Cent. Eur. J. Chem. • 8(4) • 2010 • 877-885
DOI: 10.2478/s11532-010-0050-y

VERSITA

Central European Journal of **Chemistry**

Modeling the activity of 2-phenylnaphthalene inhibitors using self-training artificial neural networks

Research Article

Zahra Garkani-Nejad*, Naser Jalili-Jahani

Department of Chemistry, Faculty of Science, Vali-e-Asr University, 7718897111 Rafsanjan, Iran

Received 07 July 2009; Accepted 14 March 2010

Abstract: The present study investigates the quantitative structure-activity relationship (QSAR) of 2-phenylnaphthalene ligands on an estrogen receptor (ER_α). A data set comprising 70 derivatives of 2-phenylnaphthalene is used. The most suitable parameters, classified as topological, geometric and electronic are selected using a combination of genetic algorithm and multiple linear regression (GA-MLR) methods. Then, selected descriptors are used as inputs for a self-training artificial neural network (STANN). Analysis of the results suggests that the STANN model shows superior results compared to the multiple linear regressions (MLR) by accounting for 91.0% of the variances of the antiseptic potency of the 2-phenylnaphthalene derivatives. The accuracy of the 8-4-1 STANN model is illustrated using leave-multiple-out (LMO) cross-validation and Y-randomization techniques.

Keywords: *Quantitative structure-activity relationship • 2-phenylnaphthalenes • Genetic algorithm Regression analysis • Self-training artificial neural networks*

1. Introduction

Estrogens are a group of naturally-occurring steroid hormones that play an indispensable role in the growth, development and preservation of various tissues. Previously, the common assumption was that estrogen-mediated events were regulated by only one estrogen receptor, now known as ER_α [1]. However, the discovery of a second estrogen receptor subtype (ER_β) in 1996 resulted in an ardent interest in clarifying ER_β function and identifying various aspects of estrogen biology mediated by it [2,3].

Significant sequence homology is observed in the DNA and ligand binding domains (LBDs) of ER_α and ER_β, despite the incongruity of the expression patterns of the two subtypes. ER_β, though widely encountered in numerous tissues, is predominantly found in ovarian granulosa cells, lung, bladder, and prostate tissues, while ER_α is mainly expressed in uterus, kidney and ovarian theca cells [4-6].

During recent years, researchers have aimed their attention at identifying selective ER_β ligands from various classes of molecules, though only a few groups of molecules have been reported to have ER_β selectivity. 2-phenylnaphthalene derivatives are among those ligands tending chiefly to ER_β as opposed to ER_α.

With a well-organized study of the effects of different substituents on the inhibitory behavior of compounds with similar scaffolds, the design of compounds with improved activity can be accomplished. Moreover, the development and application of computational procedures have facilitated the attainment of this objective. The method of quantitative structure-activity relationships (QSAR) has proven to be an effective means for investigating the inhibitory activity of various categories of compounds.

Many QSAR studies have been successfully conducted to model the activities of various types of agents [7-15]. Recently, different derivatives of six series of molecules have been reported as 2-phenylnaphthalene inhibitors [16]. The same SAR study reports the bioactivity of 2-phenylnaphthalene derivatives on ER_β [16]. In the present work, a quantitative structure-activity relationship (QSAR) study is conducted on these ligands and their bioactivity on ER_α. The purpose of this inquiry

* *E-mail: garakani@mail.vru.ac.ir*

is to select appropriate predictors using a combination of genetic algorithms and linear regression techniques. Furthermore, we attempt to assess the ability of STANN to model the bioactivity of the ligands on ER_{α}.

2. Theory

2.1. Self-training artificial neural network

A self-training artificial neural network (STANN) [17] is a procedure for updating the weights of neural nodes and training the neural networks in a parallel fashion. The details of the STANN method are described elsewhere [18,19]. In the STANN procedure, an important aspect is the existence of a neural network (network 2), which trains another network (network 1). Network 2, which is a Back-Propagation Artificial Neural Network (BP-ANN), produces the updated weights for network 1. The architecture of a STANN is shown in Fig. 1. During the training, the normalized inputs are changed by some infinitesimal amount delta (Δ). Because a sigmoid transfer function is used, which has a linear region around 0.5, it is desirable when adding the delta value to the normalized input to adjust the input towards the linear region. Thus, positive delta values should be added to normalized inputs which are less than 0.5, and negative delta values should be added to normalized inputs which are greater than 0.5. For the hidden layer, a similar procedure is used. Network 1 uses weights updated by training network 2. Thus, training of the artificial neural network 1 is not carried out with algorithmic code, but rather by a network training a network.

In two previous works, we have compared the performance of the STANN with the conventional ANN

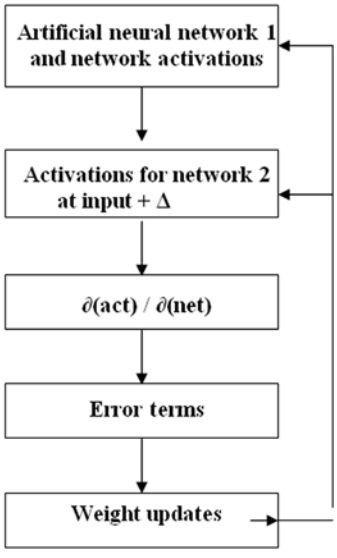

Figure 1. The architecture of a STANN

in predicting the gas chromatographic relative retention times [18] and relative response factors [19] of various organic compounds. It was shown that use of the STANN procedure reduces the number of the adjustable parameters in the network and the optimization procedure was faster compared to the conventional ANN. In a third work, we have used STANN for studying the retention behavior of different organic compounds in reversed phase liquid chromatography on different stationary phases [20].

In the present work, we have used the STANN method for investigating the nonlinear characteristics of inhibitor activity of 2-phenylnaphthalene ligands on the estrogen receptor (ER_{α}). The STANN program was written in Fortran 77 in our laboratory. A three-layer network with a sigmoid transfer function was designed. Before training the STANN, the input and output values of the networks were normalized between 0.1 and 0.9. The initial weights were selected randomly between -1.3 and +1.3. Then, the network was trained with the training set to optimize the values of the weights and biases using the BP strategy. The number of neurons in the hidden layer, the learning rate and the momentum were all optimized. To evaluate the performance of the STANN, the standard error of training or calibration and the standard error of the test set were measured. Additionally, a leave-multiple-out cross-validation method was used to evaluate the STANN model. This technique is described in the next section.

2.2. Cross-validation analysis

The consistency and reliability of any method can be explored using the cross-validation technique [21]. Two different strategies of leave-one-out (LOO) and leave-multiple-out (LMO) can be employed in this method. In the LOO strategy, by deleting one object in each case from the training set, multiple models can be produced. The predicted error sum of squares (PRESS) is a standard index to measure the accuracy of a modeling method based on the cross-validation technique. Based on the PRESS and SSY (sum of squares of deviations of the experimental values from their mean) statistics, the Q^2_{Loo} value can be easily calculated by Eq. 1:

$$Q^2_{L00} = \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n}\left(y_{\exp} - y_{pred}\right)^2}{\sum_{i=1}^{n}\left(y_{\exp} - \bar{y}\right)^2} \quad (1)$$

In the case of LMO, M represents a group of randomly selected data points which are left out at the beginning and are predicted by a model that is developed using the remaining data points. So, M molecules are considered as the prediction set. The Q^2_{LMO} value can be calculated using Eq. 2:

$$Q^2_{LMO} = \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{test}\left(y_{\exp} - y_{pred}\right)^2}{\sum_{i=1}^{train}\left(y_{\exp} - \bar{y}_{train}\right)^2} \quad (2)$$

In the present work, calculation of Q^2_{LMO} for the STANN method was based on a random selection of groups of 10 samples. The higher the Q^2_{Loo} or Q^2_{LMO} values, the higher the predictive power of the model. A more detailed description of this method can be found elsewhere [21].

3. Experimental Procedure

3.1. Data set

The six classes of compounds studied in the present work are all derivatives of the 2-phenylnaphthalene scaffold [16]. The chemical structures and logarithmic experimental activities of these compounds are shown in Fig. 2 and Table 1.

The activity parameter IC_{50} refers to the molar concentration of each ligand at 50% of ER_α inhibition. As such, this requires that the ligand interact with the ligand binding domain of ER_α. To calculate the molecular descriptors, the three-dimensional structures of the ligands under study were optimized using the semi-empirical quantum-chemical methods of the AM1 Hamiltonian method, as implemented in the Hyperchem package [22]. The molecular structures were optimized using the Polak-Ribiere algorithm until the root mean square gradient was less than 0.01.

3.2. Molecular descriptors

The selection and calculation of structural descriptors as numerical parameters that reflect chemical structures is an essential step in every QSAR study. In the present study, 12 molecular descriptors were generated with the Hyperchem package after optimizing the molecular structures. These descriptors include the van der Waals volume and surface area of the molecules as geometric descriptors, partial charges, refractivity, polarizability, electronic energy and hydration energy as electronic descriptors, molecular mass, and so on.

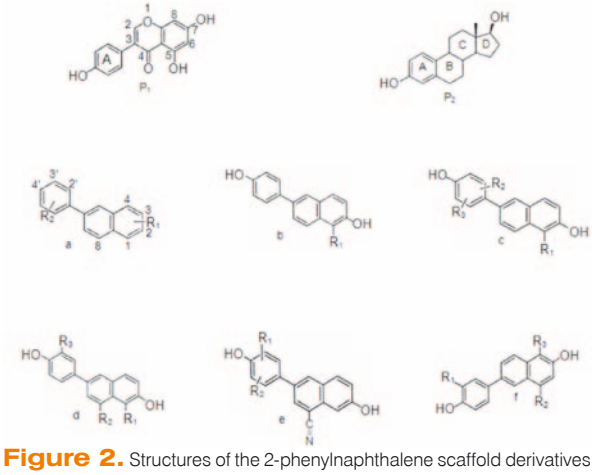

Figure 2. Structures of the 2-phenylnaphthalene scaffold derivatives together genistein and estradiol

Next, the Dragon software version web 3 was used to produce additional descriptors [23]. For each molecule, a total of 1497 descriptors were computed using this software. Descriptors that had the same values for more than 90% of the molecules were eliminated. The correlation between descriptors was then calculated. Pairs of variables with a correlation coefficient greater than 0.90 were classified as intercorrelated, and only one of them was used in developing the final model. A total of 360 descriptors were considered for further investigation after eliminating the descriptors that had the same value for all molecules or were intercorrelated.

3.3. Selection of descriptors by Genetic Algorithm

A genetic algorithm (GA) is a simulation method based on notions from Darwin's theory of evolution in that it imitates some processes observed in natural evolution. In QSAR studies, the GA method has been successfully applied for feature selection. Moreover, an approach incorporating GA with PLS (GA-PLS) has been introduced for descriptor selection in QSAR studies [24].

In the present work, we applied the GA-MLR method using the MATLAB software [25] in order to select descriptors that are most relevant to the prediction of bioactivity [26,27]. The 360 previously mentioned descriptors were used as input to the GA-MLR program, and the bioactivity of the ligands was obtained as output. In this algorithm, a population of n subsets is created, each containing a random combination of descriptors. The fitness of each subset is evaluated. Then, using techniques loosely based on biological genetics and evolution, a new population of subsets is created. The algorithm continues until a stopping criterion is reached. The fitness value of the final selected subset of descriptors and/or the number of generations in the GA-MLR program could be used as stopping criteria.

3.4. Regression analysis

A multiple linear regression procedure was used for model development [28]. For regression analysis, the data set was divided into two groups: training and prediction sets. The molecules included in these sets were selected randomly. In a previous study, we have considered the effect of the size of the test set selected from the main data set, as a percent of the main data set [29]. We have shown that if the percent of the test set is

very low, there will be an uncertainty in the correlation coefficient of the prediction set compared to the main data set. Also, if the percent of the test set is very high, the model obtained cannot be complete and cannot accurately predict the desired property. Thus, there exists an optimum range for the size of the test set. If the test set consists of between 15% and 40% of the main data set, the model constructed with the training set can predict the test set as well as the training set. As such, we randomly selected 40% of the main data set as the test set, and the remaining molecules were placed in the in training set. Therefore, a training set comprising 50 molecules was used to generate the models, and a test set comprising 20 molecules was used to evaluate the generated models. The most successful model was then chosen. For this purpose, it is common to consider the number of descriptors in the model, adjusted correlation coefficient (R^2) and standard error (SE) for the training and prediction sets. A reliable MLR model is one that yields high R^2 values, low SE and uses the lowest number of descriptors. Moreover, the model should have a high predictive ability. The specifications of the model selected are illustrated in Table 2.

Table 1. Experimental and predicted values of $LogIC_{50}$ for the 2-phenylnaphthalene inhibitor derivatives

No	Compound	R_1	R_2	R_3	Experimental $LogIC_{50}$	Calculated $LogIC_{50}$ (STANN)	Calculated $LogIC_{50}$ (MLR)
	Training set						
1	P_1[a]	-	-	-	2.597±0.199	2.611	2.515
2	P_2[b]	-	-	-	0.505±0.136	0.762	0.455
3	a	2-OH	4'-OH	-	2.324±0.152	2.042	2.174
4	a	1-OH	4'-OH	-	1.342±0.099	1.879	2.287
5	a	2-OH	2'-OH	-	3.699	3.254	2.794
6	a	2-OH	H	-	3.129±0.021	3.127	2.947
7	a	3-OH	4'-OH	-	2.645±0.225	2.543	2.645
8	a	3-OH	3'-OH	-	3.422±0.278	3.121	3.067
9	a	3-OH	H	-	3.532±0.235	3.385	3.232
10	b	Cl	-	-	1.959±0.176	1.974	2.149
11	b	Br	-	-	2.425±0.101	2.167	2.034
12	b	F	-	-	1.886±0.158	2.069	2.089
13	b	CN	-	-	3.147±0.249	2.914	3.066
14	b	Ph	-	-	3.09±0.244	3.120	3.016
15	b	OMe	-	-	2.946±0.094	2.621	2.922
16	c	H	2'-F	H	1.38±0.014	1.394	1.62
17	c	Cl	2'-F	H	1.763±0.045	1.612	1.866
18	c	H	2'-F	5'-F	1.431±0.209	1.353	1.427
19	c	H	2'-F	6'-F	2.072±0.147	1.778	1.783
20	c	Cl	2'-F	6'-F	1.544±0.062	1.389	1.538
21	c	H	2'-Cl	H	1.00±0.217	1.271	1.183
22	c	Cl	2'-Cl	H	1.556±0.277	1.501	1.609
23	c	H	2'-OMe	H	2.241±0.065	2.220	2.395
24	c	H	3'-F	H	1.964±0.118	1.787	1.882
25	c	Cl	3'-F	H	2.155±0.085	2.181	2.184
26	c	Cl	3'-Cl	H	2.551±0.079	2.596	2.444
27	c	H	3'-F	5'-F	1.964±0.245	1.953	1.844
28	d	H	F	F	1.322±0.124	1.813	1.789
29	d	Cl	F	H	1.602±0.119	1.492	1.762
30	d	Cl	F	F	2.097±0.083	2.164	2.108
31	d	H	Cl	H	1.477±0.014	1.211	1.124
32	d	Cl	Cl	H	1.633	1.764	1.717
33	d	H	CN	H	2.021±0.194	2.027	2.24
34	d	Cl	CN	H	2.037±0.115	1.995	2.192
35	d	Cl	CN	F	2.476±0.213	2.530	2.438
36	d	H	CHO	F	2.364±0.171	2.126	2.128
37	d	H	$CH=CH_2$	F	2.405±0.191	2.390	2.479
38	d	H	ethyl	F	2.371±0.057	2.478	2.537
39	d	H	$C=CCH_3$	F	2.535±0.077	2.571	2.457
40	e	2'-F	5'-F	-	2.27±0.037	2.003	1.908
41	e	2'-F	6'-F	-	1.653±0.164	1.611	1.701
42	e	3'-F	5'-F	-	2.739±0.178	2.572	2.39
43	f	F	CN	H	1.982±0.19	1.955	1.855
44	f	F	CN	Br	2.064±0.176	2.012	2.03
45	f	F	CN	CN	2.943±0.024	2.869	3.177
46	f	F	CN	Cl	1.74±0.387	2.172	2.075
47	f	F	CCMe	H	2.255±0.219	2.339	2.052
48	f	F	CHO	H	1.863±0.137	2.013	2.031
49	f	F	$CH=CH_2$	H	2.722±0.5	2.854	3.123
50	f	F	ethyl	H	2.053±0.315	2.131	2.138

Table 1 continued. Experimental and predicted values of $LogIC_{50}$ for the 2-phenylnaphthalene inhibitor derivatives

No	Compound	R_1	R_2	R_3	Experimental $LogIC_{50}$	Calculated $LogIC_{50}$ (STANN)	Calculated $LogIC_{50}$ (MLR)
	Prediction set						
51■	a	1-OH	3'-OH	-	3.162±0.215	3.128	3.11
52▲	a	H	4'-OH	-	2.805±0.201	2.931	2.964
53■	b	NO2	-	-	2.851±0.081	2.599	2.267
54▲	c	H	2'-F	6'-F	1.013±0.101	1.002	1.198
55■	c	H	3'-Cl	H	2.029±0.122	1.880	2.064
56▲	d	H	F	H	1.342±0.02	1.513	1.685
57■	d	H	CN	F	2.322±0.252	2.395	2.36
58▲	e	2'-F	H	-	1.415±0.05	1.861	2.009
59▲	f	H	CN	H	1.924±0.171	1.974	1.823
60■	f	H	CN	Cl	1.991±0.053	1.802	1.864
61▲	a	2-OH	3'-OH	-	2.362±0.172	2.594	3.059
62▲	b	Me	-	-	2.45±0.071	2.381	2.38
63■	c	H	2'-Me	H	1.602±0.011	1.759	1.657
64■	c	Cl	2'-Me	H	1.602	1.646	1.763
65■	c	Cl	3'-F	5'-F	2.715±0.105	2.758	2.486
66▲	d	Br	CN	H	2.117±0.166	2.322	2.417
67■	d	H	C=CH	F	2.393±0.334	2.261	2.28
68▲	d	F	CN	F	2.494±0.104	2.743	2.695
69■	f	F	CCH	H	1.863±0.434	1.892	1.617
70▲	f	F	CN	Me	2.609±0.145	2.381	2.353

[a] P_1 , [b]P_2 are genistein and estradiol, respectively that regarded as primary compounds in treating inflammatory diseases.
Values without SDs are for a single determination. ■ and ▲ refer to test and prediction sets in STANN model, respectively.

Table 2. Selected descriptors of multiple linear regression

Descriptor	Type of descriptor	Notation	Coefficient	Mean effect
Topological descriptors (Path/walk 2-Randic shape index)	Topological	Pw2	-50.786(±7.878)	-30.058
3D-MoRSE descriptors (Weighted by atomic masses)	Geometric	Mor02m	0.261(±0.041)	5.488
WHIM descriptors (Weighted by atomic van der waals volumes)	Geometric	E1v	13.798(±2.501)	5.472
3D-MoRSE descriptors (Weighted by atomic van der waals volumes)	Geometric	Mor28V	-4.068(±0.783)	1.374
Burden eigenvalues (Weighted by atomic Sanderson electronegativities)	Electronic	Behe5	-7.746(±1.436)	-22.508
Information indices (Bond information content ,neighborhood symmetry of 4-order)	Topological	Bic4	-13.276(±3.606)	-12.037
3D-MoRSE descriptors (Weighted by atomic Sanderson electronegativities)	Geometric	More29e	0.845(±0.252)	-0.213
Connectivity indices (Average connectivity index chi-2)	Topological	X2a	-36.841(±13.277)	-10.351
Constant			65.022(±8.802)	

$R^2_{training}$=0.825, $R^2_{prediction}$=0.724, $SE_{training}$ =0.276, $SE_{prediction}$=0.305

In order to avoid over-correlation of the regression equations, the R^2 reduction was monitored as a function of the number of descriptors used, as shown in Fig. 3. The procedure was stopped when the ΔR^2 of two consecutive regression equations was less than or equal to 0.02 (see Fig. 3). It can be seen from this figure that the change in R^2 after eight descriptors is relatively linear. Therefore, we have chosen eight descriptors as the optimum number of parameters. The descriptors in this model were Pw2, Mor02m, E1v, Mor28V, Behe5, Bic4, and More29e, the definitions of which are presented in Table 2. Based on the correlation matrix (Table 3), it can be surmised that there are no significant correlations between the selected descriptors.

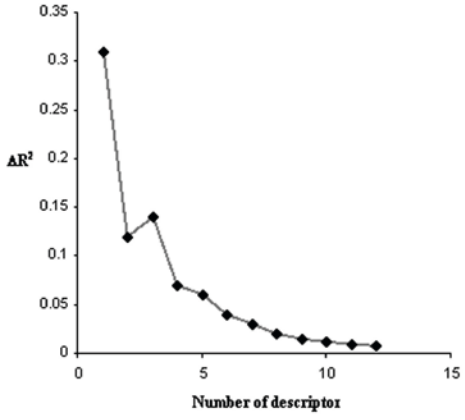

Figure 3. The plot of ΔR^2 *versus* the number of descriptors

Table 3. Correlation matrix for the eight selected descriptors

	Pw2	Mor02m	E1v	Mor28V	Behe5	Bic4	More29e	X2a
Pw2	1							
Mor02m	0.282	1						
E1v	0.092	-0.165	1					
Mor28V	-0.184	0.266	0.237	1				
Behe5	0.229	0.144	-0.145	-0.503	1			
Bic4	0.446	0.179	-0.114	0.011	-0.265	1		
More29e	-0.337	0.261	-0.124	0.183	-0.101	0.003	1	
X2a	-0.474	-0.229	0.026	0.365	-0.538	-0.113	0.222	1

3.5. STANN generation

The networks were generated using the eight descriptors appearing in the MLR model as their inputs and $\log IC_{50}$ values as their output. For STANN generation, the data set was divided into three groups: training, prediction, and test sets. The training set, comprising 50 molecules, was used for the model generation. However, the test set, comprising 10 molecules, was used to maintain the overtraining. Finally, the validation set, comprising 10 molecules, was used to evaluate the generated model. It is worth noting that the molecules in the test and prediction sets were the same as those selected as the prediction set in the GA-MLR model.

A three-layer network accompanied by a sigmoid transfer function was designed for each STANN. For STANN calculations, a program has been written in Fortran 77 in our laboratory. For optimization of the weights and bias values, the network was trained by the back propagation technique using the training set. The appropriate number of nodes in the hidden layer was identified by training the network with different numbers of nodes in the hidden layer, and selecting the optimal number. The learning rate, momentum and the number of epochs were then optimized in a similar way; the optimized conditions were found to be 0.01, 0.95, and 25100, respectively.

To evaluate the effectiveness of the outputs compared with the target values, the standard error (SE) measure was used. For evaluating the over fitting, the training of the network for the prediction of $\log IC_{50}$ should stop when the SE of the test set begins to increase while SE of training set continues to decrease. After simulation, the values of the predicted data were transformed to the true values, and standard error values were calculated from the transformed data.

4. Results and Discussion

The main purpose of the present study was to develop a QSAR model for predicting the activity parameter ($\log IC_{50}$) of 2-phenylnaphthalenes, shown in Fig. 2. This figure and Table 1 illustrate that the inhibitors encompass six different classes, with completely different substituents. Moreover, in naturally complex biological phenomena, these compounds act as inhibitors of chronic inflammatory diseases. Thus, the development of an effective and versatile QSAR model that can accurately predict$\log IC_{50}$ values is required.

At the outset, we developed a linear model of MLR, envisaging two objectives: first, the selection of appropriate descriptor variables, which was accomplished by the use of a multiple linear regression procedure. Second, we set out to evaluate the linear link between the bioactivity parameters of 2-phenylnaphthalenes and their molecular structures and properties. As Table 2 illustrates, 8 descriptors were selected out of a total of 360: Pw2, Mor02m, E1v, Mor28v, Behe5, Bic4, Mor29e, and X2a. These descriptors are classified as topological (Pw2, Bic4 and X2a), geometric (E1v Mor02m, Mor29e and Mor28v), and electronic (Behe5) [30]. This suggests that topological, geometric and electronic features may all play a role in the inhibitory activities of 2-phenylnaphthalenes.

To examine the relative importance as well as the contribution of each descriptor in the model, the value of the mean effect (MF) was calculated for each descriptor. This calculation was performed with the equation below, displayed in the last column of Table 2.

$$MF_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}} \qquad (3)$$

MF_j represents the mean effect for the considered descriptor j, β_j is the coefficient of the descriptor j, d_{ij} stands for the value of the target descriptors for each molecule and m is the descriptor number in the model. The MF value indicates the relative importance of a descriptor compared to the other descriptors in the model. Its sign corresponds to the variation direction in the value of the predicted activity as a result of a change in the descriptor value.

Table 4. Statistical results of STANN model compared to GA-MLR model

Model	$R^2_{training}$	$SE_{training}$	R^2_{test}	SE_{test}	$R^2_{prediction}$	$SE_{prediction}$
STANN	0.910	0.199	0.939[c]	0.139 [c]	0.903[b]	0.194[b]
GA-MLR	0.825	0.276	-	-	0.724[a]	0.305[a]

[a] The number of molecules in the prediction set in GA-MLR model was 20.
[b] The number of molecules in the prediction set in STANN model was 10.
[c] The number of molecules in the test set in STANN model was 10.

Fig. 4 shows the absolute relative mean effects for the MLR model. It can be asserted that the Pw2 (topological descriptor) is the most essential parameter influencing the inhibitory behavior of the molecules. In the MLR modeling, an R^2 value for the prediction set of 0.724 was obtained, which suggests that the model is able to account for 72.4% of the variances of the $logIC_{50}$. As a second step, the non-linear characteristics of the descriptors were investigated. Therefore, a self-training artificial neural network was developed, using the descriptors appearing in the MLR model as inputs. It is a common practice to optimize the parameters of the number of nodes in the hidden layer, learning rate, and momentum in developing a reliable network. Furthermore, a type of transfer function was optimized. In the present work, different numbers of neurons in the hidden layer were tested at an arbitrary learning rate and momentum, epochs, and transfer function. The number of neurons in the hidden layer with the minimum value of the SE was selected as the optimum number. Then, learning rate, momentum, epochs, and transfer function were optimized in a similar way. The experimental and calculated values of the 2-phenylnaphthalene scaffold inhibitor activities were analyzed in this work through the use of MLR and STANN methods, as shown in Table 1. The results of the STANN model compared to the MLR model are illustrated in Table 4. As shown, the R^2 value for the prediction set rose dramatically from 0.724 for the MLR models to 0.939 for the STANN model.

The consistency and reliability of a method can be explored using the cross-validation technique. The leave-multiple-out (LMO) cross-validation was carried out for both the MLR and STANN methods. M represents a group of randomly selected data points (*i.e.*10 molecules) which would be left out at the beginning of the analysis, and would be predicted by the model developed using the remaining data points. In the present work, calculations of Q^2_{L10O} and SEs for the training and prediction sets were based on 100 random selections of groups of 10 molecules (see Table 5). The cross validation results confirm the results of Table 4.

In order to ensure the robustness of the STANN model, the Y-randomization test was also performed. The dependent variable vector ($logIC_{50}$) was randomly

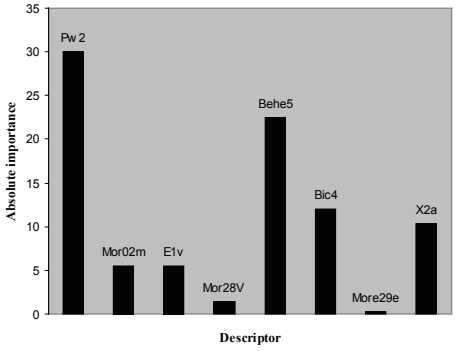

Figure 4. Absolute relative mean effects for MLR model

Table 5. Obtained statistical parameters of L10O[a] cross-validation for STANN and GA-MLR models

Model	Training set		Prediction set	
	Q^2	SE	Q^2	SE
STANN	0.869	0.223	0.918	0.231
GA-MLR	0.786	0.282	0.802	0.263

[a]Calculation of Q^2_{L10O} was based on 100 random selection of groups of 10 molecules

Table 6. R^2_P and Q^2_{LOO} values after several Y-randomization tests

Iteration	R^2_P	Q^2_{LOO}
1	0.355	0.183
2	0.239	0.082
3	0.153	0.055
4	0.078	0.036
5	0.121	0.059
6	0.033	-0.008
7	0.127	0.082
8	0.221	0.151

shuffled and a new QSAR model was developed using the original independent variable matrix. The new QSAR model is expected to have low R^2_P and Q^2_{LOO} values. Several random shuffles of the y vector were performed and the results are shown in Table 6. The R^2_P and Q^2_{LOO} values indicate that the good results for the STANN model are not due to a chance correlation or structural dependency of the training set.

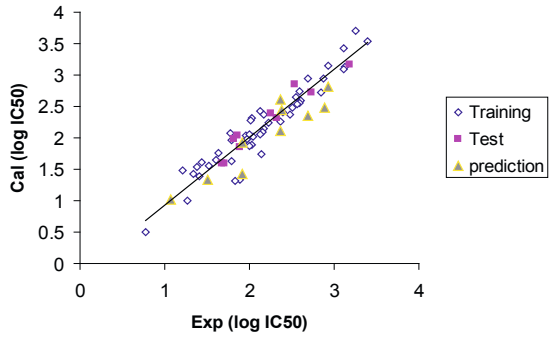

Figure 5. Experimental LogIC_{50} *versus* calculated LogIC_{50} plot

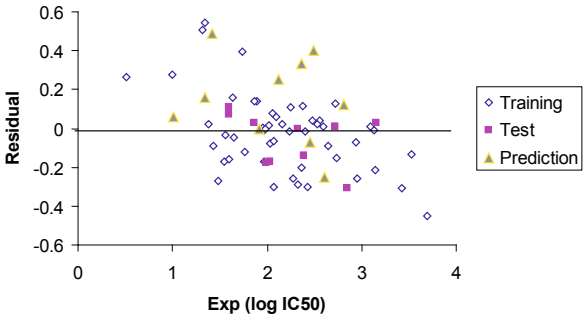

Figure 6. Experimental LogIC_{50} *versus* residual plot

Fig. 5 shows the calculated logIC_{50} versus experimental logIC_{50} for the training, test, and prediction sets. Fig. 6 shows the plot of residuals against the experimental values of logIC_{50} for the STANN model. The spread of the residuals on both sides of zero indicates the lack of systematic error in the development of the STANNs.

5. Conclusions

QSAR methodologies have been effectively utilized for creating an arithmetic link between the bioactivity of 2-phenylnaphthalene and topological, geometric, and electronic molecular descriptors. For the selection of descriptors, an amalgamation of multiple linear regression and genetic algorithm methods (GA-MLR) was used. The superior accuracy of the non-linear over the linear (MLR) model demonstrates the non-linear characteristics of the inhibitory behavior. To develop a neural network, various parameters, including the number of hidden nodes, the learning rate, the momentum, the number of epochs, and the type of transfer function were optimized. In summary, the STANN model is extremely capable of distinguishing between the inhibitory behaviors of different ligands.

References

[1] S. Green, P. Walter, V. Kumar, A. Krust, J.M. Bornert, P. Argos, P. Chambon, Nature 320, 134 (1986)
[2] G.G.J.M. Kuiper, E. Enmark, M. Pelto Huikko, S. Nilsson, J.A. Gustafsson, Proc. Natl. Acad. Sci. U.S.A. 93, 5925 (1996)
[3] S. Mosselman, J. Polman, R. Dijkema, FEBS. Lett. 392, 49 (1996)
[4] J.F. Couse, J. Lindzey, K. Grandien, J.A. Gustafsson, K.S. Korach, Endocrinology 138, 4613 (1997)
[5] S.L. Fitzpatrick, J.M. Funkhouser, D.M. Sindoni, P.E. Stevis, D.C. Deecher, A.R. Bapat, I. Merchenthaler, D.E. Frail, Endocrinology 140, 2581 (1999)
[6] G.G.J.M. Kuiper, B. Carlsson, K. Grandian, E. Enmark, J. Haggblad, S. Nilsson, J.A. Gustafsson, Endocrinology 138, 863 (1997)
[7] D. Hecht, M. Cheung, G.B. Fogel, Biosystems 92, 10 (2008)
[8] J.S. Song, T. Moon, K.D. Nam, J.K. Lee, H.G. Hahn, E.J. Choi, C.N. Yoon, Bioorg. Med. Chem. Lett. 18, 2133 (2008)
[9] M.P. Gonzalez, P. Besada, M.J. Gonzalez Moa, M. Teijeira, C. Teran, Bioorg. Med. Chem. 16, 1658 (2008)
[10] P.R. Duchowicz, A.G. Mercader, F.M. Fernandez, E.A. Castro, Chemom. Intell. Lab. Sys. 90, 97 (2008)
[11] J. Caballero, M. Fernandez, M. Saavedra, F.D. Gonzalez-Nilo, Bioorg. Med. Chem. 16, 810 (2008)
[12] R.F. Freitas, T.I. Oprea, C.A. Montanari, Bioorg. Med. Chem. 16, 838 (2008)
[13] K.M. Nikolic, J. Mol. Graph. Model. 26, 868 (2008)
[14] K. Nikolic, D. Agababa, J. Mol. Graph. Model. 27, 777 (2009)
[15] M.H. Fatemi, S. Gharaghani, Bioorg. Med. Chem. 15, 7746 (2007)
[16] R.E. Mewshaw, R.J. Edsall, C. Yang, E.S. Manas, Z.B. Xu, R.A. Henderson, J.C. Keith, H.A. Harris, J. Med. Chem. 48, 3953 (2005)
[17] Z. Garkani-Nejad, Chromatographia, 70, 869 (2009)
[18] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A 945, 173 (2002)
[19] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A 950, 183 (2002)
[20] M. Jalali-Heravi, Z. Garkani-Nejad, A. Kyani, QSAR & Comb. Sci. 27, 137 (2008)
[21] D.W. Osten, J. Chemom. 2, 39 (1988)

[22] Hyperchem, Molecular Modeling System, Hyper Cube Inc, 1993, www.hyper.com
[23] 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., Milan, Italy, (2003)
[24] K. Tang, T. Li, Chemom. Intell. Lab. Sys. 64, 55 (2002)
[25] MATLAB version 7.1. Mathworks Inc, 2005, www.mathworks.com
[26] R. Leardi, A. Lupianez, Chemolab 41, 195 (1998)
[27] R. Leardi, J. Chemom. 14, 643 (2000)
[28] J.T. Leonard, K. Roy, Bioorg. Med. Chem. 14, 1039 (2006)
[29] Z. Garkani-Nejad, J. Chromatogr. Sci. 48, (2010) In Press.
[30] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics (Wiley- VCH, Weinheim, 2009)