Journal of Pharmaceutical and Biomedical Analysis 50 (2009) 853–860

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Shuffling multivariate adaptive regression splines and adaptive neuro-fuzzy inference system as tools for QSAR study of SARS inhibitors

M. Jalali-Heravi*, M. Asadollahi-Baboli, A. Mani-Varnosfaderani

Department of Chemistry, Sharif University of Technology, P.O. Box 11155-9516, Tehran, Iran

ARTICLE INFO

Article history:
Received 28 April 2009
Received in revised form 4 July 2009
Accepted 6 July 2009
Available online 14 July 2009

Keywords:
Severe acute respiratory syndrome
Pyridine N-oxide derivatives
Quantitative structure–activity relationship
Multivariate adaptive regression splines
Adaptive neuro-fuzzy inference system

ABSTRACT

In this work, the inhibitory activity of pyridine N-oxide derivatives against human severe acute respiratory syndrome (SARS) is predicted in terms of quantitative structure–activity relationship (QSAR) models. These models were developed with the aid of multivariate adaptive regression spline (MARS) and adaptive neuro-fuzzy inference system (ANFIS) combined with shuffling cross-validation technique. A shuffling MARS algorithm is utilized to select the most important variables in QSAR modeling and then these variables were used as inputs of ANFIS to predict SARS inhibitory activities of pyridine N-oxide derivatives. A data set of 119 drug-like compounds was coded with over hundred calculated meaningful molecular descriptors. The best descriptors describing the inhibition mechanism were solvation connectivity index, length to breadth ratio, relative negative charge, harmonic oscillator of aromatic index, average molecular weight and total path count. These parameters are among topological, electronic, geometric, constitutional and aromaticity descriptors. The statistical parameters of R^2 and root mean square error (RMSE) are 0.884 and 0.359, respectively. The accuracy and robustness of shuffling MARS–ANFIS model in predicting inhibition behavior of pyridine N-oxide derivatives (pIC_{50}) was illustrated using leave-one-out and leave-multiple-out cross-validation techniques and also by *Y*-randomization. Comparison of the results of the proposed model with those of GA-PLS-ANFIS shows that the shuffling MARS–ANFIS model is superior and can be considered as a tool for predicting the inhibitory behavior of SARS drug-like molecules.

1. Introduction

The discovery of a novel human coronavirus (H-CoV) as the cause of the newly recognized severe acute respiratory syndrome (SARS) provides a new challenge to the medical community to keep control on this disease [1]. Human coronaviruses cause up to 30% of colds and they sometimes cause a lower respiratory tract disease. In contrast, animal coronaviruses are known to cause devastating epizootics of respiratory or enteric diseases in livestock and poultry [2]. The SARS coronavirus is clearly new to the human population and its RNA genome differs substantially from sequences of all known coronaviruses.

SARS, with high rates of transmission needs a rapid, sensitive and inexpensive treatment method that can be used to effectively prevent the rapid spread of the infection. Therefore, it is wise to develop safe and effective drugs against SARS-CoV as quickly as possible in case a novel widespread outbreak would occur. The development of effective drugs against SARS-CoV may also provide new strategies for the prevention or treatment of other coronavirus diseases in animals or humans [3]. SARS inhibitor has potential therapeutic value and has been extensively studied in pharmaceutical industry [4]. Recently, a total of 119 compounds that all belong to the class of the pyridine N-oxide derivatives with good inhibitory concentration has been reported against SARS-CoV [5].

To find and design new compounds with enhanced inhibitory activity, a systematic study of the different substituents on the activity of the analogues is needed. On the other hand, the growth of computational techniques has accelerated the drug design process. Many databases of inhibitors exist that have yet to be evaluated against SARS. Quantitative structure–activity relationship (QSAR) has been demonstrated as a capable tool for the investigation of bioactivity of various classes of compounds [6–12].

Experimental evaluation of inhibitory activity of newly designed compounds is time-consuming and expensive; as a result, it is of interest to develop a method for the prediction of biological activity before the synthesis. QSAR searches information relating chemical structure to biological activity by developing a mathematical model. Building of a QSAR model begins with calculating theoretical parameters or selecting structural features for the compounds involved. Nowadays, hundreds of descriptors could be generated in QSAR studies, but only some of them are statistically significant in terms of correlation with biological activity for a particular analy-

* Corresponding author. Tel.: +98 21 66165315; fax: +98 21 66012983.
E-mail address: jalali@sharif.edu (M. Jalali-Heravi).

0731-7085/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2009.07.009

sis. Therefore, variable selection techniques have become important for producing a useful predictive model. A suitable feature selection method ensures the model stability and the consistency of relationship between the descriptors and biological activity [13].

In order to make sure that the most important descriptors have been selected, shuffling cross-validation technique was used in this work. In this method, the data set was divided into several subsets, and variable selection process was performed for different combinations of these subsets. Then the most frequent descriptors in models were selected as most important variables describing the inhibitory effect.

In this study, multivariate adaptive regression spline (MARS) combined with shuffling cross-validation (SCV) was employed to select the most important parameters describing SARS inhibitors activity. The selected descriptors were then used as inputs of adaptive neuro-fuzzy inference system (ANFIS) and a hybrid model called shuffling MARS–ANFIS was developed. As final step, the generated model was used to predict the activity of pyridine N-oxide derivatives as SARS inhibitors.

2. Computational methods

2.1. Multivariate adaptive regression splines

Multivariate adaptive regression spline (MARS) is a non-parametric regression method proposed by Friedman in 1991 [14,15]. Nowadays, the MARS is used for analyzing biological, economical, sociological and other databases [16].

The main idea in MARS which makes it different from other methods is its ability for dividing the whole space of each independent variable into various sub-regions and then defining a different mathematical equation for each area. This equation relates each sub-region of independent variable to response of the system, separately. This framework makes the MARS a method that is useful for modeling non-linear and complicated systems and also applicable for the conditions which the behavior of the system is highly affected by just a specific area of independent variable.

Generally a regression couple can be presented by (X_i, Y_i, which X_i, represents for one or, n, independent variable(s) and Y_i is a dependent variable. In the MARS model, for every independent variable there is/are one or more split point(s), named t_i. For X_i greater than t_i, there is one equation named right side-basis function (BF) and for X_i less than t_i there is another equation named left side-basis function. These two left and right basis functions (spline functions) relate X_i to the dependent variable Y_i. The following equations indicate the mathematical representation of right and left basis functions:

$$[-(X_i - t_i)]_+^q = \begin{cases} (t_i - X_i)^q & \text{If } X_i < t_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$[+(X_i - t_i)]_+^q = \begin{cases} (X_i - t_i)^q & \text{If } X_i \geq t_i \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where $q(\geq 0)$ is the power to which the splines are raised and which determines the degree of smoothness of the resultant function estimate. Final response in MARS can be calculated by summing up all M basis functions with suitable coefficients (c_m) as:

$$\hat{Y} = f_M(X) = c_0 \sum_{m=1}^{M} c_m B_m(X) \quad (3)$$

where $\hat{Y}$ is the dependent variable predicted by MARS model, c_0 is a constant and $B_m(X)$ is the mth basis function.

To determine which basis function should be included in the model, MARS utilizes the generalized cross-validation (GCV). The GCV is mean squared residual error divided by a penalty dependent on the model complexity. The GCV is defined in the following way:

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_M(X_i))^2}{((1 - C(M))/n)^2} \quad (4)$$

where $C(M)$ is the complexity penalty that increases with the number of basis functions in the model and can be defined as Eq. (5):

$$C(M) = C(M+1) + dM \quad (5)$$

where M is the number of basis functions in this equation and the parameter d is a penalty for each basis function included in the model. Large value of d leads to fewer basis functions and therefore smoother function estimates. The theory behind the multivariate adaptive regression spline has been adequately described elsewhere [17].

In this study, the data set containing p observations were divided into $(p-k)$ calibration and k validation objects. The root mean square error of validation set ($RMSE_v$) has been used as fitness function for the search algorithm. Because there are various states for selecting k samples out of p, various models can be built using external validation strategy. The parameters of the generated models depend on training and validation set, as a result various variables and split points are expected to be determine in this way. We have used the most frequent variables appearing in the built models as inputs for the final modeling.

2.2. Shuffling cross-validation

In this technique, the data set would be divided into several subsets, and variable selection process and model developing would be performed for all combinations of the subsets. Then the most frequent descriptors appeared in the developed models would be selected as most important variables in describing the variation in inhibitor activity.

In the present work, the data set was randomly divided into six subsets (A–F). For variable selection procedure, four groups were applied as calibration set and the two remaining subsets were used as validation set for evaluating the selected parameters. Mathematically, there are fifteen possible states that one can select four unrepeated objects from six independent ones. The data set was divided into six subgroups, so, fifteen MARS models can be developed with various calibration and validation sets.

The molecules included in subsets of A–F are shown in Table 1 . Fifteen different combinations of calibration and validation subsets were used in the present study to develop the MARS model. The use of shuffling MARS technique guarantees that the developed model is robust and reliable and it is not obtained by chance.

2.3. Adaptive neuro-fuzzy inference system

The proposed neuro-fuzzy model in ANFIS is a multilayer neural network-based fuzzy system [18]. Its topology is shown in Fig. 1, and as can be seen the system has a total of five layers. In this connectionist structure, the input (*layer 0*) and output (*layer 5*) nodes represent the descriptors and the response, respectively, and in the hidden layers, there are nodes functioning as membership functions (MFs) and rules. This eliminates the disadvantage of a normal feed forward multilayer network, which is difficult for an observer to understand or to interpret its results. ANFIS simulates TSK (Takagi–Sugeno–Kang) fuzzy rule [19] of type-3 where the consequent part of the rule is a linear combination of input variables

Table 1
Experimental and calculated inhibitor data using shuffling MARS–ANFIS model for pyridine N-oxide derivatives.

No.	Subset[a]	R[b]	X1	X2	X3	X4	X5	Z1	Z2	Y1	Y2	Y3	Y4	Exp. pIC_{50}	MARS–ANFIS pIC_{50}
1	A	H	H	H	H	H	H	O	O	H	H	H	H	3.840	3.866
2	C	H	Me	H	H	H	H	O	O	H	H	H	H	4.060	4.272
3	B	H	H	Me	H	H	H	O	O	H	H	H	H	4.059	4.327
4	D	H	H	Me	H	H	H	O	O	H	H	H	H	3.825	3.952
5	A	H	H	H	Me	H	H	O	O	H	H	H	H	4.040	4.232
6	B	H	H	H	Me	H	H	O	–	H	H	H	H	4.352	4.845
7	E	H	Me	H	Me	H	H	O	O	H	H	H	H	4.239	4.082
8	D	H	Me	H	H	Me	H	O	O	H	H	H	H	3.938	4.359
9	C	H	Me	H	H	Me	H	O	–	H	H	H	H	3.736	3.325
10	F	H	Me	Me	H	H	Me	O	O	H	H	H	H	4.423	4.721
11	A	H	Me	Me	H	H	Me	O	–	H	H	H	H	5.294	5.128
12	E	H	H	H	Et	H	H	O	O	H	H	H	H	4.361	4.150
13	D	H	H	H	iProp	H	H	O	O	H	H	H	H	4.239	4.633
14	C	H	iProp	H	H	iProp	H	O	O	H	H	H	H	4.491	4.699
15	B	H	iProp	H	H	iProp	H	O	–	H	H	H	H	4.717	4.327
16	E	H	H	H	tBut	H	H	O	O	H	H	H	H	4.502	4.663
17	D	H	H	H	tPent	H	H	O	O	H	H	H	H	4.652	4.626
18	A	H	H	H	OMe	H	H	O	O	H	H	H	H	4.756	4.874
19	F	H	H	H	OMe	H	H	O	–	H	H	H	H	3.955	3.910
20	C	H	OMe	H	H	OMe	H	O	O	H	H	H	H	5.098	5.199
21	E	H	H	OMe	OMe	H	H	O	O	H	H	H	H	3.712	3.738
22	B	H	H	OMe	OMe	H	H	O	–	H	H	H	H	3.878	4.083
23	C	H	H	OMe	OMe	OMe	H	O	O	H	H	H	H	5.321	5.680
24	F	H	H	OMe	OMe	OMe	H	O	–	H	H	H	H	3.823	3.498
25	E	H	OMe	H	H	Me	H	O	O	H	H	H	H	3.811	4.002
26	A	H	OMe	H	H	Me	H	–	–	H	H	H	H	3.834	3.752
27	E	H	OEt	H	H	H	H	O	O	H	H	H	H	4.037	3.952
28	D	H	OEt	H	H	H	H	O	–	H	H	H	H	3.899	3.905
29	C	H	H	F	H	H	H	O	O	H	H	H	H	3.651	3.250
30	B	H	H	H	F	H	H	O	–	H	H	H	H	4.148	4.250
31	F	H	Cl	H	H	H	H	O	O	H	H	H	H	3.461	3.138
32	A	H	Cl	H	Cl	H	H	O	O	H	H	H	H	4.174	4.007
33	E	H	Cl	H	H	H	Cl	O	O	H	H	H	H	4.710	4.872
34	B	H	H	Cl	Cl	H	H	O	O	H	H	H	H	4.327	4.140
35	D	H	Cl	Cl	H	H	Cl	O	O	H	H	H	H	5.040	4.892
36	F	H	Cl	Cl	Cl	Cl	Cl	O	O	H	H	H	H	4.734	4.809
37	A	H	Cl	Cl	Me	Cl	Cl	O	O	H	H	H	H	5.546	5.770
38	E	H	Cl	H	NO_2	H	H	O	O	H	H	H	H	5.302	5.119
39	A	H	H	Br	H	H	H	O	O	H	H	H	H	4.732	4.567
40	C	H	Br	H	H	OMe	H	O	O	H	H	H	H	4.647	4.170
41	F	H	iProp	H	Br	iProp	H	O	O	H	H	H	H	4.378	4.190
42	B	H	I	H	H	H	H	O	O	H	H	H	H	4.972	4.892
43	F	H	NO_2	H	H	H	H	O	O	H	H	H	H	4.176	4.103
44	C	H	H	H	NO_2	H	H	O	O	H	H	H	H	4.355	4.349
45	E	H	H	NO_2	H	NO_2	H	O	O	H	H	H	H	4.588	4.404
46	B	H	H	NO_2	Me	H	H	O	O	H	H	H	H	4.887	4.438
47	F	H	H	Me	NO_2	H	H	O	O	H	H	H	H	4.726	4.421
48	C	H	Me	H	H	NO_2	H	O	O	H	H	H	H	4.799	4.849
49	A	H	OMe	H	H	NO_2	H	O	O	H	H	H	H	3.733	3.717
50	D	H	H	NO_2	Cl	H	H	O	O	H	H	H	H	4.281	4.439
51	F	H	CN	H	H	H	H	O	O	H	H	H	H	4.883	4.672
52	B	H	H	H	CN	H	H	O	O	H	H	H	H	4.262	4.069
53	E	H	H	H	Phe	H	H	O	O	H	H	H	H	4.359	4.188
54	D	H	OPhe	H	H	H	H	O	O	H	H	H	H	4.850	4.615
55	A	H	H	OMe	OBz	H	H	O	O	H	H	H	H	4.533	4.783
56	B	H	H	CF_3	H	H	H	O	–	H	H	H	H	4.667	4.873
57	C	H	OH	H	H	NO_2	H	–	–	H	H	H	H	3.747	3.250
58	F	Me	H	H	H	H	H	O	–	H	H	H	H	3.876	3.994
59	B	Me	H	H	Me	H	H	O	–	H	H	H	H	5.420	5.237
60	C	Me	Me	H	H	Me	H	–	–	H	H	H	H	5.646	5.860
61	D	Me	H	H	F	H	H	O	–	H	H	H	H	3.545	3.407
62	A	Me	Cl	H	H	Me	H	O	–	H	H	H	H	5.905	5.390
63	E	Et	H	H	H	H	H	O	O	H	H	H	H	6.192	5.717
64	C	Et	Me	H	H	Me	H	O	O	H	H	H	H	3.658	3.525
65	F	Prop	H	H	H	H	H	O	O	H	H	H	H	3.582	3.633
66	B	Prop	H	H	H	H	H	–	–	H	H	H	H	3.831	3.934
67	D	Prop	Me	H	H	Me	H	–	–	H	H	H	H	3.622	3.473
68	A	Hept	H	Me	Me	Me	H	–	–	H	H	H	H	4.495	4.380
69	F	Hept	Me	H	H	Me	H	O	O	H	H	H	H	4.252	4.272
70	D	Undec	Me	H	H	Me	H	O	O	H	H	H	H	5.043	4.767
71	C	Isobut	Me	H	H	Me	H	O	O	H	H	H	H	3.691	3.399
72	E	C_3H_6	Me	H	H	Me	H	O	O	H	H	H	H	4.731	4.549
73	B	C_6H_5	Me	H	H	H	H	O	O	H	H	H	H	3.922	3.603
74	C	C_6H_5	Me	H	H	Me	H	O	O	H	H	H	H	3.859	3.781

Table 1 (*Continued*)

No.	Subset[a]	R[b]	X1	X2	X3	X4	X5	Z1	Z2	Y1	Y2	Y3	Y4	Exp. pIC_{50}	MARS–ANFIS pIC_{50}
75	D	C_6H_5	Me	H	H	Me	H	–	–	H	H	H	H	3.915	3.749
76	F	CH_2Ph	Me	H	H	Me	H	–	–	H	H	H	H	3.976	4.003
77	A	CN	Me	H	H	Me	H	–	–	H	H	H	H	5.111	5.216
78	D	CH_2CO_2H	Me	H	H	Me	H	O	O	H	H	H	H	3.900	3.617
79	F	Br	Me	H	H	Me	H	O	O	H	H	H	H	3.567	3.598
80	B	CO_2CH_3	Me	H	H	H	H	O	–	H	H	H	H	3.643	3.576
81	C	CO_2CH_3	Me	H	H	Me	H	O	O	H	H	H	H	3.769	3.392
82	D	CO_2CH_3	H	OPh	H	H	H	O	O	H	H	H	H	3.969	3.860
83	F	CF_3	Me	H	H	Me	H	O	O	H	H	H	H	4.186	4.199
84	A	CH_2OMe	Me	H	H	Me	H	O	O	H	H	H	H	3.926	3.480
85	E	Me, Cl	H	H	H	H	H	O	O	H	H	H	H	4.143	3.895
86	C	Me, Cl	Me	H	H	Me	H	O	O	H	H	H	H	4.474	4.177
87	B	Me	H	H	Me	H	Me	O	O	H	H	Me	H	4.151	3.881
88	D	H	H	H	H	H	H	O	O	H	H	H	Me	4.087	4.356
89	F	Me	H	H	Me	H	Me	O	O	H	H	H	Me	3.973	3.970
90	A	Me	H	H	Me	H	H	O	–	H	H	H	Me	3.905	3.804
91	E	Me	H	H	H	H	H	O	O	H	H	H	Me	3.817	3.616
92	D	Me	H	Me	H	H	H	O	O	H	H	H	Me	4.213	4.120
93	B	Me	H	H	Me	H	Et	O	O	H	H	H	Me	4.423	4.289
94	F	Cl	H	H	H	Cl	H	–	–	H	H	H	Me	4.203	4.127
95	C	H	H	H	H	H	Me	O	O	H	H	H	Me	3.986	3.957
96	E	Cl	H	H	H	H	H	O	O	H	H	H	Me	3.665	3.634
97	A	Me	NO_2	H	H	H	H	O	O	H	H	H	Me	4.152	4.113
98	B	Me	H	Me	H	H	Me	O	O	H	H	H	Me	4.526	4.321
99	F	Cl	H	H	H	H	Me	O	O	H	H	H	Me	3.832	3.577
100	D	Me	NO_2	H	H	H	Me	O	O	H	H	H	Me	3.914	3.719
101	E	Me	H	H	Me	H	H	O	O	H	H	H	OMe	3.545	3.801
102	A	Me	H	H	Me	H	H	O	–	H	H	H	OMe	4.585	4.150
103	C	Me	H	H	Me	H	H	–	–	H	H	H	OMe	3.672	3.983
104	E	Me	H	H	Me	H	H	–	–	H	H	H	OH	3.595	3.088
105	B	H	H	OMe	H	H	H	O	O	H	H	t-Bu	H	3.604	3.629
106	D	H	H	OMe	H	H	H	–	–	H	H	t-Bu	H	3.946	4.304
107	E	H	H	H	H	H	H	–	–	Cl	H	H	H	3.801	3.871
108	A	Me	H	H	Me	H	H	O	O	Cl	H	H	H	3.720	3.324
109	B	Me	H	H	Me	H	Me	O	O	Cl	H	H	H	4.892	4.610
110	F	H	H	H	H	H	H	–	–	H	Cl	H	H	4.860	4.651
111	C	Me	H	H	Me	H	H	O	O	H	Cl	H	H	3.633	3.495
112	E	Me	H	H	Me	H	H	O	–	H	Cl	H	H	4.709	4.628
113	A	H	H	H	H	H	Cl	O	O	H	Cl	H	H	4.325	4.674
114	D	H	H	H	H	H	H	O	O	H	H	H	Cl	5.141	4.740
115	F	H	H	H	H	H	H	O	–	H	H	H	Cl	5.277	5.216
116	C	Me	H	H	Me	H	H	–	–	H	H	H	Cl	4.860	4.638
117	B	Me	H	H	Me	H	Me	O	O	H	H	H	Cl	3.824	4.007
118	D	Me	H	H	Me	H	Cl	O	O	H	H	H	Cl	5.283	5.216
119	A	H	H	H	H	H	H	–	–	H	H	H	NO_2	4.363	4.643

[a] A–F subsets.
[b] Substituted groups in pyridine N-oxide derivatives is shown in Fig. 2.

and a constant. For a Sugeno fuzzy model a common rule set with the fuzzy if-then rules is as following:

Rule 1: IF x is A_1 and y is B_1 THEN

$$f_1 = p_{1X} + q_{1Y} + r_1$$

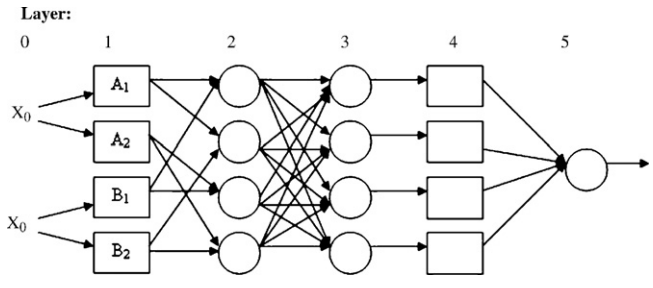

Fig. 1. A typical ANFIS structure.

Rule 2: IF x is A_2 and y is B_2 THEN

$$f_2 = p_{2X} + q_{2Y} + r_2$$

For simplicity, we assume that the examined fuzzy inference system has two inputs x and y and one output. The ANFIS contains five layers (Fig. 1):

Layer 1. The fuzzy part of ANFIS is mathematically incorporated in the form of membership functions (MFs). A membership function $\mu_{Ai}(x)$ can be any continuous and piecewise differentiable function that transforms the input value x into a membership degree, that is to say a value between 0 and 1. The most widely applied membership functions are the generalized bell (gbell MF) or the Gaussian function in Eqs. (6) and (7), which are described by the three parameters, a–c. Therefore, Layer 1 is the *fuzzification* layer in which each node represents a membership:

$$\mu_{Ai}(x) = \frac{1}{1 + \left[((x - c_i)/a_i)^2\right]^{bi}} \quad (6)$$

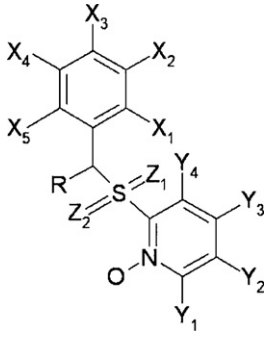

Fig. 2. Main skeleton with different functional positions of pyridine N-oxide derivatives.

$$\mu_{Ai}(x) = \exp\left[-((x - c_i)/a_i)^2\right] \quad (7)$$

As the values of the parameters $\{a_i, b_i \text{ and } c_i\}$ change, the bell-shaped functions vary accordingly, thus exhibiting various forms of membership functions on linguistic label *Ai*. Parameters in this layer are referred to as premise parameters.

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

$$O_{2,1} = w_i = \mu_{Ai}(x) \times \mu_{Bi}(y) \quad \text{for } i = 1, 2 \quad (8)$$

Every node in this layer computes the multiplication of the input values and gives the product as the output as in the above equation. The membership values represented by $\mu_{Ai}(x)$ and $\mu_{Bi}(y)$ are multiplied in order to find the firing strength of a rule where the variables *x* and *y* have linguistic values *Ai* and *Bi*, respectively.

Layer 3. This layer is the normalization layer which normalizes the strength of all rules according to Eq. (9):

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1 + w_2} \quad \text{for } i = 1, 2 \quad (9)$$

where w_i is the firing strength of the *i*th rule which is computed in layer 2. Node *i* computes the ratio of the *i*th rule's firing strength to the sum of all rules' firing strengths. 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:

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i(p_i x + q_i y + r_i) \quad (10)$$

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.

Table 2
Selecting the important variables using shuffling MARS method.

Run	Calibration set	R^2_{Cal}	$RMSE_{Cal}$	Validation set	R^2_{Val}	$RMSE_{Val}$
1	A+B+C+D	0.834	0.241	E+F	0.767	0.458
2	A+B+C+E	0.820	0.268	D+F	0.810	0.372
3	A+B+D+E	0.831	0.279	C+F	0.805	0.367
4	A+C+D+E	0.835	0.253	B+F	0.751	0.476
5	B+C+D+E	0.803	0.226	A+F	0.740	0.450
6	A+B+C+F	0.833	0.240	D+E	0.802	0.393
7	A+B+D+F	0.819	0.273	C+E	0.783	0.422
8	A+C+D+F	0.843	0.226	B+E	0.745	0.449
9	B+C+D+F	0.825	0.282	A+E	0.730	0.470
10	A+B+E+F	0.839	0.228	C+D	0.804	0.418
11	A+C+E+F	0.813	0.265	B+D	0.806	0.416
12	B+C+E+F	0.826	0.250	A+D	0.784	0.464
13	A+D+E+F	0.837	0.242	B+C	0.787	0.466
14	B+D+E+F	0.821	0.255	A+C	0.769	0.471
15	C+D+E+F	0.818	0.235	A+B	0.750	0.483
Mean		0.827	0.251		0.776	0.438

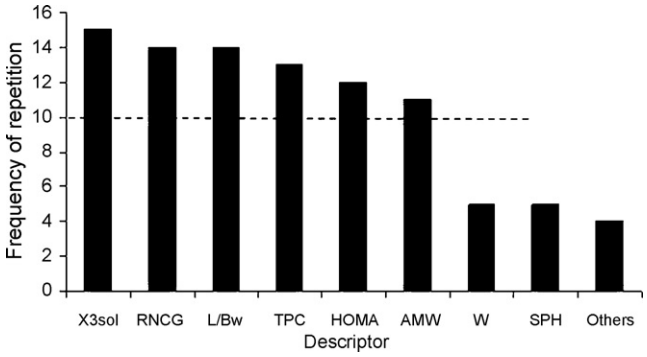

Fig. 3. The selected descriptors and the frequency of each one in the shuffling MARS models.

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:

$$\text{overall output} = O_{5,i} = \sum_i \bar{w}_i f_i = \frac{\sum_i \bar{w}_i f_i}{\sum_i w_i} \quad (11)$$

Thus we have constructed an ANFIS system that is functionally equivalent to Sugeno fuzzy model. This system is used in the present QSAR study due to its transparency and efficiency.

3. Data set collection and descriptor generation

A set of 119 variously functionalized pyridine N-oxide was collected along with their activity data [5]. The IC_{50} values were converted to pIC_{50} values and used as dependent variables in the QSAR study. The main skeleton with different functional positions for pyridine N-oxide derivatives is shown in Fig. 2. A list of inhibitory activities is given in Table 1. Prior to the calculation of the molecular descriptors, the 3D structures of the studied compounds were optimized using semi-empirical quantum-chemical methods of PM3 implemented in Hyperchem computer program [20]. In this work, over hundred meaningful descriptors were calculated for each compound, which encoded different aspects of the molecular structures. These descriptors were consisted of constitutional, topological, electronic, geometric and empirical descriptors. Pairs of descriptors that were highly correlated ($R > 0.90$) encoded

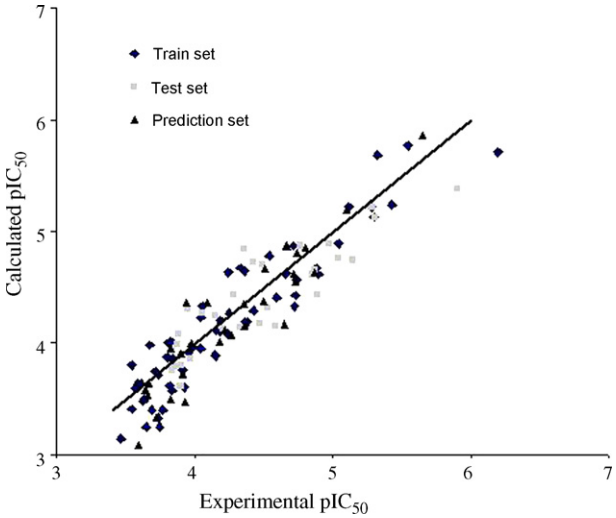

Fig. 4. Plot of the shuffling MARS–ANFIS calculated pIC_{50} values against the experimental ones for the training, test and validation sets.

similar information, and therefore one of them has been eliminated. Descriptors with constant or almost constant values for all molecules were also eliminated. All these molecular descriptors were generated using Dragon3 software [21]. Table 2 shows 15 different combinations of calibration and validation subsets used for the variable selection via shuffling MARS. Fig. 3 shows the selected descriptors and the frequency of each descriptor that has been appeared in the shuffling MARS models. Shuffling MARS–ANFIS algorithm was written in our laboratory using MATLAB 7.0 [22] and run on a personal computer (Intel Pentium processor 4/1.8 GHz 1 GB RAM).

4. Results and discussion

4.1. Shuffling MARS–ANFIS modeling

First, all 119 molecules studied in this work were sorted according to their biological activity. Then the molecules were divided into six groups, five groups of them consisted of twenty molecules each and one consisted of nineteen molecules. Each group was selected in such a way that it consisted of all range of inhibitory activity from weak to highly active compounds. In the variable selection procedure, four groups were applied as calibration set and the two remaining subsets were used as validation set for evaluating the selected parameters. The data set was divided into six subgroups, so, we can make 15 MARS models with various calibration and validation sets. Because these calibration and validation sets contain different molecules, various descriptors are expected to be selected by MARS search strategy, in each model. In the calibration procedure, the forward selection and backward deletion algorithm uses the parameter, root mean square of validation set ($RMSE_v$) as an index for evaluating the selected split points. Statistical parameters obtained for 15 models are shown in Table 2. The selected descriptors and the frequency of each descriptor in shuffling-MARS models are shown in Fig. 3. Inspection of this figure shows that parameters of solvation connectivity index (X3sol), length to breadth ratio (L/Bw), relative negative charge (RNCG), harmonic oscillator of aromatic index (HOMA), average molecular weight (AMW) and total path count (TPC) have appeared more frequently (more than 10 runs) in the 15 runs compared to the other descriptors. These six descriptors are among topological, electronic, geometric, constitutional and aromaticity descriptors. The detailed description of these descriptors is given in Reference [23]. The most important selected variables (six variables) using the shuffling MARS algorithm were used as inputs for developing the ANFIS model to predict the value of $p\mathrm{IC}_{50}$ for the SARS inhibitors.

The ANFIS modeling involves two steps: (a) structure identification and (b) parameter identification. The former is related to finding a suitable number of rules and a proper partition of the feature space. The latter is concerned with the adjustment of system parameters, such as membership function (MF) parameters, linear coefficients, and so on. It is concluded that by increasing the number of MFs per input, the number of rules increases accordingly [13]. For the first stage of ANFIS modeling grid partitioning was used for partitioning the features. The number and type of membership functions were optimized using RMSE as a criterion for the test set.

For the ANFIS modeling, data set was divided into three groups: training, test and prediction sets. All molecules were randomly included in these sets. The training set, consisted of 70 molecules and was used for the model generation. However, the test set, consisted of 30 molecules, was used to take care of the overtraining. The prediction set, consisted of 19 molecules, was used to evaluate the generated model.

The predicted values of pyridine N-oxide inhibition behavior obtained using shuffling MARS–ANFIS model are listed in Table 1.

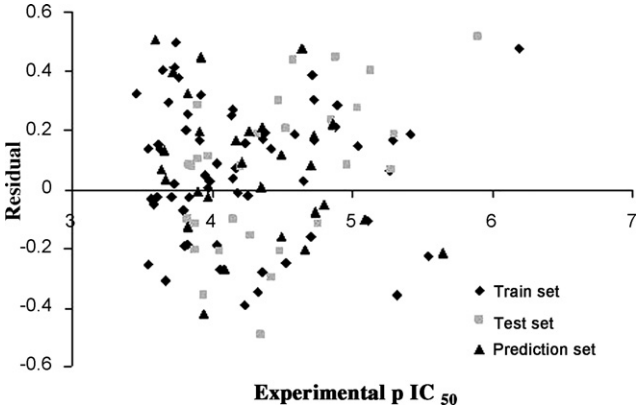

Fig. 5. Plot of residuals versus experimental values of $p\mathrm{IC}_{50}$ for the shuffling MARS–ANFIS.

This table shows that the calculated $p\mathrm{IC}_{50}$ is a good estimate of experimental $p\mathrm{IC}_{50}$. The correlation between the experimental and calculated values of $p\mathrm{IC}_{50}$ is shown in Fig. 4. The adjusted R^2 for train, test and prediction set in shuffling MARS–ANFIS model are 0.856, 0.862 and 0.870, respectively. Also the RMSE for train, test and prediction set are 0.285, 0.337 and 0.382, respectively. The residuals of the calculated values of $p\mathrm{IC}_{50}$ are plotted against the experimental values in Fig. 5. The propagation of the residuals in both sides of zero line indicates that no symmetric error exists in the development of the QSAR model. From this figure, one can find there is no out-layer in the generated shuffling MARS–ANFIS model.

4.2. Validation of shuffling MARS–ANFIS model

Second step of this work was investigating the validity of the generated model. The consistency and reliability of a method can be explored using the cross-validation techniques [24]. The cross-validation techniques of leave-one-out (LOO-CV) and leave-multiple-out (LMO-CV) were used to assess the consistency of the model. In order to examine the robustness of the developed model, the *Y*-randomization test was performed in this contribution. In LOO-CV algorithm, one compound was left in each step as prediction set and the model was developed using the remaining molecules as training set [24]. The accuracy of cross-validation results is extensively acceptable in the literature considering Q^2_{LOO} value using Eq. (11):

$$Q^2 = \frac{\mathrm{PRESS}}{\mathrm{SSY}} = 1 - \frac{\sum_{i=1}^{n}(y_{\mathrm{exp}} - y_{\mathrm{pred}})^2}{\sum_{i=1}^{n}(y_{\mathrm{exp}} - \bar{y})^2} \quad (11)$$

In this sense, a high value for the statistical parameter ($Q^2 > 0.5$) is considered as proof of high predictive ability of the model [25]. However, several authors suggest that a high value of Q^2_{LOO} appears to be necessary but not sufficient [26]. Consequently, we also used LMO-CV and *Y*-randomization techniques. In the case of LMO, *M* represents a group of randomly selected data points which would leave out at the beginning and would be predicted by the model which was developed using the remaining data points. Therefore, *M* molecules are considered as prediction set. The R^2_{LMO} can be calculated by using Eq. (12):

$$R^2_{\mathrm{LMO}} = \frac{\mathrm{PRESS}}{\mathrm{SSY}} = 1 - \frac{\sum_{i=1}^{\mathrm{test}}(y_{\mathrm{exp}} - y_{\mathrm{pred}})^2}{\sum_{i=1}^{\mathrm{test}}(y_{\mathrm{exp}} - y_{\mathrm{train}})^2} \quad (12)$$

In the present contribution, we have performed leave-12-out (L12O) and leave-18-out (L18O) cross-validations. A group of 12 and 18 compounds was randomly selected, respectively from the training set. Then each group was left out and was predicted by the

Table 3
Statistics using LOO-CV and LMO-CV methods for comparing the results of shuffling MARS–ANFIS method with GA-PLS-ANFIS method.

Method	LOO		L12O[a]		L18O[a]	
	Q^2	$RMSE_p$	R^2[b]	$RMSE_p$	R^2	$RMSE_p$
Shuffling MARS–ANFIS	0.892	0.331	0.884	0.359	0.870	0.380
GA-PLS-ANFIS[c]	0.813	0.446	0.787	0.489	0.785	0.494

[a] Calculation of R^2_{LMO} was based on 1000 random selections of groups of 12 and 18 samples.
[b] All R^2 are adjusted coefficient regression.
[c] Selected variables: X3sol, TPC, RNCG and AROM.

Table 4
Mean values of R^2_p and Q^2_{LOO} after performing 100 Y-randomization tests.

Method	Mean of R^2_p	Mean of Q^2_{LOO}
Shuffling MARS–ANFIS	0.185	0.096
GA-PLS-ANFIS	0.236	0.143

model developed from the remaining observations. This procedure was carried out 1000 times. Table 3 shows the results for LOO and LMO cross-validations. High values for Q^2_{LOO} and R^2 indicate the consistency of the developed model. In order to assess the robustness of the shuffling MARS–ANFIS, the *Y*-randomization test was applied in this contribution. The dependent variable vector pIC_{50} was randomly shuffled and a new QSAR model was developed using the original variable matrix. The new QSAR model is expected to show a low value for R^2_p and Q^2_{LOO}. One hundred random shuffles of the *y* vector were performed for which the results are shown in Table 4. The poor values for the mean of $R^2{}_p$ and Q^2_{LOO} indicate that the good results of the shuffling MARS–ANFIS model are not due to a chance correlation or structural dependency of the training set.

4.3. Comparison of shuffling MARS–ANFIS with GA-PLS-ANFIS

For further investigation, GA-PLS technique is also used to select the most important descriptors in the present work. The theories behind this algorithm are discussed elsewhere [27]. To find the best model, GA-PLS were run many times with different settings of initial populations. The best models of GA-PLS with best fitness were selected. Fig. 6 shows the result of GA-PLS variable selection after 3000 runs. This figure shows the most important descriptors are X3sol, TPC, RNCG and AROM (aromaticity). The selected descriptors appeared in GA-PLS model were used in developing ANFIS model to predict the value of pIC_{50}. The results of $Q^2{}_{LOO}$, $R^2{}_{LMO}$ and $RMSE_p$ for LOO, L12O and L18O in GA-PLS-ANFIS model are summarized in Table 3. This table shows that the best model also has four variables for GA-PLS technique. The poor values for the mean of adjusted R^2_p and Q^2_{LOO} in Table 4 confirm that the good results of the GA-PLS-ANFIS model are not due to a chance correlation and the developed model is reliable.

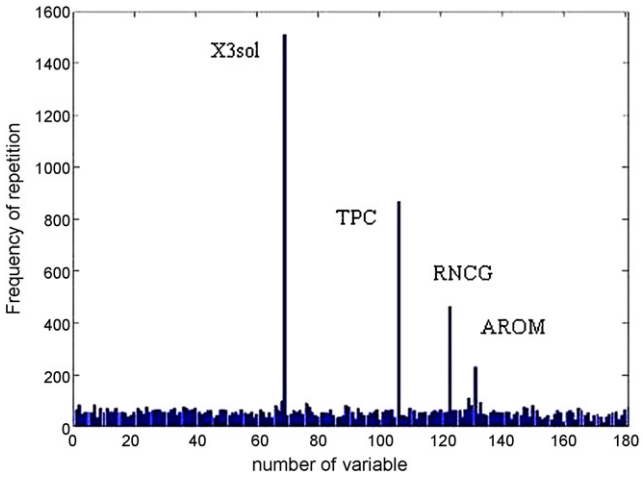

Fig. 6. Selected variables using GA-PLS method after 3000 runs.

It is clear from Table 3 that the results of LOO, L12O and L18O for the shuffling MARS–ANFIS model are superior compared with those of the GA-PLS-ANFIS. However, the shuffling MARS–ANFIS model has 6 descriptors and the GA-PLS-ANFIS model has 4 descriptors, but the adjusted R^2 is relatively independent from the number of variables. It is obvious that the RMSE of both LOO and LMO has been reduced about 50% using shuffling MARS–ANFIS.

4.4. Descriptors appeared in QSAR model

The most repeated variable in the shuffling MARS–ANFIS model is X3sol which is among salvation connectivity indices. These molecular descriptors are defined to model salvation entropy and describe dispersion interactions in solution.

The next important variable selected by the shuffling MARS–ANFIS model is relative negative charge (RNCG). This descriptor is the partial charge of the most negative atom divided by the total negative charge and is defined by the following equation.

$$\mathrm{RNCG} = \frac{Q^-_{max}}{Q^-_{total}} \quad (13)$$

Different hetero atoms such as nitrogen, oxygen and halogen affect Q_{total} and Q_{max} dramatically. Also the presence of donor–acceptor atoms for H bond influences the value of both Q_{total} and Q_{max}. Therefore, the presence of these functions is important in inhibitor–isozyme interaction.

It is shown that another important factor in inhibition mechanism is Length to breadth ratio (L/Bw) of the inhibitor [28]. Length to breadth ratio is defined as the ratio of the longest (L) to the shortest (B) side of the rectangle that envelopes the molecular structure and at the same time maximizes the L/B ratio. This shape parameter not only accounts for the distance between extreme atoms along the principle axis but also for the distribution of all atoms around the molecule center.

The parameter TPS is the total path count of the H-depleted molecular structure and is a useful quantitative measure of molecular complexity. The TPS parameter for molecules with simple structures is smaller than those calculated for molecules with various branching in their structures [23].

The parameter HOMA is harmonic oscillator model of aromacity index and is among resonance indices. The resonance indices are theoretical quantities to explain the stability of benzene and predicting the degree of delocalization of conjugated systems [23].

The last parameter which has been used for modeling and has acceptable frequency of repetition in shuffling-MARS approach is the average molecular weight (AMW). This parameter is calculated by dividing the molecular weight by the number of atoms in the considered molecule. This parameter is a simple molecular descriptor which encodes information on elemental composition of the molecule.

5. Conclusions

A cumbersome step in every QSAR studies is selecting suitable descriptors using a feature selection method. This is more serious when the data set under study is diverse or the mechanism of the process is complex. The data set considered in this work consisted of drug-like molecules inhibiting SARS and consequently, the mechanism of their action could be complicated. An approach

of shuffling MARS–ANFIS was successfully applied for predicting the inhibitor activity of pyridine N-oxide derivatives against SARS. The reasons behind this success could be: (1) the strength of the shuffling MARS as feature selection technique. It is shown that the six parameters of AMW, X3sol, LBw, RNCG, HOMA and TPC chosen by shuffling MARTS affect significantly the inhibition process of the drug-like molecules. (2) The role of ANFIS as mapping model which has the power of prediction of the inhibition behavior. It is a general framework that combines two technologies, namely neural networks and fuzzy systems; by using fuzzy techniques, both numerical and linguistic knowledge can be combined into a fuzzy rule, which require extensive trails and errors for the optimization of their architecture. The shuffling MARS–ANFIS has been testified to be an effective method for variable selection and developing model by using the cross-validation techniques of leave-one-out, leave-multiple-out and also *Y*-randomization. Comparing the results of GA-PLS-ANFIS with those for shuffling MARS–ANFIS reveals that the latter model selects the best variables to predict the inhibition action of pyridine N-oxide derivatives. The appearance of the above-mentioned parameters in the model indicates that type of the atoms, size of the molecule, complexity of the compound, aromacity and elemental composition of the molecule are playing roles in the mechanism of inhibition.

References

[1] N. Lee, D. Hui, A. Wu, P. Chan, P. Cameron, G.M. Joynt, J. Sung. N. Engl. J. Med. 348 (2003) 1986–1994.
[2] S.M. Poutanen, D.E. Low, B. Henry, S. Finkelstein, D. Rose, K. Green, R. Tellier, R.C. Brunham, A.J. McGeer, N. Engl. J. Med. 348 (2003) 1995–2005.
[3] U. Bacha, J. Barrila, A. Velazquez-Campoy, S.A. Leavitt, E. Freire, Bioche 43 (2004) 4906–4912.
[4] A.J. Dooley, N. Shindo, B. Taggart, J.G. Park, Y.P. Pang, Bioorg. Med. Chem. Lett. 16 (2006) 830–833.
[5] J. Balzarini, E. Keyaerts, L. Vijgen, F. Vandermeer, J. Antimicrob. Chemother. 55 (2006) 472–481.
[6] H. González-Díaz, F. Prado-Prado, F.M. Ubeira, Curr. Top. Med. Chem. 8 (2008) 1676–1690.
[7] H. González-Díaz, Y. González-Díaz, L. Santana, F.M. Ubeira, E. Uriarte, Proteomics 8 (2008) 750–778.
[8] H. González-Díaz, S. Vilar, L. Santana, E. Uriarte, Curr. Top. Med. Chem. 7 (2007) 1015–1029.
[9] J.F. Wang, D.Q. Wei, K.C. Chou, Curr. Top. Med. Chem. 8 (2008) 1656–1665.
[10] S. Vilar, G. Cozza, S. Moro, Curr. Top. Med. Chem. 8 (2008) 1555–1572.
[11] A.M. Helguera, R.D. Combes, M.P. González, M.N. Cordeiro, Curr. Top. Med. Chem. 8 (2008) 1628–1655.
[12] M.P. González, C. Terán, L. Saíz-Urra, M. Teijeira, Curr. Top. Med. Chem. 8 (2008) 1606–1627.
[13] T. Ghafourian, M.T.D. Cronin, SAR. QSAR. Environ. Res. 16 (2005) 171–190.
[14] Y. Zhou, H. Leung, J. Syst. Software 80 (2007) 1349–1361.
[15] E. Deconinck, D. Coomansb, Y. Vander Heyden, J. Pharm. Biomed. Anal. 43 (2007) 119–130.
[16] Q.S. Xu, F. Daeyaert, P.J. Lewi, D.L. Massart, Chem. Int. Lab. Syst. 82 (2006) 24–30.
[17] E. Deconinck, Q.S. Xu, R. Put, D. Coomansb, D.L. Massart, Y. Vander Heyden, J. Pharm. Biomed. Anal. 39 (2005) 1021–1030.
[18] W. Shi, Q. Shen, W. Kong, B. Ye, Eur. J. Med. Chem. 42 (2007) 81–86.
[19] M. Sugeno, G.T. Kang, Fuzzy Sets Syst. 28 (1988) 15–24.
[20] Hyperchem, Molecular Modeling System, Hyper Cube, Inc. and Auto Desk, Inc., 1993, Developed by Hyper Cube, Inc.
[21] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Software Dragon: Calculation of Molecular Descriptors, Department of Environmental Sciences, University of Milano-Bicocca, and Talete, srl. http://disat.unimib.it/chm/Dragon.htm, Milan, Italy, 2003.
[22] MATLAB 7.0, http://www.mathworks.com/products/matlab/.
[23] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley/VCH, Weinheim, 2000.
[24] D.W. Osten, J. Chemom. 2 (1998) 39–48.
[25] S. Wold, Quant. Struc. Act. Relat. 10 (1991) 191–193.
[26] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269–276.
[27] M. Jalali-Heravi, A. Kyani, Eur. J. Med. Chem. 42 (2007) 649–659.
[28] C.T. Supuran, A. popescu, M. Ilisiu, Eur. J. Med. Chem. 31 (1996) 439–447.