inputs (PCA scores) with the target (bioactivities). Fuzzy inference system (FIS) [24] is a knowledge representation where each fuzzy rule describes a local behavior of the system. The network structure that implements FIS is referred to as ANFIS and employs hybrid learning rules to train a Sugeno-style FIS [25] with linear rule outputs.

2. Computational methods

2.1. MIA–QSAR modeling

The MIA–QSAR procedure has been fully detailed elsewhere [22], thus only a brief description is given here for the series of TIBO derivatives, obtained from the literature [9]. The structures of compounds **1**–**70** (Table 1) were systematically built by using appropriate software, ACD/ChemSketch [26], and then converted to bitmaps in 288 × 316 pixels windows, with resolution of 102 × 102 points per inch. All the molecular structures were fixed by a common point among them in a given coordinate (Fig. 1), since the shapes should be superimposed afterwards, as a 2D alignment to allow maximum similarity. In our dataset, the pixel located at the 130 × 110 coordinate (common to the whole series), was used as reference in the alignment step. Each 2D image was read and converted into binaries (double array in Matlab [27]), and the predictors block was built by grouping the 70 treated images, giving a 70 × 288 × 316 array. The 3D array was unfolded to a 2-way array (70 × 91,008), in order to obtain PCA scores for further analysis with ANFIS. PCA was performed for the training set, and the scores of validation and test sets were predicted by this PCA model. Also, the X-matrix was used for regression against the bioactivity values through partial least squares (PLS) and multilinear partial least squares (N-PLS).

2.2. ANFIS model

2.2.1. ANFIS architecture

The proposed neuro-fuzzy model in ANFIS is a multilayer neural network-based fuzzy system [28]. Its topology is shown in Fig. 2a and b, and the system has a total of five layers. In this connectionist structure, the input and output nodes represent the descriptors and the response, respectively, and in the hidden layers, there are nodes functioning as membership functions (MFs) and rules. For simplicity, we assume that the examined fuzzy inference system has two inputs, x and y, and one output. For a Sugeno fuzzy model, a common rule set with two fuzzy if-then rules is as follows:

Rule 1: If x is A_1 and y is B_1, then $f_1 = p_1x + q_1y + r_1$
Rule 2: If x is A_2 and y is B_2, then $f_2 = p_2x + q_2y + r_2$

Fig. 2b illustrates the reasoning mechanism for the Sugeno model and the corresponding equivalent ANFIS architecture, where nodes of the same layer have similar functions (the output of the ith node in layer l is denoted as $O_{l,i}$).

Layer 0: The input layer. It has n nodes where n is the number of inputs to the system.

Layer 1: It is the fuzzification layer in which each node i in this layer is an adaptive node output defined by

$$O_{1,i} = \mu_{A_i}(x) \text{ for } i = 1,2 \text{ or } O_{1,i} = \mu_{B_{i-2}}(y) \text{ for } i = 3,4 \qquad (2)$$

Where x (or y) is the input to node, and A_i (or B_{i-2}) is a fuzzy set associated with this node. In other word, outputs of this layer are the membership values of the premise part. Here the membership functions for A_i and B_i can be any appropriate parameterized membership function that usually we choose $\mu Ai\ (x)$ to be bell-shaped with maximum equal to 1 and minimum equal to 0, such as the generalized bell function

$$\mu_{A_i}(x) = \frac{1}{1+\left[\left(\frac{x-c_i}{a_i}\right)^2\right]^{b_i}} \qquad (3)$$

or the Gaussian function

$$\mu_{A_i}(x) = \exp\left\{-\left[\left(\frac{x-c_i}{a_i}\right)^2\right]^{b_i}\right\} \qquad (4)$$

As the values of the parameters $\{a_i, b_i, c_i\}$ change, the bellshaped functions vary accordingly, thus exhibiting various forms of membership functions on linguistic label A_i. Parameters in this layer are referred to as premise parameters.

In this layer, there exist $n \times p$ nodes where n is the number of input variables and p is the number of membership functions. For example, if size is an input variable and there exists two linguistic values for size, which are SMALL and LARGE, then two nodes are kept in the first layer and they denote the membership values of input variable size to the linguistic values SMALL and LARGE.

Layer 2: Every node in this layer is a fixed node labeled II (Fig. 2b), whose output is the product of all the incoming signals:

$$O_{2,i} = w_i = \mu_{A_i}(x) \times \mu_{B_i}(y) \quad \text{for } i = 1,2 \qquad (5)$$

Each node output represents the firing strength of a rule.

Layer 3: Every node in this layer is a fixed node labeled N. The ith node calculates the ratio of the ith rule's firing strength to the sum of all rules' firing strengths:

$$O_{3,i} = \overline{w}_i = \frac{w_i}{w_1 + w_2} \quad i = 1,2 \qquad (6)$$

For convenience, outputs of this layer are called normalized firing strengths.

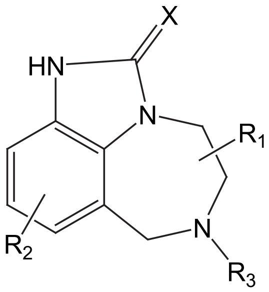
Layer 4: Every node i in this layer is an adaptive node with a node function

$$O_{4,i} = \overline{w}_i f_i = \overline{w}_i(p_i x + q_i y + r_i) \qquad (7)$$

where w_i is a normalized firing strength from layer 3 and $\{p_i, q_i, r_i\}$ is the parameter set of this node. Parameters in this layer are referred to as consequent parameters.

Layer 5: The single node in this layer is a fixed node labeled Σ, which computes the overall output as the summation of all incoming signals:

Table 1
Compounds used in the MIA–QSAR analyses, and experimental, fitted and predicted pIC$_{50}$ (IC$_{50}$ in mol L^{-1}).[a]



No.	R$_1$	X	R$_2$	R$_3$[b]	Exp.	PLS	N-PLS	RBFNN	ANFIS
1	H	S	8-Cl	DMA	7.34	7.70	7.78	7.33	7.24
2	H	S	9-Cl	DMA	6.79	6.75	5.85	6.99	6.40
3	5-Et	O	H	2-MA	4.30	4.60	4.46	4.90	4.67
4*	5-*i*-Pr	O	H	2-MA	5.00	5.63	5.18	5.59	4.84
5**	5-*i*-Pr	O	H	DMA	5.00	4.81	4.86	5.29	5.11
6	5,5-Di-Me	O	H	2-MA	4.64	4.76	4.94	4.65	5.19
7*	4-Me	O	H	2-MA	4.49	4.41	3.79	4.27	4.62
8*	4-Me	S	9-Cl	2-MA	6.17	5.63	5.95	6.36	5.99
9*	4-Me	S	9-Cl	CH$_2$CH(CH$_2$)$_2$	5.66	6.05	6.35	5.94	5.85
10*	4-*i*-Pr	O	H	*n*-Pr	4.13	4.14	3.76	4.14	4.26
11	4-*i*-Pr	O	H	2-MA	4.90	5.16	4.66	5.28	4.92
12	4-*n*-Pr	O	H	*n*-Pr	3.74	3.09	3.57	4.48	3.69
13*	4-*n*-Pr	O	H	2-MA	4.32	4.76	4.38	4.39	3.98
14	7-Me	O	H	*n*-Pr	4.08	4.32	4.29	5.00	4.05
15**	7-Me	O	H	DMA	4.92	5.16	5.17	5.27	5.43
16	7-Me	O	8-Cl	DMA	6.84	6.82	6.54	6.47	6.47
17*	7-Me	O	9-Cl	DMA	6.79	6.01	5.97	6.57	6.89
18**	7-Me	S	H	*n*-Pr	5.61	4.93	5.12	5.72	5.50
19*	7-Me	S	H	DMA	7.11	6.89	6.39	6.92	6.91
20	7-Me	S	8-Cl	DMA	7.92	7.86	7.76	7.38	7.44
21*	7-Me	S	9-Cl	DMA	7.64	7.07	7.27	7.50	7.75
22	4,5-Di-Me (*cis*)	O	H	DMA	4.25	4.58	4.36	4.82	4.79
23	4,5-Di-Me (*cis*)	S	H	DMA	5.65	5.43	5.30	5.81	5.66
24*	4,5-Di-Me (*trans*)	S	H	CH$_2$CH(CH$_2$)$_2$	4.87	4.12	5.20	4.63	4.98
25	4,5-Di-Me (*trans*)	S	H	DMA	4.84	4.92	5.80	4.83	4.61
26	4-Keto-5-Me	S	9-Cl	*n*-Pr	4.30	4.91	4.91	4.51	4.48
27	4,5-Di-benzo	S	H	CH$_2$CH(CH$_2$)$_2$	5.00	4.43	5.30	4.98	4.64
28*	5,7-Di-Me (*trans*)	S	H	DMA	7.38	7.36	6.33	6.81	7.05
29*	5,7-Di-Me (*cis*)	S	H	DMA	5.94	5.92	5.88	5.55	6.53
30	5,7-Di-Me (*R,R*; *trans*)	O	9-Cl	DMA	6.64	6.09	5.91	6.38	6.77
31**	5,7-Di-Me (*R,R*; *trans*)	S	9-Cl	DMA	6.32	7.15	7.22	6.24	6.51
32	5,7-Di-Me (*S,S*; *trans*)	O	9-Cl	DMA	5.30	5.60	5.86	5.50	5.53
33	4,7-Di-Me (*trans*)	S	H	DMA	4.59	4.64	5.10	5.40	4.95
34	5,6-CH$_2$C(=CHCH$_3$)CH$_2$ (*S*)	S	9-Cl	–	5.42	5.29	4.91	5.45	5.34
35**	6,7-(CH$_2$)$_4$	S	9-Cl	–	5.70	5.94	5.31	6.20	5.87
36*	5-Me (*S*)	S	8-Cl	DMA	8.30	8.21	7.81	7.85	7.53
37**	5-Me (*S*)	O	9-Cl	DMA	6.74	6.34	6.04	5.97	7.02
38**	5-Me (*S*)	S	9-Cl	DMA	7.37	7.40	7.34	7.32	7.07
39	5-Me (*S*)	S	9-Cl	CH$_2$CH(CH$_2$)$_2$	7.47	7.49	7.41	7.06	7.14
40	5-Me (*S*)	S	H	CH$_2$CH(CH$_2$)$_2$	7.22	7.28	7.50	7.41	7.15
41*	5-Me	O	H	*n*-Pr	4.22	4.69	4.84	4.93	4.19
42	5-Me	S	H	*n*-Pr	5.78	5.79	5.86	6.49	5.85
43	5-Me	O	H	2-MA	4.46	4.41	4.24	5.50	4.34
44	5-Me	S	H	DMA	7.01	6.54	6.43	6.82	6.99
45	5-Me (*S*)	O	H	DMA	5.48	5.50	5.21	6.17	5.52
46	5-Me (*S*)	S	H	2-MA	7.58	7.27	6.97	7.44	7.52
47	H	O	H	DMA	4.90	4.91	5.21	5.19	5.14
48	H	O	H	2-MA	4.33	4.55	4.48	4.37	4.64
49**	H	O	H	*n*-Pr	4.05	4.12	4.40	4.33	4.25
50**	H	O	H	2-EA	4.43	4.57	4.44	4.53	4.64
51	5-Me (*S*)	S	H	DMA	7.36	6.54	6.43	7.33	7.20
52	5-Me (*S*)	O	H	Allyl	4.15	4.35	4.69	4.16	4.00
53*	5-Me (*S*)	O	H	*n*-Bu	4.00	5.03	4.85	4.87	3.97
54	5-Me (*S*)	S	8-F	DMA	8.24	8.01	8.36	7.74	7.99
55	5-Me (*S*)	O	8-Br	DMA	7.32	7.50	7.57	6.91	7.23
56	5-Me (*S*)	S	8-Br	DMA	8.52	8.32	7.80	7.76	8.34
57	5-Me (*S*)	S	8-Me	DMA	7.87	7.69	8.29	7.66	7.73
58**	5-Me (*S*)	S	8-OMe	DMA	7.47	7.74	7.69	7.09	7.31
59	5-Me (*S*)	S	9,10-Di-Cl	DMA	7.59	7.36	7.28	7.27	7.41

Table 1 (*continued*)

No.	R_1	X	R_2	R_3[b]	Exp.	PLS	N-PLS	RBFNN	ANFIS
60	5-Me (*S*)	O	8-CN	DMA	5.94	5.96	6.11	6.17	6.06
61*	5-Me (*S*)	S	8-CN	DMA	7.25	6.99	7.32	6.51	6.53
62	5-Me (*S*)	O	8-Me	DMA	6.00	6.39	6.31	5.97	6.55
63	5-Me (*S*)	S	10-OMe	DMA	5.33	5.71	5.53	5.86	5.56
64*	5-Me (*S*)	O	10-OMe	DMA	5.18	4.65	4.16	5.63	5.23
65**	5-Me (*S*)	S	10-Br	DMA	5.97	6.14	6.20	6.09	6.19
66	5-Me (*S*)	S	8-CHO	DMA	6.73	6.96	7.33	6.45	6.52
67**	5-Me (*S*)	O	8-I	DMA	7.06	6.66	6.32	6.54	6.56
68	5-Me (*S*)	S	8-I	DMA	7.32	7.69	7.54	7.40	6.53
69*	5-Me (*S*)	O	8-C≡CH	DMA	6.36	6.16	6.14	5.94	6.56
70	5-Me (*S*)	S	8-C≡CH	DMA	7.53	7.20	7.36	7.21	7.23

[a] Compounds marked with an asterisk pertain to the test set; compounds marked with two asterisks were used as a validation set in ANFIS.
[b] DMA = 3,3-Dimethylallyl; 2-MA = 2-methylallyl; 2-EA = 2-ethylallyl.

$$\text{overall output} = O_{5,i} = \sum \overline{w}_i f_i = \frac{\sum\limits_i \overline{w}_i f_i}{\sum\limits_i w_i} \tag{8}$$

Thus, we have constructed an ANFIS system that is functionally equivalent to Sugeno fuzzy model, which will be used in the present MIA–QSAR study.

2.2.2. *Hybrid learning algorithm*

From the proposed ANFIS architecture (Fig. 2b), it is observed that given the values of premise parameters, the overall output can be expressed as linear combinations of the consequent parameters. More precisely, the output *f* in Fig. 2b can be rewritten as;

$$\begin{aligned} f &= \frac{w_1}{w_1+w_2} f_1 + \frac{w_2}{w_1+w_2} f_2 \\ &= \overline{w}_1(p_1x+q_1y+r_1) + \overline{w}_2(p_2x+q_2y+r_2) \\ &= (\overline{w}_1x)p_1 + (\overline{w}_1y)q_1 + \overline{w}_1r_1 + (\overline{w}_2x)p_2 + (\overline{w}_2y)q_2 + \overline{w}_2r_2 \end{aligned} \tag{9}$$

which is linear in the consequent parameters p_1, q_1, r_1, p_2, q_2, and r_2. To train the above ANFIS system, the following error measure will be used:

$$E = \sum_{K=1}^{n} \left(f_K - \widehat{f}_K\right)^2 \tag{10}$$

where f_K and $\widehat{f}_K$ are the *k*th desired and estimated outputs and *n* is the total number of data in the training data set. The learning algorithms of ANFIS consist of the following two parts: (a) the learning of the premise parameters by back propagation and (b) the learning of the consequence parameters by least squares estimation. More specifically, in the forward pass of the hybrid learning algorithm, functional signals go forward till layer 4 and the consequent parameters are identified by the least squares estimate. In the backward pass, the error rates propagate backward, and the premise parameters are updated by the gradient descent. During the learning process, the parameters associated with the membership functions will change. The computation of these parameters (or their adjustment) is facilitated by a gradient vector, which provides a measure of how well the fuzzy inference system is modeling the input/output data for a given set of parameters [29].

2.2.3. *Building the ANFIS model*

In this step, the data set is divided into three subsets: the training, validation and test subsets. The training and validation sets were used for the construction of the ANFIS model, and then the generated model was applied to the test set. We have used the validation set for stopping training in order to prevent overtraining problem on the model; training was stopped when the error (eq. (10)) of the validation set begins to increase while error (eq. (10)) of training set continues to decrease. Likewise, test set was used to evaluate the ANFIS model. Fig. 3 shows the architecture of the best ANFIS model, with two gaussian MFs for input 1–2 and one gaussian MFs for input 3, that has the lowest testing error.

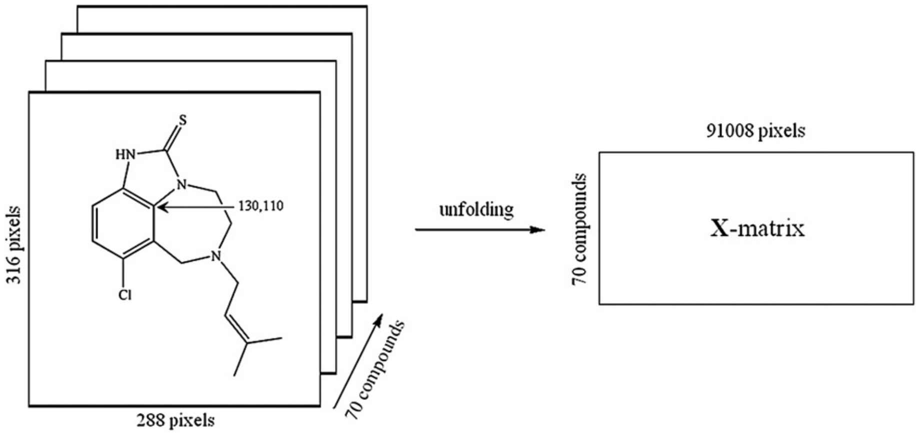


Fig. 1. Superposition of the 70 chemical structures (2D images) and unfolding step to give the X-matrix. The arrow in structure indicates the coordinate of a pixel in common among the whole series of compounds, used in the 2D alignment step.

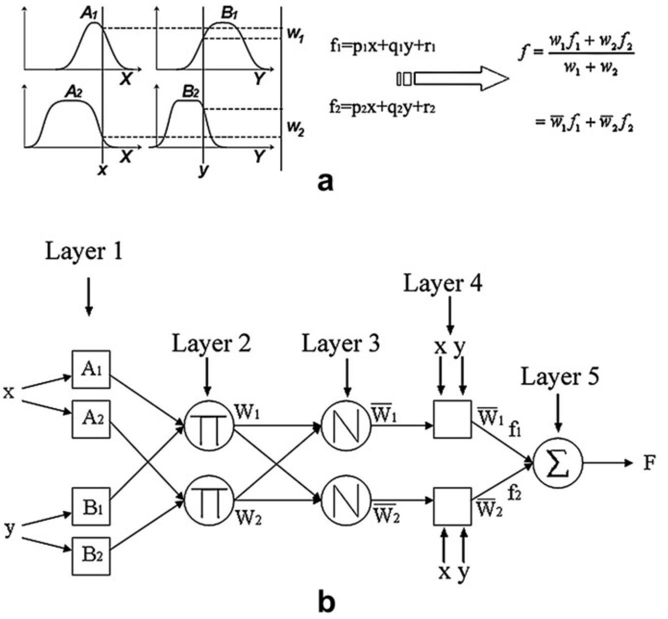


Fig. 2. (a) A two-input first-order Sugeno fuzzy model with two rules, (b) equivalent ANFIS Architecture.

3. Results and discussion

The MIA–QSAR method, when coupled to the well-known PLS and N-PLS regression methods, has presented prediction results comparable to 3D methodologies. The 2D alignment is simpler than a 3D one. Since this method is a ligand-based approach, it should be applied to a congeneric series, *i.e.* to compound sets containing a minimum of similarity. For more complex/diverse structures, but having a minimum of similarity, the congruent substructure may be superimposed (aligned) and new active compounds may be proposed. This may be performed by mixing the substructures of two different data sets containing a small moiety in common, as illustrated in a study on anti-HIV compounds [30]. Also, optical stereoisomers may be differentiated by drawing bonds differently at the chiral center [31]. In this work, MIA descriptors of a series of TIBO derivatives were applied to derive QSAR models by using PLS and N-PLS as regression methods. The predictions for external set compounds using the MIA–QSAR/PLS and N-PLS models built were found to be at least comparable to those internal/external validation data of literature, in which q^2 or r^2_{pred} varied from 0.612 to 0.885 [4–9]. The statistics for the PLS/N-PLS-based models are depicted in Table 2. Although the high predictive ability of such models, there is an increasing search for more accurate QSAR methodologies, in order to be used as reliable strategies for the design of drugs. In line with this, many descriptors usually do not explain considerable variance in Y because most regression methods, such as PLS and N-PLS, do not account for non-linearities. A recent study on check-point kinase WEE1 inhibitors illustrates well the importance of considering nonlinearity during the regression step in a QSAR modeling [20].

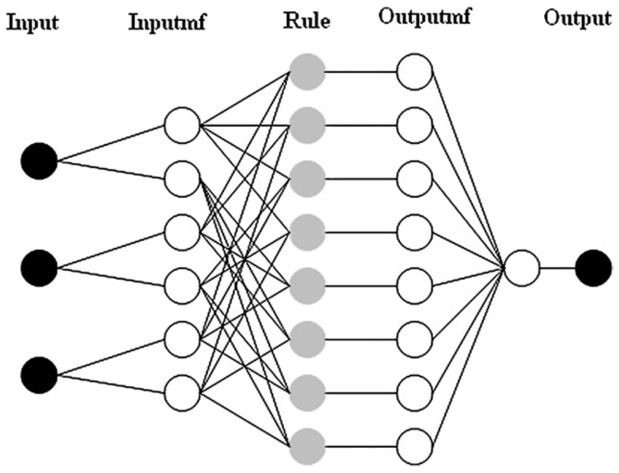


Fig. 3. ANFIS architecture for a three-input Sugeno fuzzy model with eight rules.

Table 2
Correlation matrix of selected descriptors.

	pIC_{50}	PC4	PC3	PC2
pIC_{50}	1			
PC4	0.2255	1		
PC3	0.1534	0	1	
PC2	0.0518	0.0201	0.00005	1

Thus, to obtain a more accurate model, relevant information was extracted from the X-matrix by selecting suitable regressors from the original descriptors pool. In the PCA-ranking method, the criterion for the selection of the space of the PCs containing the important response matrix is the correlation of the PCs with bioactivities. Therefore, we have inspected the scores in the space of the correlated PCs for choosing the response matrix with highest variances in this space, as shown in Table 3. The selected descriptors were regressed against the bioactivities through ANFIS; the lower prediction capability of MIA descriptors coupled to the linear methods above may be due to non-linearity, which is accounted for by ANFIS. This procedure gave excellent correlation both in calibration and validation/prediction (Table 2 and Fig. 4), and thus it is supposed to be reliably used to predict the bioactivities of novel, proposed analogs. The In order to compare these results to another non-linear method, ANFIS estimation was found to be slightly superior to radial basis functions neural network, RBFNN (Tables 1 and 3), which has been widely used for modeling and classification [32,33]. The RBFNN was constructed for the selected PCs. While estimation and prediction improved nearly 10% when using PCA–ANFIS in comparison to PLS and N-PLS, the hypothesis of systematic

Table 3
Statistical results for the MIA–QSAR models.

Parameters	Sets	PLS	N-PLS	RBFNN	ANFIS
RMSEP	Training set	0.331	0.449	0.436	0.287
	Validation set			0.363	0.279
	Test set	0.487	0.586	0.444	0.326
RSEP(%)	Training set	5.390	7.324	7.069	4.655
	Validation set			6.058	4.662
	Test set	8.154	9.807	7.433	5.461
MAE(%)	Training set	7.052	8.489	9.170	7.540
	Validation set			15.671	14.370
	Test set	14.729	16.518	14.430	11.576
Fisher	Training set	750.881	383.791	548.308	1005.606
	Validation set			105.124	190.859
	Test set	107.566	75.261	157.141	268.520
T test	Training set	27.402	19.591	23.416	31.711
	Validation set			10.253	13.815
	Test set	10.371	8.675	12.536	16.387
r^2	Training set	0.938	0.885	0.935	0.964
	Validation set			0.913	0.950
	Test set	0.871	0.825	0.908	0.944
q^2_{LOO}	Training set	0.698	0.605		
$q^2_{\mathrm{L-20\%-O}}$	Training set	0.690	0.568		

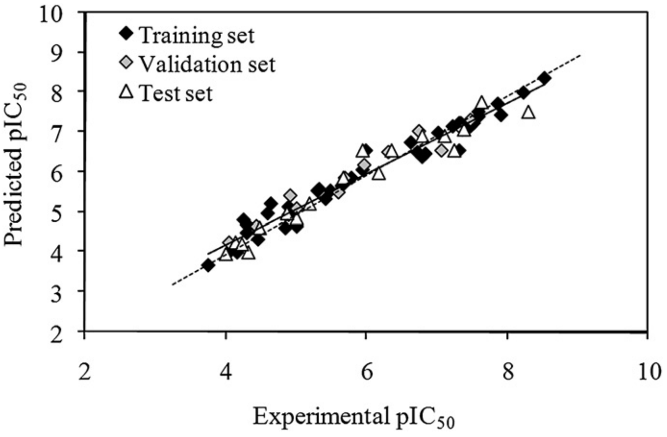


Fig. 4. Plot of experimental *versus* fitted/predicted pIC_{50} using the MIA–QSAR/PCA–ANFIS model.

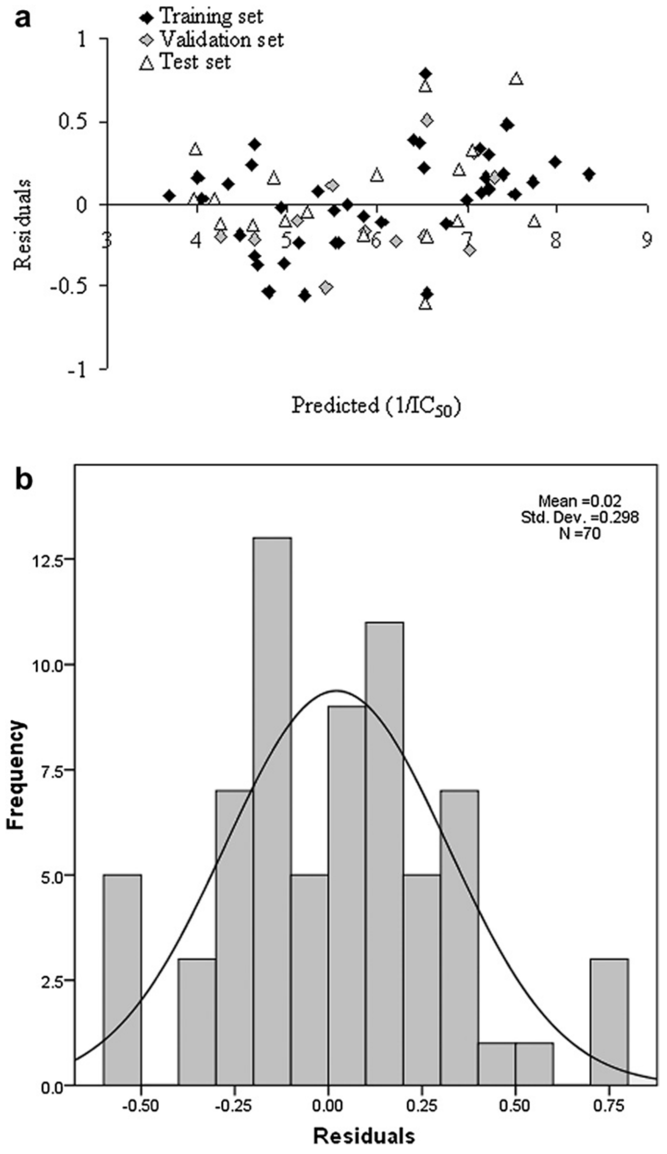


Fig. 5. Residuals analysis: (a) residuals obtained from the MIA–QSAR/PCA–ANFIS predictions; (b) histogram demonstrating the Gaussian distribution of model residuals.

error may be discarded by analyzing the residuals distribution in the plot of Fig. 5.

In summary, the proposed MIA–QSAR method coupled to PCA-ranking and ANFIS regression gave a highly predictive model, which accounts for non-linear behavior not considered by regression methods of widespread use, like MLR and PLS. Moreover, the present data required non-linear local modeling rather than the global one, justifying the improved results to those obtained previously by means of artificial neural network (ANN) [9]. Our model may be useful to predict the activity a novel compound before synthesizing it. This new structure may be derived from substructures of the training set compounds, *e.g.* a miscellany of those molecules with high bioactivities, such as compounds **20** and **36**. A rapid inspection of Table 1 reveals that S instead of O as X substituent plays an important role in enhancing the bioactivities of the TIBO derivatives, as well as the presence of halogens as R_2 substituents and DMA as R_3 substituents. However, no significant changes in biological activities are observed between 5-Me and 7-Me as R_1 substituents. Thus, a new compound containing the basic scaffold plus 7-Me, S, 8-Br and DMA as R_1, X, R_2 and R_3 substituents, respectively, has not yet been tested and may have its biological activity reliably predicted by using our MIA–QSAR/PCA–ANFIS model.

4. Conclusion

The ANFIS procedure was capable of accurately correlating the inputs (PCA scores) with the bioactivities, increasing the predictive ability of MIA descriptors when using the well-known PLS and N-PLS regression methods. Also, the predictive performance of the proposed method showed to be advantageous when compared to existing models, namely those based on CoMFA, CoMSIA, OCWLGI and classical descriptors, suggesting that the present methodology may be useful to solve other QSAR problems, especially those involving non-linearities.

Acknowledgements

CNPq is gratefully acknowledged for the fellowship (to M.P.F.), as is FAPEMIG for the financial support of this research.

References

[1] AIDS Epidemic Update (2007).http://www.unaids.org/en/KnowledgeCentre/HIVData/EpiUpdate/EpiUpdArchive/2007/Default.asp.
[2] J.T. Leonard, K. Roy, QSAR Comb. Sci. 23 (2004) 23–35.
[3] K. Roy, J.T. Leonard, Bioorg. Med. Chem. 12 (2004) 745–754.
[4] J. Huuskonen, J. Chem. Inf. Comput. Sci. 41 (2001) 425–429.
[5] Z. Zhou, J.D. Madura, J. Chem. Inf. Comput. Sci. 44 (2004) 2167–2178.
[6] A.A. Toropov, A.P. Toropova, I.V. Nesterov, O.M. Nabiev, J. Mol. Struct. (Theochem) 640 (2003) 175–181.
[7] S. Hannongbua, P. Pungpo, J. Limtrakul, P. Wolschann, J. Comput. Aided Mol. Des. 13 (1999) 563–577.
[8] V.P. Solov'ev, A. Varnek, J. Chem. Inf. Comput. 43 (2003) 1703–1719.
[9] A.S. Mandal, K. Roy, Eur. J. Med. Chem. 44 (2009) 1509–1524.
[10] M.P. Freitas, S.D. Brown, J.A. Martins, J. Mol. Struct. 738 (2005) 149–154.
[11] M.P. Freitas, Org. Biomol. Chem. 4 (2006) 1154–1159.
[12] M.P. Freitas, Med. Chem. Res. 16 (2007) 461–467.
[13] J.R. Pinheiro, M. Bitencourt, E.F.F. da Cunha, T.C. Ramalho, M.P. Freitas, Bioorg. Med. Chem. 16 (2008) 1683–1690.
[14] M.P. Freitas, Chemom. Intell. Lab. Syst. 91 (2008) 173–175.
[15] M.P. Freitas, R. Rittner, QSAR Comb. Sci. 27 (2008) 582–585.
[16] J.E. Antunes, M.P. Freitas, R. Rittner, Eur. J. Med. Chem. 43 (2008) 1632–1638.
[17] J.E. Antunes, M.P. Freitas, E.F.F. da Cunha, T.C. Ramalho, R. Rittner, Bioorg. Med. Chem. 16 (2008) 7599–7606.
[18] M. Goodarzi, M.P. Freitas, QSAR Comb. Sci. 27 (2008) 1092–1097.
[19] M. Bitencourt, M.P. Freitas, Med. Chem. 5 (2009) 79–86.
[20] R.A. Cormanich, M. Goodarzi, M.P. Freitas, Chem. Biol. Drug Des. 73 (2009) 244–252.
[21] M. Goodarzi, M.P. Freitas, Chemom. Intell. Lab. Syst. 96 (2009) 59–62.

[22] M.P. Freitas, E.F.F. da Cunha, T.C. Ramalho, M. Goodarzi, Curr. Comput.-Aided Drug Des. 4 (2008) 273–282.

[23] K. Pearson, Philos. Magazine 6 (1901) 559–572.

[24] L.A. Zadeh, Inf. Contr 8 (1965) 338–353.

[25] M. Sugeno, G. Kang, Fuzzy Set. Sys. 28 (1988) 15–33.

[26] ACD/ChemSketch Version 11.02. Advanced Chemistry Development, Inc., Toronto, Ont., Canada, 2008.

[27] Matlab Version 7.5. MathWorks Inc., Natick, MA, 2007.

[28] Y.L. Loukas, J. Med. Chem. 44 (2001) 2772–2783.

[29] V. Centner, O. de Noord, D. Massart, Anal. Chim. Acta. 376 (1998) 153–168.

[30] J.R. Pinheiro, M. Bitencourt, E.F.F. da Cunha, T.C. Ramalho, M.P. Freitas, Bioorg. Med. Chem. 16 (2008) 1683–1690.

[31] M. Goodarzi, M.P. Freitas, Separ. Purif. Technol. 68 (2009) 363–366.

[32] X.J. Yao, A. Panaye, 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.

[33] J. Shi, F. Luan, H.X. Zhang, M.C. Liu, Q.X. Guo, Z.D. Hu, B.T. Fan, QSAR Comb. Sci. 25 (2006) 147–155.