Available online at www.sciencedirect.com

ScienceDirect

European Journal of Medicinal Chemistry 42 (2007) 649–659

EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY

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

Original article

Application of genetic algorithm-kernel partial least square as a novel nonlinear feature selection method: Activity of carbonic anhydrase II inhibitors

Mehdi Jalali-Heravi*, Anahita Kyani

Department of Chemistry, Sharif University of Technology, Azadi Avenue, P.O. Box 11365-9516, Tehran, Iran

Received 31 August 2006; received in revised form 2 December 2006; accepted 5 December 2006
Available online 12 January 2007

Abstract

This paper introduces the genetic algorithm-kernel partial least square (GA-KPLS), as a novel nonlinear feature selection method. This technique combines genetic algorithms (GAs) as powerful optimization methods with KPLS as a robust nonlinear statistical method for variable selection. This feature selection method is combined with artificial neural network to develop a nonlinear QSAR model for predicting activities of a series of substituted aromatic sulfonamides as carbonic anhydrase II (CA II) inhibitors. Eight simple one- and two-dimensional descriptors were selected by GA-KPLS and considered as inputs for developing artificial neural networks (ANNs). These parameters represent the role of acceptor-donor pair, hydrogen bonding, hydrosolubility and lipophilicity of the active sites and also the size of the inhibitors on inhibitor–isozyme interaction. The accuracy of 8-4-1 networks was illustrated by validation techniques of leave-one-out (LOO) and leave-multiple-out (LMO) cross-validations and *Y*-randomization. Superiority of this method (GA-KPLS-ANN) over the linear one (MLR) in a previous work and also the GA-PLS-ANN in which a linear feature selection method has been used indicates that the GA-KPLS approach is a powerful method for the variable selection in nonlinear systems.

Keywords: Carbonic anhydrase II inhibitors; Genetic algorithm-kernel partial least square; QSAR; Artificial neural networks

1. Introduction

Since Hansch's seminal work on quantitative structure–activity relationships (QSAR) analysis [1,2], many different QSAR methods including 2D and 3D QSAR approaches have been developed. These models have been used to study action mechanisms of chemical–biological interactions in modern drug discovery [3]. However, a common problem for these models is choosing an optimal set of structural descriptors. A large number of structural descriptors can be generated by existing softwares [4,5], but choosing adequate descriptors for QSAR/QSPR studies is difficult and challenging. To overcome this problem a powerful variable selection technique is needed. This technique should be able to choose the suitable parameters and discard the others by considering a predefined criterion. In most conventional QSAR methods, a linear relationship is assumed, which in most situations is valid for relatively small and congeneric sets of compounds. This assumption may not be true for all cases especially for diverse sets of data. Frequent optimization search algorithms such as stepwise-forward and stepwise-backward MLR depend on an assumed linear relationship between the dependent variable and one or more descriptors, while there may be a nonlinear relationship between the physical–chemical properties of compounds and their structural descriptors. There are some published papers suggesting that genetic algorithms (GAs) are useful in data analysis [6–9]. Rogers and Hopfinger applied this method in QSAR analysis for the first time and

* Corresponding author. Tel.: +98 21 6616 5315; fax: +98 21 6601 2983.
E-mail address: jalali@sharif.edu (M. Jalali-Heravi).

0223-5234/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2006.12.020

proved that GAs are very powerful tools with many merits that other methods did not have [10]. GAs are able to reduce the noise introduced by the descriptors that do not influence the studied structure–property relationship and improves the robustness and generalization ability of the constructed model. GAs are developed to mimic some of the processes observed in natural evolution, which are an efficient strategy to search for the global optima of the solutions. Genetic algorithms have also been applied as feature selection method in regression analysis [11,12]. Moreover, an approach combining GAs with PLS (GA-PLS) is used in some cases for variable selection in QSAR and QSPR studies [13,14]. However, all of these methods are used for feature selection in linear systems. On the other hand, some researchers have used genetic algorithm neural network (GNN) as a nonlinear feature selection, which uses neural network as a nonlinear fitness function [15,16]. In the case of GNN the ANN parameters together with GAs parameters should be optimized. This means that the optimization of the genetic algorithm-based neural network models is a complex procedure [17].

Recently, a new nonlinear PLS technique called kernel partial least square (KPLS) was developed for dealing with the problem of nonlinearity [18]. KPLS differs from the nonlinear PLS algorithms, in that the original output data are nonlinearly transformed into a higher dimensional feature space and the linear PLS modeling is performed in this space [19]. Compared to other nonlinear approaches, the main advantage of KPLS is that it avoids nonlinear optimization by utilizing the kernel function corresponding to the inner product in the feature space. Therefore, this method can be considered as a fast and effective method for nonlinear systems. Two abilities make KPLS superior to other nonlinear methods: first, its simplicity which is same as the standard PLS because of using only linear algebra and second, its application in a wide range of nonlinearities because of having different kinds of kernels. Based on these merits, compared to the linear PLS, KPLS is able to have a better performance in regressing and classifying data in nonlinear systems [19].

The main aim of this paper was to use the advantages of KPLS by introducing it as a part of a novel nonlinear feature selection method for nonlinear systems. This method is based on combining the KPLS as a nonlinear fitness function with the GAs as a powerful optimization method. We refer to this technique as GA-KPLS. By using this method, we expect to improve both the predictive ability and simplicity of the model and also to obtain the best variables for the nonlinear systems. Therefore, a new model (GA-KPLS-ANN) was developed by combining GA-KPLS with the neural networks. In order to ascertain the ability of this model it was used in predicting the biological activity of a set of aromatic sulfonamides as carbonic anhydrase II inhibitors [20,21].

Sulfonamides represent an important class of biologically active compounds. The antibacterial sulfonamides continue to play an important role in chemotherapy, alone or in combination with other drugs. The sulfonamides that inhibit the zinc enzyme carbonic anhydrase (CA, EC 4. 2. 1. 1) possess many applications as diuretic, antiglaucoma or antiepileptic drugs [22]. The aromatic/heterocyclic sulfonamides act as carbonic anhydrase inhibitors and other types of derivatives show hypoglycemic activity, anticancer properties or may act as inhibitors of the aspartic HIV protease being used for the treatment of AIDS and HIV infection [20].

The data set studied in this work consists of 114 molecules for which the biological activities were reported as $\log \mathrm{IC}_{50}$ values [20,21]. QSAR studies of these compounds have recently been reported but they were restricted to the linear regression models using two- and three-dimensional descriptors [20,21]. In the present work, the best variables among simple zero-, one- or two-dimensional descriptors were selected using GA-KPLS and then a nonlinear ANN model was developed to predict the $\log \mathrm{IC}_{50}$ of these compounds. Our hybrid method (GA-KPLS-ANN) shows a considerable improvement in comparison with both GA-PLS-ANN and recently reported regression models [20,21].

2. Methods

2.1. Genetic algorithms

A detailed description of the genetic algorithms (GAs) can be found in the literature [23,24]. Genetic algorithms are simulated methods based on ideas from Darwin's theory of natural selection and evolution (the struggle for life). In GAs a chromosome (or an individual) can be defined as an enciphered entity of a candidate solution, which is expressed as a set of variables.

GAs consist of the following basic steps: (1) A chromosome is represented by a binary bit string and an initial population of chromosomes is created in a random way; (2) A value for the fitness function (here PLS and KPLS) of each chromosome is evaluated; (3) Based on the values of the fitness functions, the chromosomes of the next generation are produced by selection, crossover and mutation operations.

2.2. Partial least square (PLS)

A detailed theory behind PLS algorithm is given in the literature [25,26]. The basic concept of the PLS is briefly discussed here to establish a background for a better understanding of the KPLS method.

PLS models a linear relationship between a set of input variables (predictors), $x_i \in R^N, i = 1, \ldots, n$ and a set of output variables (responses), $y_i \in R^M, i = 1, \ldots, n$ by means of latent variables [19].

It is assumed that $X(n \times N)$ contains the descriptors that can be used for predicting the activities $Y(n \times M)$. It is well known that PLS decomposes the data matrices X and Y into a two matrices product plus residual in a single process. The matrices E and F contain residuals for X and Y, respectively:

$$X = TP' + E \tag{1}$$

$$Y = UQ' + F \tag{2}$$

where T and U are score matrices and P' and Q' are loading matrices for X and Y, respectively. These two equations can be written as a multiple regression model:

$$Y = XB + G \tag{3}$$

where matrix B contains the PLS regression coefficients [25] and can be calculated as below:

$$B = X^T U \left(T^T X X^T U\right)^{-1} T^T Y \tag{4}$$

Unlike the classical PLS algorithm, the modified PLS algorithm normalizes the latent vectors of t and u rather than the weight vectors of w and c [19]. The modified PLS algorithm is shown in Fig. 1a.

2.3. Kernel partial least square

While PLS can be performed on linear systems, kernel partial least square (KPLS) can map nonlinear data in a higher-dimensional space called feature space (F) in which they can be modeled linearly. KPLS is formulated in this feature space to extend linear PLS to its nonlinear kernel form. In the nonlinear transformation, input variables x_i, $i = 1,\ldots, n$ were mapped into feature space F:

$$x_i \in R^N \rightarrow \Phi(x_i) \in F \tag{5}$$

where it is assumed that $\sum \Phi(x_k) = 0$, and $\Phi(\bullet)$ is a nonlinear transformation function that projects the input vectors from the input space to F. The feature space can have the arbitrary large and even infinite dimension. Φ is the $(n \times S)$ matrix whose ith row is the vector $\Phi(x_i)$ in a S-dimensional feature space F. For the sake of comparison, the algorithms of PLS and KPLS are shown in Fig. 1. It is obvious that KPLS algorithm can be derived directly from the PLS algorithm by modifying the steps of 2 and 3 and using the matrix Φ of transformed input data instead. Through the introduction of the kernel trick, $\Phi(x_i)^T \Phi(x_j) = K(x_i, x_j)$, one can avoid both performing nonlinear mappings and computing dot products in the feature space. $\Phi\Phi^T$ is a $(n \times n)$ kernel Gram matrix K created the cross dot products between all mapped input data points, $i = 1,\ldots, n$.

The deflation in final step is based on rank-one reduction of the K and Y matrices using a new extracted score vector t as given below:

$$K \leftarrow \left(I - tt^T\right) K \left(I - tt^T\right) = K - tt^T K - K tt^T + tt^T K tt^T \tag{6}$$

where I is an n-dimensional identity matrix.

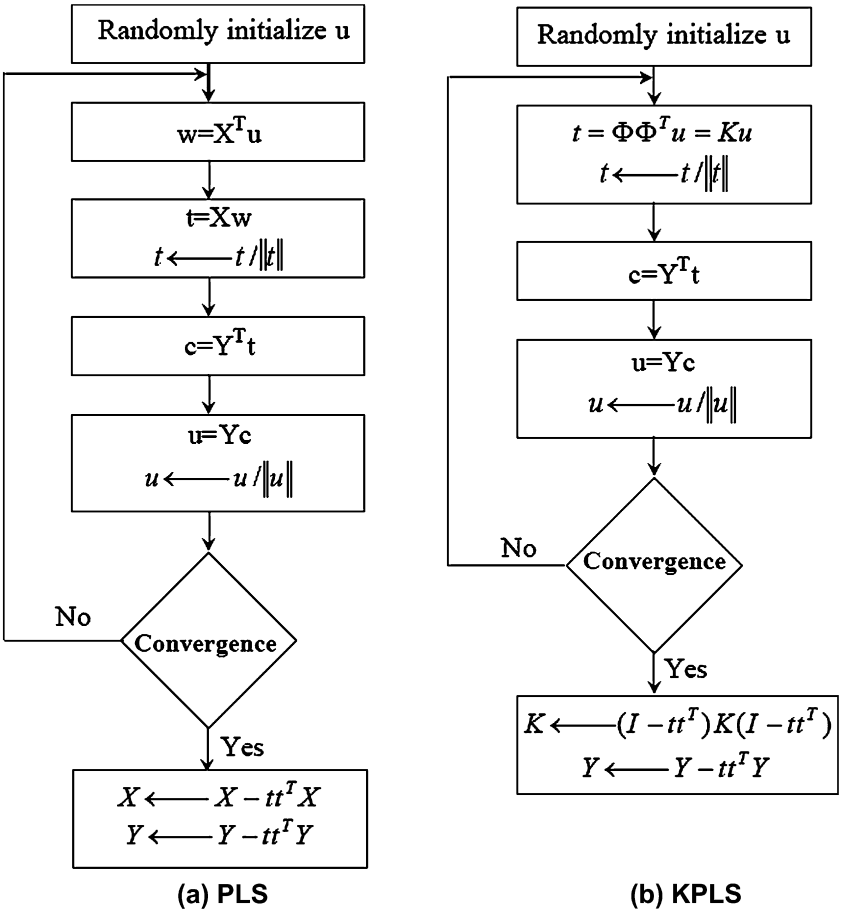

Fig. 1. Comparison of the (a) PLS and (b) KPLS algorithms.

The regression coefficient B in the KPLS algorithm can be obtained from Eq. (7).

$$B = \Phi^T U\left(T^T K U\right)^{-1} T^T Y \tag{7}$$

For a data consisted of n_t dependent variables of test set, the following equations can be used to predict the training and test data, respectively:

$$\widehat{Y} = \Phi B = KU\left(T^T K U\right)^{-1} T^T Y \tag{8}$$

$$\widehat{Y}_t = \Phi_t B = K_t U\left(T^T K U\right)^{-1} T^T Y. \tag{9}$$

Here, Φ_t is the matrix of the mapped test points and K_t is the $(n_t \times n)$ test matrix whose elements are $K_{ij} = K(x_i, x_j)$, where x_i is the ith test vector and x_j is the jth training vector.

Before applying KPLS, mean centering in the higher-dimensional space should be performed by substituting the kernel matrices K and K_t with $\tilde{K}$ and $\tilde{K}_t$:

$$\tilde{K} = \left(I - \frac{1}{n} I_n I_n^T\right) K \left(I - \frac{1}{n} I_n I_n^T\right) \tag{10}$$

$$\tilde{K}_t = \left(K_t - \frac{1}{n} I_{n_t} I_n^T K\right)\left(I - \frac{1}{n} I_n I_n^T\right) \tag{11}$$

where I is an n-dimensional identity matrix, I_n and I_{n_t} represent vectors whose elements are ones, with lengths n and n_t, respectively.

Different kernel functions are available as:

$$\text{Polynomial kernel}: \quad k(x,y) = \langle x,y \rangle^d$$

$$\text{Sigmoid kernel}: \quad k(x,y) = \tanh(\beta_0 \langle x,y \rangle + \beta_1)$$

$$\text{Radial basis kernel}: \quad k(x,y) = \exp\left(-\frac{\|x-y\|^2}{c}\right)$$

where d, β_0, β_1,and c should be predefined by user [19]. Among different types of kernels, radial basis kernel is more common.

3. Experimental

3.1. Data set

A data set consisted of 114 substituted aromatic sulfonamides as carbonic anhydrase II inhibitors was taken from Refs. [20,21]. The structure of these compounds is given in Fig. 2. The inhibition effects expressed as $\log \mathrm{IC}_{50}$ in terms of nanomolar affinity for the investigated carbonic anhydrase isozymes are given in Table 1. Log IC_{50} was used as dependent parameter in developing GA-KPLS-ANN model. In order to evaluate the generated ANN model, we used leave-one-out cross-validation (LOO-CV-ANN). In this algorithm, one compound was left in each step as prediction set and the model was developed using the remaining molecules as training set. The left-out data is then used to perform the prediction. For a further exhaustive testing of the predictive power of the model, in addition to LOO-CV, three different leave-multiple-out cross-validation ANN algorithms (LMO-CV-ANN) were also carried out. Here we performed leave-10-out (L10O), leave-14-out (L14O) and leave-20-out (L20O) cross-validations. A group of 10, 14 and 20 compounds was randomly selected from the training set. Then each group was left out and was predicted by the model developed from the remaining observations. This procedure was carried out 200 times. In order to study the robustness of GA-KPLS-ANN model, we also used Y-randomization test [27]. In this test the $\log \mathrm{IC}_{50}$ is randomly shuffled and a new QSAR model is developed using the original descriptor matrix. The new QSAR model is expected to have low R^2 and Q^2 values.

3.2. Descriptor generation

The numerical descriptors are responsible for encoding important features of the structure of the molecules and can be categorized as three-, two-, one- and zero-dimensional parameters. These parameters can be divided into geometric, topological, functional and constitutional descriptors. In the present work, we intended to select the best descriptors among the simple calculated ones. To reach this goal, we have calculated only zero-, one- and two-dimensional descriptors by using Dragon software [5]. A total of 275 descriptors were calculated for each compound. These descriptors were obtained from a large number of descriptors after removing the parameters, which have more than 10% constant or zero values.

3.3. Genetic algorithm kernel partial least square (GA-KPLS)

In this paper, we used GA-KPLS as a nonlinear feature selection method and compared it with GA-PLS as a linear technique. The genetic algorithms follow Leardi's method [13].

The cross-validation technique was used for evaluating the descriptors selected by GAs in each step. The data set was divided into approximately q equal deletion groups (here q set to be 5). One group is left out as a test set, then the PLS and KPLS models were developed with $(q-1)$ groups as training set. This procedure is repeated until all the objects have been predicted once. In other word a leave-group-out cross-validation has been performed. The n selected descriptors in each chromosome were evaluated by fitness function of PLS and KPLS based on the following equation:

$$\text{Fitness} = \sqrt{\frac{\text{CUMPRESS}}{m-n}} \tag{12}$$

Where CUMPRESS and m are the cumulative predictive sum of square error and the number of compounds in data set, respectively.

In this paper a radial basis kernel function, $k(x,y) = \exp(\|x-y\|^2/c)$, was selected as the kernel function with

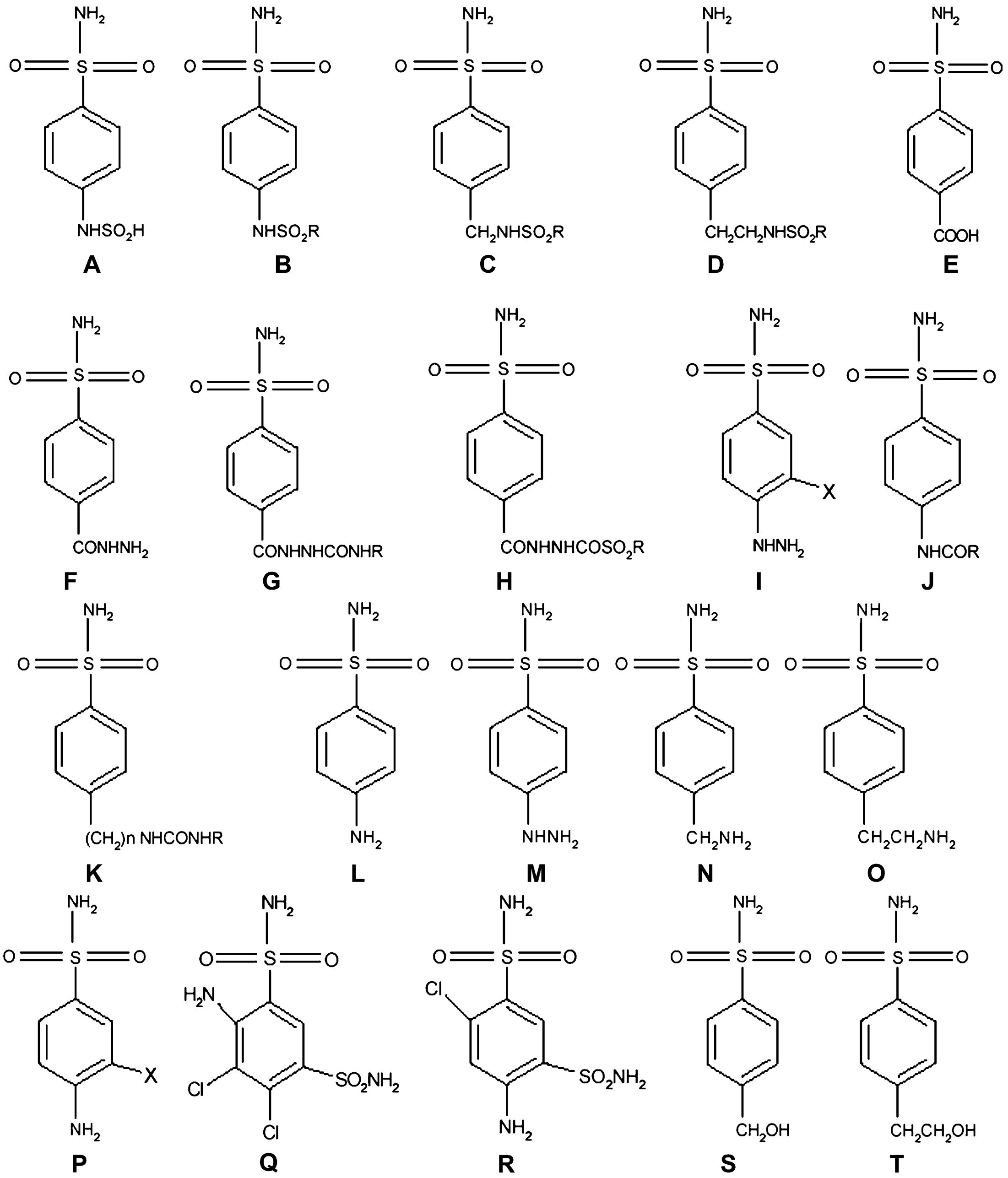

Fig. 2. Structures of sulfonamides as CA II inhibitors studied in this work.

$c = rm\sigma^2$, where r is a constant that can be determined by considering the process to be predicted (here r set to be 1), m is the dimension of the input space and σ^2 is the variance of the data [19]. It means that the value of c depends on the system under study. The GA-KPLS and GA-PLS programs were written in m-file and were compiled using a MATLAB 6.5 compiler.

3.4. Artificial neural network

A detailed description of the theory behind a neural network has been adequately described elsewhere [28–31]. The feed forward back propagation artificial neural network (BP-ANN) and the program for ANN evaluation based on leave-one-out and leave-multiple-out cross-validation was written in MATLAB 6.5. The output layer represents $\log \mathrm{IC}_{50}$ of the CA II inhibitors. The descriptors were selected using GA-PLS and GA-KPLS and were considered as inputs for ANN. In this investigation the sigmoid function was used as transfer function. The initial weights of the network were randomly selected from a uniform distribution that ranged between −0.3 and +0.3. The initial values of biases were set to be 1. These values were optimized based on the Levenberg–Marquardt optimization as a fast function in updating the weights of the network [32,33].

Before training, the inputs were normalized between −2 and 2 and output between 0.2 and 0.8. The network parameters such as the number of nodes in hidden layer, learning rate and

Table 1
Experimental and GA-KPLS-ANN calculated values of log IC_{50} of CA II inhibitors studied in this work

Drug[a]	Type[b]	R	Experimental log IC_{50}	Calculated log IC_{50}
1[c]	A	Me	1.806	1.863
2[c]	A	$PhCH_2$	1.230	1.408
3[c]	A	Me_2N	1.079	1.217
4[e]	A	Ph	1.690	1.671
5[d]	A	4-Me–C_6H_4	1.623	1.667
6[e]	A	2,4,6-Me_3–C_6H_2	1.740	1.745
7[e]	A	4-F–C_6H_4	0.954	1.470
8[c]	A	4-Cl–C_6H_4	0.954	1.070
9[c]	A	4-Br–C_6H_4	0.903	0.665
10[c]	A	4-MeO–C_6H_4	1.477	1.246
11[c]	A	2-HOOC–C_6H_4	0.845	0.985
12[c]	A	2-O_2N–C_6H_4	1.041	0.980
13[c]	A	3-O_2N–C_6H_4	1.000	1.022
14[c]	A	4-O_2N–C_6H_4	1.000	1.092
15[c]	A	4-AcNH–C_6H_4	1.914	1.938
16[c]	A	3-H_2N–C_6H_4	1.634	1.622
17[c]	A	4-H_2N–C_6H_4	1.663	1.607
18[c]	A	4-Cl-3-O_2N–C_6H_3	1.079	1.234
19[d]	B	Me	1.898	1.858
20[c]	B	$PhCH_2$	1.447	1.410
21[c]	B	Me_2N	1.491	1.203
22[c]	B	Ph	1.887	1.928
23[c]	B	4-Me–C_6H_4	1.881	1.678
24[c]	B	2,4,6-Me_3–C_6H_2	1.826	1.789
25[c]	B	4-F–C_6H_4	1.000	1.029
26[c]	B	4-Cl–C_6H_4	0.954	1.040
27[c]	B	4-Br–C_6H_4	0.954	1.020
28[c]	B	4-MeO–C_6H_4	1.602	1.269
29[c]	B	2-HOOC–C_6H_4	0.845	1.010
30[c]	B	2-O_2N–C_6H_4	1.000	0.871
31[c]	B	3-O_2N–C_6H_4	0.954	1.052
32[c]	B	4-O_2N–C_6H_4	1.000	0.830
33[c]	B	4-AcNH–C_6H_4	2.004	1.974
34[c]	B	3-H_2N–C_6H_4	1.672	1.676
35[c]	B	4-H_2N–C_6H_4	1.699	1.660
36[c]	B	4-Cl-3-O_2N–C_6H_3	1.176	1.254
37[c]	C	Me	1.708	1.876
38[d]	C	$PhCH_2$	1.000	1.002
39[c]	C	Me_2N	0.903	0.869
40[c]	C	Ph	1.602	1.407
41[c]	C	4-Me–C_6H_4	1.491	1.514
42[c]	C	2,4,6-Me_3–C_6H_2	1.716	1.661
43[c]	C	4-F–C_6H_4	0.845	1.017
44[e]	C	4-Cl–C_6H_4	0.845	0.807
45[c]	C	4-Br–C_6H_4	0.699	0.685
46[c]	C	4-MeO–C_6H_4	1.255	0.995
47[c]	C	2-HOOC–C_6H_4	0.778	0.985
48[e]	C	2-O_2N–C_6H_4	1.000	1.082
49[c]	C	3-O_2N–C_6H_4	0.954	0.945
50[c]	C	4-O_2N–C_6H_4	0.778	0.672
51[e]	C	4-AcNH–C_6H_4	1.887	1.675
52[c]	C	3-H_2N–C_6H_4	1.544	1.564
53[e]	C	4-H_2N–C_6H_4	1.519	1.553
54[d]	C	4-Cl-3-O_2N–C_6H_3	1.041	1.047
55[c]	D	Me	1.602	1.413
56[c]	D	$PhCH_2$	0.903	0.792
57[d]	D	Me_2N	0.778	0.810
58[c]	D	Ph	1.447	1.403
59[c]	D	4-Me–C_6H_4	1.431	1.499
60[c]	D	2,4,6-Me_3–C_6H_2	1.602	1.653
61[c]	D	4-F–C_6H_4	0.699	0.763
62[c]	D	4-Cl–C_6H_4	0.699	0.938
63[c]	D	4-Br–C_6H_4	0.477	0.633
64[c]	D	4-MeO–C_6H_4	1.176	1.499
65[d]	D	2-HOOC–C_6H_4	0.699	0.984
66[e]	D	2-O_2N–C_6H_4	0.778	0.799
67[d]	D	3-O_2N–C_6H_4	0.903	0.981
68[e]	D	4-O_2N–C_6H_4	0.602	0.786
69[c]	D	4-AcNH–C_6H_4	1.875	1.869
70[c]	D	3-H_2N–C_6H_4	1.477	1.471
71[d]	D	4-H_2N–C_6H_4	1.477	1.479
72[d]	D	4-Cl-3-O_2N–C_6H_3	0.954	0.819
73[c]	E	–	2.412	2.169
74[d]	F	–	2.093	2.149
75[c]	G	3,4-$Cl_2C_6H_3$	1.114	0.783
76[c]	G	4-Ac–C_6H_4	1.176	0.894
77[e]	G	4-EtOOC–C_6H_4	0.954	1.008
78[c]	G	4-Br–C_6H_4	0.863	0.859
79[e]	G	4-Ph–C_6H_4	1.041	0.897
80[d]	G	4-PhO–C_6H_4	1.255	0.933
81[d]	G	4-$PhCH_2$–C_6H_4	1.176	0.976
82[c]	H	Ph	1.826	1.367
83[c]	H	2-Me–C_6H_4	1.732	1.336
84[c]	H	4-Me–C_6H_4	0.991	1.327
85[c]	H	4-F–C_6H_4	0.978	1.065
86[d]	H	4-Cl–C_6H_4	0.959	1.202
87[d]	I	F	1.708	1.407
88[c]	I	Cl	1.881	1.702
89[e]	J	Me	2.391	2.382
90[c]	J	CF_3	2.124	1.828
91[c]	J	Et	2.366	2.379
92[d]	J	*n*-Pr	2.356	2.376
93[d]	J	*i*-Pr	2.412	2.407
94[c]	J	*n*-Bu	2.330	2.373
95[c]	J	*t*-Bu	2.362	2.348
96[d]	J	*n*-C_5H_{11}	1.799	2.310
97[d]	J	Ph	1.568	1.987
98[e]	J	C_6F_5	1.230	1.191
99[c]	K	Ph, $n=0$	2.380	1.983
100[c]	K	Ph, $n=1$	2.021	1.748
101[c]	K	Ph, $n=2$	1.875	1.670
102[e]	K	3,4-$Cl_2C_6H_3$, $n=2$	1.114	0.597
103[c]	L	–	2.477	2.278
104[c]	M	–	2.505	2.171
105[e]	N	–	2.230	2.250
106[c]	O	–	2.204	2.239
107[c]	P	F	1.778	1.777
108[c]	P	Cl	2.041	2.130
109[e]	P	Br	1.602	1.602
110[e]	P	I	1.845	1.638
111[e]	Q	–	1.447	1.371
112[d]	R	–	1.875	1.746
113[c]	S	–	2.097	2.134
114[e]	T	–	2.041	1.700

[a] Molecules defined by c, d and e refer to the training, test and validation sets, respectively.

[b] The structures of sulfonamides are given in Fig. 2.

momentum were optimized based on obtaining the minimum standard error of training (SE_T) and standard error of prediction (SE_P) [31]. We used leave-one-out cross-validation (LOO-CV) method for the evaluation of the ANN model and compared the statistics of new nonlinear hybrid method of GA-KPLS-ANN with GA-PLS-ANN and previously reported linear regression models [20,21]. Also we used leave-multiple-out cross-validation (LMO-CV) for comparing the consistency of the generated model and also comparing its results with GA-PLS-ANN for which the variable selection method is linear.

4. Results and discussion

4.1. Choosing a strategy for developing a model

One of the challenging parts in developing models is choosing suitable parameters encoding different aspects of the molecular structure. A large number of structural descriptors can be calculated using existing softwares such CODESSA and Dragon [4,5]. However, nowadays the main problem is choosing the most adequate and interpretable parameters needed for developing the models among a large number of them.

In developing a model, three strategies can be considered: (1) linear feature selection/linear modeling; (2) linear feature selection/nonlinear modeling and (3) nonlinear feature selection/nonlinear modeling. The strategy (1) is only applicable for modeling the processes with linear characteristics. The strategies (2) and (3) can be applied for the nonlinear systems, but a question arises that which one can have a better performance. The main objective of the present work was looking for a proper answer to this question. Several researchers have shown that artificial neural networks are able to model nonlinear systems [34–36]. Therefore, one expects to get a reasonable model by combining a suitable linear/nonlinear feature selection method with the neural networks. In the present work, we have shown that the strategy (3) is the most successful one by introducing a new nonlinear feature selection method. We have combined for the first time GAs as the optimization method with KPLS as a robust nonlinear statistical method to generate nonlinear feature selection method of GA-KPLS. Also, for the sake of comparison, the results of GA-PLS as linear feature selection technique were compared with the results of this method.

4.2. Sulfonamides as a data set

A data set consisted of log IC_{50} for 114 substituted aromatic sulfonamides as carbonic anhydrase II (CA II) (Table 1 and Fig. 2) was chosen to assess the performance of these methods. Two reasons were behind choosing this data set. First, sulfonamides represent an important class of biologically active compounds. These compounds inhibit the zinc enzyme carbonic anhydrase. Also, the hypoglycemic sulfonamides were extensively used in the treatment of some forms of diabetes and antithyroid drugs [20]. Second reason was a recent paper published by Supuran et al. [20]. They have developed a linear QSAR for a total of 47 *para*-substituted aromatic sulfonamides and their result was valuable from point of comparing multi-parametric regressions with our model.

4.3. The development of feature selection methods

Table 2 shows the specifications of the GAs that were same for both GA-PLS and GA-KPLS techniques. These parameters were optimized in a way that gives the lowest fitness error. The difference between these two methods was based on fitness function, linear for PLS and nonlinear for KPLS. We performed both GA-PLS and GA-KPLS on matrices X (114×275) and Y (114×1). Then by comparing the fitness values of chromosomes the best model was chosen. Eight and ten descriptors were appeared in the best models for GA-KPLS and GA-PLS, respectively. The definitions and corresponding notations, based on Dragon software [5], of the selected descriptors of both feature selection methods are shown in Table 3.

4.4. The power of GA-KPLS-ANN model

Next step of this work was building a model to predict the log IC_{50} of CA II inhibitors. Artificial neural network was used as feature mapping method to build the nonlinear model. At first 20% of the data set was randomly chosen as prediction set in a way to be a good representative of other sulfonamides in the training set. It means that these compounds were selected randomly from all inhibitors in data set after sorting them based on log IC_{50}.

The ten and eight descriptors chosen by GA-PLS and GA-KPLS feature selection methods were considered as inputs for developing BP-ANNs. The network parameters were optimized for both GA-PLS-ANN and GA-KPLS-ANN models using the procedure given elsewhere [31]. Table 4 shows the architecture and specifications of the optimized ANNs. In order to evaluate the robustness of the networks leave-one-out cross-validation (LOO-CV) method was used after the optimization of networks parameters. In LOO-CV, each object of the data set is taken away one at a time and the log IC_{50} of the deleted object is predicted from the model developed using the remaining molecules. Q^2, which is a measure of the model fit to the cross-validation set, can be calculated as:

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

$$\mathrm{RMSE} = \sqrt{\frac{\mathrm{PRESS}}{n}}$$

Table 2
Specifications of GAs for both GA-PLS and GA-KPLS techniques

Population size	256
No. of generation	100
Crossover type	Single
Mutation rate	0.001
Crossover frequency	0.01
q Parameter for CV	5

Table 3
Definitions and notations of descriptors for GA-PLS and GA-KPLS

GA-PLS		GA-KPLS	
Descriptor	Notation[a]	Descriptor	Notation[a]
No. of CH_3R/CH_4	C-001	No. of CH_3R/CH_4	C-001
No. of R−CH−R	C-024	No. of R−CX−R	C-026
Molecular weight	MW	No. of H attached to C1(sp^3)/C0 (sp^2)	H-047
No. of RCO−N</>N−X = X	N-072	No. of oxygen double bonds	O-058
No. of donor atoms for H bond with N and O	nHDon	No. of substituted aromatic C (sp^2)	nCaR
No. of sulfonamides	nSO_2N	No. of donor atoms for H bond with N and O	nHDon
Wiener-type index from mass weighted distance matrix	Whetm	Balaban-type index from Z weighted distance matrix (Barysz index)	JhetZ
E-state topological parameter	TIE	Eigenvalue sum from Z weighted distance matrix (Barysz index)	SEigZ
Balaban-type index from Z weighted distance matrix (Barysz index)	JhetZ		
Connectivity index chi-4	X4		

[a] The notations are based on Dragon software.

where PRESS is predictive sum of squares of the residuals and SSY is the sum of squares of the response variables corrected for the mean. The statistical parameters obtained by LOO-CV for GA-PLS-ANN, GA-KPLS-ANN and the linear QSAR models are compared in Table 5. It can be seen from this table that statistical results for GA-KPLS-ANN model are superior to other methods. Inspection of the results of the table reveals a higher R^2 and Q^2 values and lower RMSEs for the GA-KPLS-ANN method compared with their counterparts for GA-PLS-ANN model. Also, a lower number of variables have appeared in the former model. This clearly shows the strength of GA-KPLS as a nonlinear feature selection method.

The accuracy of cross-validation results is extensively accepted in the literature considering the Q^2 value. In this sense, a high value of the statistical characteristic ($Q^2 > 0.5$) is considered as proof of the high predictive ability of the model [37]. However, several authors suggest that a high value of Q^2 appears to be a necessary but not sufficient condition for a model to have a high predictive power and consider that the predictive ability of a QSAR model can only be estimated using a sufficiently large collection of compounds that was not used for building the model [38]. We believe that applying only LOO-CV is not sufficient to evaluate the predictive ability of a model. Thus we employed a two-step validation protocol. The model is first validated internally using the data set. The data set was divided into training (calibration), test and validation sets after sorting based on the log IC_{50} values. The training set consisted of 76 molecules and the test and validation sets, each consisted of 19 molecules. The test and the validation sets adequately represent the training set. The training set was used for model generation. The test set was applied to deal with overfitting of the network, whereas validation set in which its molecules have no role in model building was used for the evaluation of the predictive ability of the network. Correlation (R^2) values of 0.891, 0.850, and 0.845 were obtained for the training, test and validation sets, respectively. Also, RMSEs of 0.176, 0.214 and 0.205, respectively were obtained for these sets. The GA-KPLS-ANN calculated values of log IC_{50} for the training, test and validation sets are presented in Table 1.

Table 4
Architecture and specifications for hybrid methods of GA-PLS-ANN and GA-PLS-ANN

	GA-PLS-ANN	GA-KPLS-ANN
No. of nodes in the input layer	10	8
No. of nodes in the hidden layer	6	4
No. of nodes in output layer	1	1
Learning rate	0.2	0.1
Momentum	0.5	0.1
Transfer function	Sigmoid	Sigmoid
Update weight function	Levenberg-Marquardt μ=0.1, λ=100	Levenberg-Marquardt μ=0.9, λ=100

To further check the reliability of the proposed model we have also used LMO-CV. Based on this technique, a number of modified data sets were created by deleting a small group of objects in each step (here 10, 14 and 20 objects) and then the model was evaluated by measuring its accuracy in predicting the responses of the deleted group (the ones that have not been utilized in the development of the model). The results of L10O, L14O and L20O for two methods GA-PLS-ANN and GA-KPLS-ANN are reported in Table 6. The consistency in the statistics of R^2, RMSE, R^2_{cv} and $RMSE_{cv}$ for different data sets of L10O, L14O and L20O reveals the stability and robustness of both models GA-KPLS-ANN and GA-PLS-ANN. However, the superiority of GA-KPLS-ANN over that of GA-PLS-ANN is obvious from Table 6. This is due to higher values of R^2 and R^2_{cv} and lower values of RMSEs for the former model compared with those of the latter one.

The GA-KPLS-ANN model was further evaluated by applying the *Y*-randomization. Several random shuffles of the *Y*(log IC_{50}) were chosen and the feature selection and modeling process were performed for all cases. The results are shown in Table 7. The low values for R^2 and Q^2 show that the good statistical results in GA-KPLS-ANN model are not due to a chance correlation or structural dependency of the training set [39].

In order to further examine the predictive ability of GA-KPLS-ANN model, we decided to use the same data set used in previous works. Therefore, we chose 47 sulfonamides as data set 1 from Ref. [20] and 72 sulfonamides as data set 2 from Ref. [21]. Four descriptors were selected using GA-KPLS for developing an ANN model for the data set 1 (see Table 5). It can be seen from Table 5 that although the number

Table 5
Statistical parameters for leave one out comparison of linear and nonlinear methods

Model	No. compounds	No. variables	R^2	RMSE	Q^2	$RMSE_{cv}$	Ref.
GA-PLS-ANN	114	10	0.851	0.204	0.713	0.277	Present work
GA-KPLS-ANN	114	8	0.899	0.163	0.800	0.229	Present work
	72[a]	8	0.951	0.081	0.852	0.162	Present work
	47[b]	4	0.913	0.088	0.822	0.101	Present work
Linear regression	72[a]	8	0.682	–	0.594	0.240	21
Linear regression	47[b]	4	0.728	0.263	0.668	0.291	20

[a] This data set referred as set 2 in the text.
[b] This data set referred as set 1 in the text.

of variables are same for GA-KPLS-ANN and linear regression model, but the former model shows a much higher R^2 value (0.913 vs 0.728) and much lower RMSE (0.088 vs 0.263) compared to the latter one. Similar improvements can be seen for the set 2. It should be mentioned that the descriptors used in regression model for the data set 1 were as simple as the descriptors selected by GA-KPLS from the point of dimensionality. However, the variables selected by GA-KPLS for the data set 2 are simple one- and two-dimensional compared with the three-dimensional parameters of previous work [21]. This superiority shows the power of the GA-KPLS as a feature selection method and also demonstrates the nonlinear characteristic of inhibition mechanism of sulfonamides on CA II isozyme.

4.5. The interpretation of the inhibition mechanism

The GA-KPLS selected descriptors are Balaban type index (JhetZ), eigenvalue distance matrix (SEigZ) and also the number of CH_3R (C-001), R–CX–R (C-026), oxygen double bonds (O-058), H attached to C (sp^3)/C (sp^2) (H-047), substituted aromatic C (sp^2) (nCaR) and donor atoms for H-bonds with N and O atoms (nHDon). All these parameters are simple and also are able to encode different aspects of the inhibition mechanism of the sulfonamides.

Supuran et al. showed the importance of HOMO and LUMO energies and donor-acceptor pairs on inhibition [40]. These factors play some roles on stability of inhibitor-isozyme and interaction energy. The appearance of nHDon among other descriptors in the model shows the importance of hydrogen bonding between inhibitor and isozyme. Different electron withdrawing groups such as nitro or oxygen double bond on sulfonyl substituents and also halogens of R–CX–R would stabilize the inhibition–isozyme pair. It has been shown that sulfonamides with these substituents have higher activity compared with the unsubstituted derivatives [20,21]. Descriptors such as nCaR, C-001, H-047, C-026 and O-058 are important from pharmaceutical point of view, because they can induce lipophilicity (nCaR, C-001, H-047) or higher hydrosolubility (O-058, C-026) to the CA II inhibitors. The appropriate balance between hydro and liposolubility is important for inhibitor activity. This is correlated with the architecture of the CA active site [21,40]. It means that a good CA inhibitor must possess moieties with ability to interact with both hydrophobic and hydrophilic substituents. It was found that another important factor in the interaction between inhibitor and isozyme is the size of the inhibitors [40,41]. Appearance of Balaban type index and eigenvalue distance matrix confirms that the size of the inhibitor plays a major role on inhibition mechanism. It is noteworthy that all these eight simple descriptors are one- or two-dimensional ones. Another point is the fact that most of the sulfonamides studied in the previous works and this one have single bond(s). This means that barrier to rotation is not large and rotation of some parts of the molecules with respect to the other parts is probable. Consequently, calculation of three-dimensional parameters using AM1 Hamiltonian is not accurate. The one- and two-dimensional descriptors of the present work are superior in this respect.

Fig. 3 shows the plot of the predicted GA-KPLS-ANN values of log IC_{50} for the inhibitors in the data set. The residuals of the calculated values of log IC_{50} are plotted against the experimental ones in Fig. 4 for the training, test and validation sets. The propagation of the residuals in both sides of zero line indicates that no systematic error exists in the development of GA-KPLS-ANN.

5. Conclusion

Our main focus of the present work was finding a strategy to select the best variables for developing nonlinear models. The previous works showed that the strategy of linear feature

Table 6
Statistics for LMO-CV for comparison of GA-PLS-ANN and GA-KPLS-ANN methods

	GA-KPLS-ANN				GA-PLS-ANN			
	R^2	RMSE	R^2_{cv}	$RMSE_{cv}$	R^2	RMSE	R^2_{cv}	$RMSE_{cv}$
L10O	0.898	0.160	0.875	0.173	0.873	0.171	0.836	0.197
L14O	0.907	0.152	0.865	0.182	0.875	0.175	0.841	0.195
L20O	0.918	0.142	0.868	0.183	0.868	0.182	0.840	0.195

Table 7
R^2 and Q^2 values for GA-KPLS-ANN model after several *Y*-randomization tests

Model	R^2	Q^2
1	0.316	0.082
2	0.014	0.018
3	0.152	0.052
4	0.255	0.058
5	0.063	0.018
6	0.049	0.011
7	0.051	0.003
8	0.189	0.003
9	0.019	0.147
10	0.114	0.005

selection/linear modeling is successful in developing the QSAR models for the systems with linear characteristics. However, two strategies have been presented in the literature for nonlinear systems: (1) linear feature selection-nonlinear modeling and (2) nonlinear feature selection-nonlinear modeling. Although both strategies can be applied for the nonlinear systems, but a question arises that which one can have a better performance. The main objective of the present work was looking for a proper answer to this question. To assess the performance of the strategy (2) needs a reliable nonlinear feature selection technique. Therefore, the novel nonlinear feature selection method of GA-KPLS that combines genetic algorithms with kernel partial least square is introduced in this paper. The power of this algorithm was demonstrated by selecting the best set of simple one- and two-dimensional descriptors. The performance of this technique was compared with GA-PLS as a linear one. The results of GA-KPLS-ANN compared with those for GA-PLS-ANN and also linear QSAR model suggest that GA-KPLS holds promise for applications in choosing of variables for nonlinear systems. The validation procedures utilized in this work (LOO-CV, LMO-CV and *Y*-randomization) illustrated the accuracy and robustness of the produced model not only by calculating its fitness on sets of training data, but also by testing the predictive ability of the model. The descriptors appearing in the recommended model reveal the role of acceptor–donor pair, hydrogen bonding, hydrosolubility and lipophilicity properties of inhibitor active sites and also the size of inhibitors on inhibitor–isozyme interaction. This paper clearly shows the power of the nonlinear feature selection method of GA-KPLS in choosing variables for the modeling of sulfonamides as CA II inhibitors. However, researches are underway in our laboratory to check the performance of this method using different nonlinear systems.

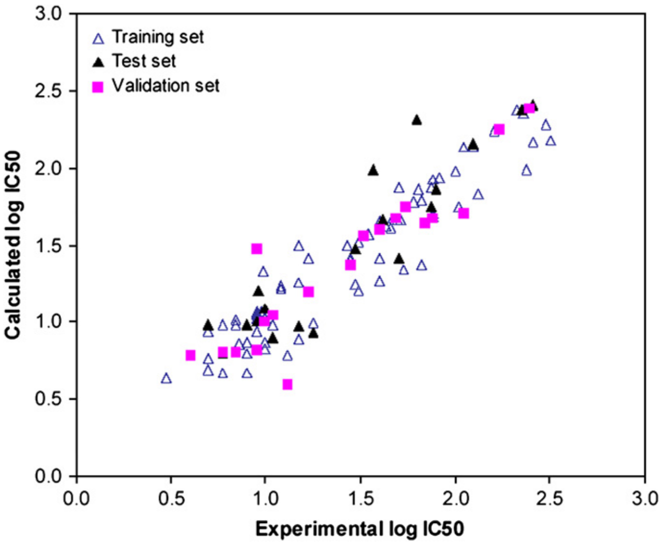

Fig. 3. Plot of the ANN calculated values of log IC_{50} against the experimental ones.

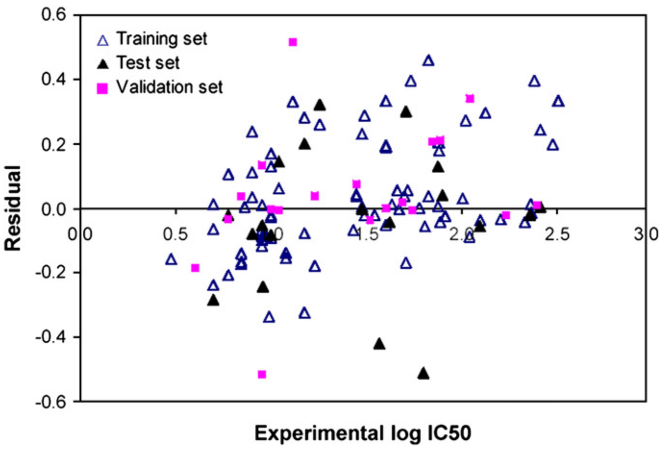

Fig. 4. Plot of residuals vs experimental values of log IC_{50} for the ANN model.

References

[1] C. Hansch, R. Muir, T. Fujita, P.P. Maloney, J. Am. Chem. Soc. 85 (1963) 2817–2824.

[2] C. Hansch, T. Fujita, J. Iwasa, J. Am. Chem. Soc. 86 (1964) 5175–5180.

[3] C. Hansch, D. Hoekman, H. Gao, Chem. Rev. 96 (1996) 1045–1075.

[4] A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA Version 2.0 Reference Manual, Comprehensive Descriptors for Structural and Statistical Analysis, University of Florida, U.S.A., 1994.

[5] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Dragon Web Version 3.0, Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences – University of Milano, TALETE srl – Milano, Italy, 2003.

[6] M. Vracko, S.C. Basak, K. Geiss, F. Witzmann, J. Chem. Inf. Model. 46 (2006) 130–146.

[7] S.J. Cho, M.A. Hermsmeier, J. Chem. Inf. Comput. Sci. 42 (2002) 927–936.

[8] T.J. Hou, J.M. Wang, N. Liao, X.J. Xu, J. Chem. Inf. Comput. Sci. 39 (1999) 775–781.

[9] M.H. Fatemi, M. Jalali-Heravi, E. Konuze, Anal. Chim. Acta 486 (2003) 101–108.

[10] D. Rogers, A.J. Hopfinger, J. Chem. Inf. Comput. Sci. 34 (1994) 854–866.

[11] R. Leardi, J. Chemom. 8 (1994) 65–79.

[12] H. Kubinyi, Quant. Struct.–Act. Relat. 13 (1994) 285–294.

[13] R. Leardi, A.L. Gonzales, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.

[14] R. Leardi, J. Chemom. 14 (2000) 643–655.

[15] M.J. González, J. Caballero, A. Tundidor-Camba, A.H. Helguera, Bioorg. Med. Chem. 14 (2006) 200–213.

[16] B. Hemmateenejad, M.A. Safarpour, F. Taghavi, J. Mol. Struct. (Theochem) 635 (2003) 183–190.

[17] B. Hemmateenejad, Chemom. Intell. Lab. Syst. 75 (2005) 231–245.
[18] R. Rosipal, L.J. Trejo, J. Mach. Learn. Res. 2 (2001) 97–123.
[19] K. Kim, J.M. Lee, I.-B. Lee, Chemom. Intell. Lab. Syst. 79 (2005) 22–30.
[20] G. Melagraki, A. Afantitis, H. Sarimveis, O. Igglessi-Markopoulou, C.T. Supuran, Bioorg. Med. Chem. 14 (2006) 1108–1114.
[21] B.W. Clare, C.T. Supuran, Eur. J. Med. Chem. 34 (1999) 463–474.
[22] V.K. Agrawal, Sh. Bano, C.T. Supuran, Eur. J. Med. Chem. 39 (2004) 593–600.
[23] D.B. Hibbert, Chemom. Intell. Lab. Syst. 19 (1993) 277–293.
[24] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 19 (1993) 1–33.
[25] S. Wold, M. Sjostorm, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[26] K. Tang, T. Li, Anal. Chim. Acta 476 (2003) 85–92.
[27] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2002) 69–77.
[28] S. Haykin, Neural Network, Prentice-Hall, Englewood Cliffs, NJ, 1994.
[29] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, VCH, Weinheim, 1999.
[30] N.K. Bose, P. Liang, Neural Network, Fundamentals, McGraw-Hill, New York, 1996.
[31] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A 927 (2001) 211–218.
[32] A. Guven, S. Kara, Expert Syst. Appl. 31 (2006) 199–205.
[33] M.T. Hugan, M.B. Menhaj, IEEE Trans. Neural Netw. 5 (1994) 989–993.
[34] M. Jalali-Heravi, A. Kyani, J. Chem. Inf. Comput. Sci. 44 (2004) 1328–1335.
[35] L. Douali, D. Villemin, D. Cherqaoui, J. Chem. Inf. Comput. Sci. 43 (2003) 1200–1207.
[36] D.K. Agrafiotis, W. Cedeno, V.S. Lobanov, J. Chem. Inf. Comput. Sci. 42 (2002) 903–911.
[37] S. Wold, Quant. Struct.–Act. Relat. 10 (1991) 191–193.
[38] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269–276.
[39] J.G. Topliss, R.P. Edwards, J. Med. Chem. 22 (1979) 1238–1244.
[40] C.T. Supuran, A. Popescu, M. Ilisiu, Eur. J. Med. Chem. 31 (1996) 439–447.
[41] C.T. Supuran, B.W. Clare, Eur. J. Med. Chem. 32 (1997) 311–319.