Genetic Algorithm Applied to the Selection of Factors in Principal Component-Artificial Neural Networks: Application to QSAR Study of Calcium Channel Antagonist Activity of 1,4-Dihydropyridines (Nifedipine Analogous)

Bahram Hemmateenejad,[§] Morteza Akhond,[§] Ramin Miri,[‡] and Mojtaba Shamsipur*,[†]

Department of Chemistry, Shiraz University, Shiraz, Iran, Department of Medicinal Chemistry, Faculty of Pharmacy, Medicinal Science University of Shiraz, Shiraz, Iran, and Department of Chemistry, Razi University, Kermanshah, Iran

Received December 24, 2002

A QSAR algorithm, principal component-genetic algorithm-artificial neural network (PC-GA-ANN), has been applied to a set of newly synthesized calcium channel blockers, which are of special interest because of their role in cardiac diseases. A data set of 124 1,4-dihydropyridines bearing different ester substituents at the C-3 and C-5 positions of the dihydropyridine ring and nitroimidazolyl, phenylimidazolyl, and methylsulfonylimidazolyl groups at the C-4 position with known Ca^{2+} channel binding affinities was employed in this study. Ten different sets of descriptors (837 descriptors) were calculated for each molecule. The principal component analysis was used to compress the descriptor groups into principal components. The most significant descriptors of each set were selected and used as input for the ANN. The genetic algorithm (GA) was used for the selection of the best set of extracted principal components. A feed forward artificial neural network with a back-propagation of error algorithm was used to process the nonlinear relationship between the selected principal components and biological activity of the dihydropyridines. A comparison between PC-GA-ANN and routine PC-ANN shows that the first model yields better prediction ability.

1. INTRODUCTION

The influx of extracellular Ca^{2+} through L-type potential dependent calcium channel is responsible for the regulation of many physiological functions, including smooth and cardiac muscle contraction.[1−3] The discovery that the 1,4-dihydropyridine (DHP) class of calcium channel antagonist inhibits this Ca^{2+} influx represented a major therapeutic advance in the treatment of cardiovascular diseases such as hypertension, angina pectoris, and other spastic smooth muscle disorders.[1] The dihydropyridine class of compounds in which nifedipine is the prototype has been the subject of many QSAR studies.[4−8] It is reported that changes in the substitution pattern at the C-3, C-4, and C-5 positions of nifedipine alter the activity and tissue selectivity. Gaudia et al. reported a QSAR study on the 4-phenyl substituted nifedipine analogous by combination of substituent constants and molecular descriptors.[8] Recently, Viswanadhan et al. reported a comparative study between linear and nonlinear methods for the QSAR study of 4-phenyl substituted DHP compounds.[10] In a recent work, we used some theoretically derived descriptors for the QSAR study of Ca^{2+} channel antagonist activity of 4-nitroimidazolyl dihydropyridines.[11] Two linear methods (PLS and MLR) were used for modeling the relationships between biological activity and molecular descriptors. However, some nonlinear relationships were observed, and for linearization, the second power of the descriptors was incorporated into the data matrix. Our attempts to extend the models to the higher numbers of DHPs and incorporating the 4-phenylimidazolyl and 4-methylsulfonylimidazolyl DHPs into the models failed because of the highly nonlinear and complex relationship between the descriptors and activity.

In contrast, to PLS and MLR, the artificial neural networks (ANN) are capable of recognizing highly nonlinear relationships. Over the past few years, the ANN modeling technique has attracted an increasing interest as a very promising method for classification and multivariate calibration problems.[12,13] The flexibility of ANN enables it to discover more complex relationships in experimental data, when it is compared with the traditional statistical models. Hence, the ANNs provide proper analytical alternatives to conventional techniques and interesting approaches to the QSAR and QSPR studies.[14−17] The principal component-artificial neural network (PC-ANN) was proposed by Gemperline et al., to improve training speed and decrease the overall calibration error.[18] In this method, the input data are subjected to principal component analysis (PCA) before being introduced into the neural network, and the most significant principal components of the original data matrix are selected and used as ANN input.

Obtaining the number of significant principal components is the main problem in the PCA based methods.[19] Different methods have been employed to select the significant PCs, for the calibration purposes. The simplest and most common one is a top-down variable selection where the factors are ranked in the order of decreasing eigenvalues. The factor with the highest eigenvalue is considered as the most significant one, and, subsequently, the factors are introduced into the calibration model until no further improvement of

* Corresponding author phone: (+98)-831-4223307; fax: (+98)-831-4228439; e-mail: mshamsipur@yahoo.com.
[†] Razi University.
[‡] Medicinal Science University of Shiraz.
[§] Shiraz University.

10.1021/ci025661p CCC: $25.00
Published on Web 06/25/2003

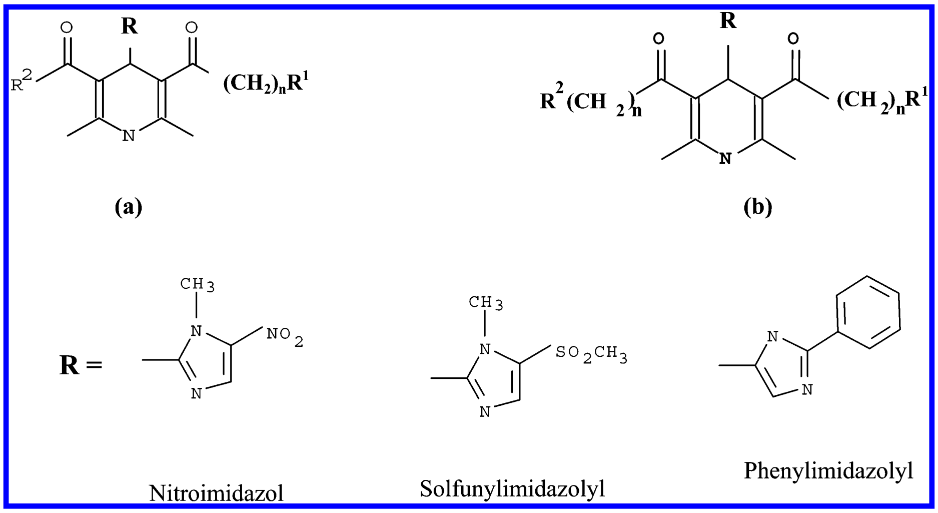

Figure 1. Molecular structure of the 1,4-dihydropyridines used in this study: (a) unsymmetrical and (b) symmetrical.

the calibration model is obtained.[20,21] However, the magnitude of an eigenvalue is not necessarily a measure of its significance for the calibration (see ref 21 and references therein). In the other method, called correlation ranking, the factors are ranked by their correlation coefficient with the property to be correlated (i.e., the dependent variable) and selected by the procedure discussed for eigenvalue ranking.[22] Better results are often achieved by this method. Very recently, Depczynski et al.[23] used genetic algorithm (GA) for the selection of variables in the principal component regression (PCR) and found the best set of PCs that gives lower calibration error. They compared the result of GA factor selection with the two above-mentioned methods and showed that GA gives a result very close to the correlation ranking. A GA is a stochastic method to solve optimization problems defined by a fitness criteria applying evolution hypothesis of Darwin and different genetic functions, i.e., crossover and mutation.[24−26]

A routine method for the selection of factors in PC-ANN is the eigenvalue ranking. Here, we aimed to use genetic algorithm for the selection of the set of factors that resulted in the best model. The data used are 124 new 1,4-dihydropyridines which have different ester substituents at the C-3 and C-5 positions of the dihydropyridine ring and nitroimidazolyl, phenylimidazolyl, and methylsulfonylimidazolyl groups at the C-4 position with known Ca^{2+} channel antagonist activity. Ten different sets of descriptors (837 descriptors) were calculated for each molecule. The principal component analysis was done on each descriptor group to compress the descriptors into principal components. The first most significant factors of each group (describing more than 95% of the original data matrix) were selected and used as input for the ANN, and the best set of PCs was selected by GA. The results compared with the case were all descriptors gathered in a data matrix and used as input for PC-ANN with the eigenvalue ranking method for the selection of PC. The results indicate that the first model represented higher prediction ability and contains lower calibration error.

2. EXPERIMENTAL SECTION

2.1. Activity Data and Descriptor Generation. The biological data used in this study are the calcium channel antagonist activity in Guinea-pig Ileal, log(1/IC_{50}), of a series of 1,4-dihydropyrines, which have different ester substituents at the C-3 and C-5 positions of DHP ring and nitroimidazolyl, phenylimidazolyl, and methylsulfonylimidazolyl groups at C-4 position (total number of 124 compounds). The basic structures of these compounds are shown in Figure 1. The synthesis and determination of biological activity of some of these compounds were done in this research group,[27−29] and other data were taken from the literature (Table 1).[30−33] The log(1/IC_{50}) values of these compounds are listed in Table 1.

Molecular descriptors define the molecular structure and physicochemical properties of molecules by a single number. A wide variety of descriptors have been reported in the literature for use in the QSAR analysis.[34−39] There is a recently increased use of theoretical descriptors in QSAR studies. Here, 837 descriptors including constitutional descriptors,[34] topological indices,[35,36] topological charge indices,[38] geometrical descriptors,[35] molecular walk counts,[37] Burden’s eigenvalue descriptors,[38] autocorrelation descriptors,[39] and physicochemical parameters and liquid properties[34] were generated for each compound (Table 2).

The molecular structures were drawn by the ChemSketch 3.5 freeware.[40] Most of the descriptors were calculated by the Dragon software.[41] The physicochemical parameters such as hydrophobicity, surface area, molecular volume, etc. were generated for each compound by HyperChem version 6.03[42] and ChemSketch 3.5 freeware.[40]

2.2. PC-ANN. A total number of 837 descriptors were calculated for each compound using Dragon, HyperChem, and ChemSketch software, as shown in Table 2. To decrease the redundancy existing in the descriptors data matrix, the correlation of descriptors with each other and with the log1/IC_{50} of the drugs was examined, and collinear descriptors (i.e. $r > 0.9$) were detected. Among the collinear descriptors, one with the lowest correlation with drug activity was removed from the data matrix. At this step, the number of descriptors was reduced to 685. The remaining descriptors were gathered in a new data matrix (**D**). The data set was classified into training ($\mathbf{D}_t$), prediction ($\mathbf{D}_p$), and validation ($\mathbf{D}_v$) sets randomly (the number of molecules used in the training, prediction, and validation sets was 80, 29, and 15, respectively). The same classification was done on the activity data.

Table 1. Activity (Log1/IC_{50}) Data of 4-(1-Methyl-5-nitro-2-imidazolyl) dihydropyridine Derivatives and the Corresponding Predictive Values by PC-ANN and PC-GA-ANN Models

					predicted Log1/IC_{50}							predicted Log1/IC_{50}	
compd no.	R^1	R^2	*n*	Log1/IC_{50}	PC-ANN	PC-GA-ANN	compd no.	R^1	R^2	*n*	Log1/IC_{50}	PC-ANN	PC-GA-ANN
Nitroimidazolyl													
1[a]	*n*-butyl	ethyl	0	11.49	11.5	11.53	39[a]	c-hexyl	c-hexyl	1	9.31	9.67	8.26
2[a]	*n*-butyl	*n*-butyl	0	11.29	12.35	11.17	40[b]	c-hexyl	c-hexyl	2	8.88	8.31	9.4
3[b]	i-propyl	ethyl		11.29	10.88	11.01	41[a]	c-hexyl	c-hexyl	3	7.55	8.53	8.41
4[a]	OCH3	i-propyl	2	12.21	13.1	12.63	42[b]	c-hexyl	c-hexyl	4	7.09	6.21	7.91
5[b]	OH	methyl	2	11.89	12.8	12.3	43[c]	*n*-propyl	*n*-propyl	0	10.05	8.88	10.89
6[a]	COCH3	methyl	3	11.83	11.53	11.61	44[a]	c-propyl	c-propyl	1	7.57	7.86	7.61
7[b]	CN	i-propyl	2	11.83	12.63	12.18	45[a]	c-pentyl	c-pentyl	3	7.89	7.12	8.41
8[a]	OH	i-propyl	2	11.53	12.8	12.1	46[b]	phenyl	phenyl	1	8.58	8.02	8.91
9[a]	OH	ethyl	2	11.45	10.93	11.66	47[a]	phenyl	phenyl	2	8.42	8.25	7.32
10[a]	OCH3	methyl	2	11.38	11.8	12.1	48[a]	p-tolyl	p-tolyl	2	7.28	7.99	7.05
11[a]	OCH3	ethyl	2	11.13	11.59	11.63	49[a]	phenyl	phenyl	3	7.35	7.2	8.01
12[c]	c-hexyl	methyl	0	9.01	7.94	7.91	50[a]	phenyl	phenyl	4	6.89	6.37	6.34
13[a]	c-hexyl	ethyl	0	8.40	8.8	8	51[a]	phenyl	phenyl	5	5.95	5.43	6.56
14[c]	c-hexyl	methyl	1	8.57	7.3	8.39	52[a]	i-butyl	i-butyl	0	11.36	12.52	11.98
15[a]	c-hexyl	ethyl	1	8.26	9.62	9.03	53[a]	ONO2	methyl	2	11.59	12.35	11.22
16[a]	c-hexyl	methyl	2	8.40	7.1	7.84	54[a]	ONO2	ethyl	2	10.93	12.1	11.15
17[a]	c-hexyl	ethyl	2	7.82	7.91	7.09	55[b]	ONO2	i-propyl	2	12.02	11.23	12.39
18[a]	c-hexyl	methyl	3	8.33	8.05	8.64	56[b]	ONO2	methyl	3	11.33	12.06	11.86
19[a]	c-hexyl	ethyl	3	8.28	8.26	8.57	57[a]	ONO2	ethyl	3	11.11	11.93	10.83
20[b]	c-hexyl	methyl	4	8.30	8.96	8.45	58[c]	ONO2	i-propyl	3	11.75	11.53	12
21[c]	c-hexyl	ethyl	4	8.22	7.35	7.35	59[a]	ONO2	methyl	4	11.60	12.48	11.2
22[a]	c-pentyl	methyl	3	8.73	8.67	9.29	60[a]	ONO2	ethyl	4	10.87	10.48	10.76
23[c]	c-pentyl	ethyl	3	8.40	8.69	7.5	61[b]	ONO2	i-propyl	4	10.65	12.25	10.04
24[b]	c-propyl	methyl	1	7.48	7.84	7.92	62[a]	$CH(CH_2ONO_2)_2$	methyl	0	11.41	12.25	12.1
25[b]	c-propyl	ethyl	1	7.13	6.49	6.62	63[c]	$CH(CH_2ONO_2)_2$	ethyl	0	11.66	12.11	10.92
26[a]	phenyl	methyl	1	7.70	7.75	7.74	64[a]	$CH(CH_2ONO_2)_2$	i-propyl	0	11.61	11.87	11.2
27[c]	phenyl	ethyl	1	7.70	8.18	7.69	65[a]	$N(CH_3)_2$	methyl	3	8.14	7.69	8.1
28[b]	phenyl	methyl	2	7.70	6.63	7.08	66[a]	$N(CH_3)_2$	ethyl	3	8.96	9.25	9.36
29[b]	phenyl	ethyl	2	7.56	6.84	6.84	67[b]	$N(CH_3)_2$	i-propyl	3	9.33	8.62	8.56
30[a]	p-tolyl	methyl	2	7.90	8.82	8.46	68[a]	$N(CH_3)_2$	methyl	2	8.50	8.75	7.84
31[a]	p-tolyl	Ethyl	2	7.32	7.47	7.89	69[a]	$N(CH_3)_2$	ethyl	2	9.31	9.15	10.32
32[a]	phenyl	methyl	3	7.61	8.24	7.95	70[b]	$N(CH_3)_2$	i-propyl	2	10.13	11.53	10.96
33[a]	phenyl	ethyl	3	7.42	7.38	7.58	71[a]	*tert*-butyl	methyl	0	9.74	9.44	10.04
34[a]	phenyl	methyl	4	7.57	8.87	7.25	72[b]	*tert*-butyl	ethyl	0	10.31	10.23	9.56
35[a]	phenyl	ethyl	4	7.21	7.86	7.46	73[a]	*n*-pentyl	ethyl	0	12.13	12.51	12.63
36[c]	phenyl	methyl	5	7.14	7.13	6.53	74[a]	methyl	ethyl	0	11.13	11.11	11.32
37[a]	phenyl	ethyl	5	7.08	6.14	6.45	75[c]	*n*-pentyl	methyl	0	12.33	13.1	11.86
38[a]	c-hexyl	c-hexyl	0	7.85	8.85	7.45							
Phenylimidazolyl													
76[b]	methyl	c-hexyl	0	9.76	9.17	10.26	87[a]	ethyl	p-tolyl	2	9.24	9.86	8.9
77[a]	methyl	c-hexyl	1	8.55	7.68	8.95	88[c]	i-propyl	p-tolyl	2	9.01	10.04	9.34
78[a]	methyl	c-hexyl	2	8.52	9.05	8.88	89[a]	methyl	Me	0	9.56	8.76	8.5
79[a]	methyl	c-hexyl	3	8.45	9.59	9.29	90[b]	c-hexyl	c-hexyl	0	9.19	9.18	9.22
80[b]	methyl	c-hexyl	4	8.31	8.21	7.99	91[a]	c-hexyl	c-hexyl	1	9.39	10.13	9.26
81[b]	methyl	phenyl	1	8.57	9.23	9.1	92[a]	c-hexyl	c-hexyl	2	9.74	8.49	8.91
82[a]	methyl	phenyl	2	9.28	9.19	9.37	93[c]	c-hexyl	c-hexyl	3	10.92	10.73	11.43
83[a]	methyl	phenyl	3	9.23	8.46	8.97	94[a]	c-hexyl	c-hexyl	4	9.19	9.97	9.49
84[a]	methyl	phenyl	4	8.21	9.27	9.11	95[a]	phenyl	phenyl	2	7.19	7.57	6.94
85[c]	methyl	phenyl	5	8.14	8.91	8.54	96[b]	phenyl	phenyl	5	7.05	7.05	7.16
86[a]	methyl	p-tolyl	2	9.48	9.14	8.84							
Methylsulfonylimidazolyl													
97[a]	methyl	i-propyl	1	6.15	6.75	7.02	111[a]	ethyl	c-hexyl	0	5.62	6.7	5.58
98[a]	methyl	phenyl	1	8.66	8.07	7.9	112[a]	ethyl	c-hexyl	1	6.47	7.22	6.37
99[c]	methyl	phenyl	2	9.73	9.01	9.02	113[b]	ethyl	c-penyl	3	5.90	5.91	5.61
100[a]	methyl	c-hexyl	0	5.85	6.25	5.4	114[b]	ethyl	methyl	2	5.14	5.7	5
101[a]	methyl	c-hexyl	1	6.33	6.34	6.52	115[a]	ethyl	i-propyl	0	5.42	6.17	5.49
102[a]	methyl	c-penyl	3	5.82	5.45	5.57	116[a]	ethyl	*n*-butyl	0	5.92	4.96	5.94
103[a]	methyl	methyl	1	5.21	5.93	5.61	117[a]	ethyl	*tert*-butyl	0	5.45	6	5.86
104[a]	methyl	*n*-propyl	0	5.74	6.89	5.73	118[a]	methyl	methyl	0	8.24	8.27	8.45
105[a]	methyl	i-propyl	0	5.26	4.75	5.2	119[b]	methyl	methyl	1	7.55	6.87	7.36
106[b]	methyl	*n*-butyl	0	5.64	6.68	6.29	120[a]	methyl	methyl	2	5.50	5.13	5.71
107[a]	methyl	*tert*-butyl	0	5.26	6.34	5.2	121[b]	methyl	methyl	3	6.38	7.37	7.24
108[a]	ethyl	i-propyl	1	6.91	6.87	7.27	122[a]	methyl	methyl	4	8.37	8.38	8.36
109[a]	ethyl	phenyl	1	6.90	6.44	6.24	123[b]	i-propyl	i-propyl	1	6.91	7.59	7.52
110[c]	ethyl	phenyl	2	7.64	7.21	7.19	124[b]	*tert*-butyl	*tert*-butyl	0	8.19	8.65	8.18

[a] The compounds used in the training set. [b] The compounds used in the prediction set. [c] The compounds used in the validation set.

Table 2. Brief Description of Type of Descriptors Used in the Study

descriptor	molecular descriptors	no. of descriptors
constitutional	molecular weight, no. of atoms, no. of non-H atoms, no. of bonds, no. of heteroatoms, no. of multiple bonds, no. of aromatic bonds, no. of functional groups (hydroxy, amine, aldehyde, carbonyl, nitro, nitroso, ...), no. of rings, no. of circuits, no. of H-bond donors, no. of H-bond acceptors, chemical composition	37
topological indices	molecular size index, molecular connectivity indices, information contents, Kier shape indices, path/walk-Randic shape indices, Zagreb indices, Schultz indices, Balaban J index, Wiener indices, information contents	69
molecular walk counts	molecular walk counts of order 1−10, self-re-turning of order 1−10	20
Burden eigenvalues	positive and negative Burden eigenvalues weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume	64
charge topological indices	order 1−10 of Galvez charge topological indices, mean topological charge indices order 1−10, global topological charge index, maximum, minimum, average and total charges, local dipole index	28
autocorrelation descriptors	Broto-Moreau autocorrelation of a topological structure, Moran autocorrelation, Geary autocorrelation, H-autocorrelation weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume, leverage autocorrelation weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume, R-autocorrelation weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume	291
molecular profile indices	Randic molecular profile no. 1−20, Randic shape profile no. 1−20	40
three-dimensional and geometrical descriptors	3-D MoRSE signals weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume, 3D-Wiener index, average geometrical distances, molecular eccentricity, spherocity, average shape profile index, distance-distance index	178
VHIM descriptors	unweighted size, shape, symmetry and accessibility directional indices; size, shape, symmetry and accessibility directional indices weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume; total size, shape symmetry and accessibility indices	99
chemical descriptors	LogP, polarizability, density, molar refractivity, parachor, surface tension, molecular volume, molecular surface area, hydration energy, unsaturation index, aromatic ratio	11

The training data matrix was subjected to principal component analysis (PCA) using the singular value decomposition procedure (SVD)

$$\mathbf{D}_t = \mathbf{U}_t\, \mathbf{S}_t\, \mathbf{V}_t^{\mathrm{T}} \tag{1}$$

where $\mathbf{U}_t$ and $\mathbf{V}_t$ are the orthonormal matrices spanning the respective row and column spaces of the data matrix ($\mathbf{D}_t$). $\mathbf{S}_t$ is a diagonal matrix whose elements are the square root of the eigenvalues. The superscripts "T" denote the transpose of the matrix. The eigenvectors included in $\mathbf{U}_t$ are named as principal components (PC). The PCs of the prediction and validation sets were calculated by the equation below:

$$\mathbf{U}_{p/v} = \mathbf{D}_{p/v}\, \mathbf{S}_t^{-1}\, \mathbf{V}_t \tag{2}$$

The extracted PCs were used as the predictor variables (input) for neural network model.

A feed-forward neural network with back-propagation of error algorithm was constructed to model the structure activity relationship. Our network had an input layer, a hidden layer, and an output layer. The input vectors were the PCs ranked by decreasing their corresponding eigenvalues. The number of nodes in the input layer depended on the number of PCs introduced in the network. The PCs were successively introduced into the network, and, in each case, the number of nodes in the hidden layer was optimized. There was only one node in the output layer (i.e., log1/IC$_{50}$). A bias unit with a constant activation of unity was connected to each unit in the hidden and output layers. The ANN models were confined to a single hidden layer, because the network with more than one hidden layer would be harder to train. The training and prediction data sets were used to optimize the network performance. To ensure that the overfitting and underfitting of the ANN model did not occur, for each configuration, the fitness function (η), calculated from both the root-mean-square errors of training and prediction (RMSET and RMSEP, respectively), was used to evaluate the performance of each neuron. This fitness function, recently proposed by Depczynski et al.,[23] was used in this paper

$$\eta = \{[(m_t - n - 1)\,\mathrm{RMSET}^2 + m_p\,\mathrm{RMSEP}^2]/(m_t + m_p - n - 1)\}^{1/2} \tag{3}$$

where m_t and m_p are the number of compounds in the training and prediction sets, respectively, and n represents the number of selected PCs. The training of each network was stopped after observing no improvement in η.

2.3. PC-GA-ANN. The descriptors chosen by the procedure discussed in the previous section (after removing the collinear ones) were classified into 10 groups. Each group of descriptors was then subjected to PCA in order to reduce its dimensionality, and, subsequently, the significant principal components (PCs) of each group, which explain most of the variances in the original data (>95%), were extracted. It should be noted that, like the procedure discussed in the PC-ANN section, the data were classified into training, prediction, and validation sets.

The genetic algorithm was used to select the set of PCs which resulted in the best fitted model. The genetic algorithm applied in this paper uses a binary representation as the coding technique for a given problem; the presence or absence of a PC in a chromosome is coded by either 1 or 0.[24−26] The GA performs its optimization by variation and selection via evaluation of the fitness function defined in the previous section. The operators used here were crossover and mutation. The probability for the application of these operators was varied linearly with the generation renewal.

The ANN model used here was similar to that discussed in the previous section. The input vectors were the set of PCs that selected by the GA in which the number of nodes was dependent on the number of selected PCs. The number of nodes in the hidden layer was optimized through the learning procedure, and there was only one node in the output layer. A bias unit with a constant activation of unity was connected to each unit in the hidden and output layers. Several network configurations were tested, each with a different number of hidden layer elements. The ANN models were confined to a single hidden layer, because the network with more than one hidden layer would be harder to train. To ensure that the overfitting and underfitting of the ANN model did not occur, for each configuration, the fitness function (η) was calculated from the training and prediction data. For each chromosome of the GA, the training was stopped after observing no improvement in the fitness.

3. RESULTS AND DISCUSSION

The compounds used in this study have the same molecular backbone, 2,6-dimethyl-1,4-dihydropyridine (DHP, Figure 1). These compounds have been the aim of many structure activity relationships. However, the QSAR study on the corresponding compounds with varying substituents at the the C-3 and C-5 positions of the DHPs is rare. A total of 124 DHP compounds with known activity was used in this study (Table 1).

All of the descriptors used in this work have been well documented and hence are not discussed here. Table 2 lists the types of descriptors included in the present study and briefly describes them. The first nine groups were calculated by the Dragon software (i.e., the topological and information based descriptors). The indices indicated include those derived from distance matrices, measures of graph complexity, structural information measures, and complementary information measures. The last group is the chemical descriptors, which was driven by the Hyperchem software.

3.1. PC-ANN. The input data for the PC-ANN algorithm were all the calculated descriptors in a single matrix. The results obtained by the principal component analysis on the matrix of descriptors are given in Table 3. The PCs were introduced into the ANN model successively, to decrease their eigenvalues. It should be noted that, in each step, the number of nodes in the hidden layer was varied to reach the lowest fitness function (η), defined in section 2.2. Meanwhile, for each configuration, the correlation coefficient between the predicted and experimental values of log(1/IC_{50}) for both the training and prediction sets were monitored. The last columns of Table 3 indicate the ηs obtained in the presence of the entered PCs and the optimized number of nodes used in the hidden layer (n_H). As it is seen, by introducing more PCs in the ANN model, η is decreased. However, after introducing more than eight PCs no further improvement in η was observed. The number of nodes in the hidden layer is shown to be 3−6, and no systematic relationship between n_H and the number of PCs is observed. The statistical parameters for the PC-ANN model are shown in Table 4. As shown, the designed PC-ANN model has an accepted quality and could predict the activity of the molecules of the validation set with small errors. The optimized PC-ANN model was used to predict the activity of the components, which were not used in any step of ANN modeling (validation set). In Figure 2, the predicted activities of compounds obtained by optimized PC-ANN model are plotted against the experimentally determined activities.

Table 3. Results of PCA of the Descriptors Data Matrix

PC	eigenvalue × 10^{-2}	variance explained % of	cumulative %	n_H	η
1	3.01	34.44	34.44	4	1.38
2	1.42	16.92	51.36	5	1.03
3	1.08	13.09	64.45	3	0.89
4	0.87	11.21	75.66	4	0.82
5	0.72	8.86	84.52	5	0.78
6	0.52	6.38	90.90	6	0.74
7	0.36	4.91	95.81	4	0.71
8	0.22	2.62	98.43	5	0.70
9	0.10	1.05	99.48	3	0.71
10	0.05	0.33	99.81	5	0.71
11	0.01	0.06	99.87	5	0.71

Table 4. Statistical Parameters of the Models Used

parameter	PC-ANN	PC-GA-ANN1	PC-GA-ANN2
RMSET	0.659	0.402	0.397
RMSEP	0.735	0.597	0.457
RMSEV	0.761	0.631	0.562
RSET	7.95	4.27	3.81
RSEP	8.46	5.27	4.61
RSEV	7.90	6.64	5.85
r_t^2	0.906	0.929	0.941
r_p^2	0.878	0.919	0.934
r_v^2	0.842	0.891	0.900

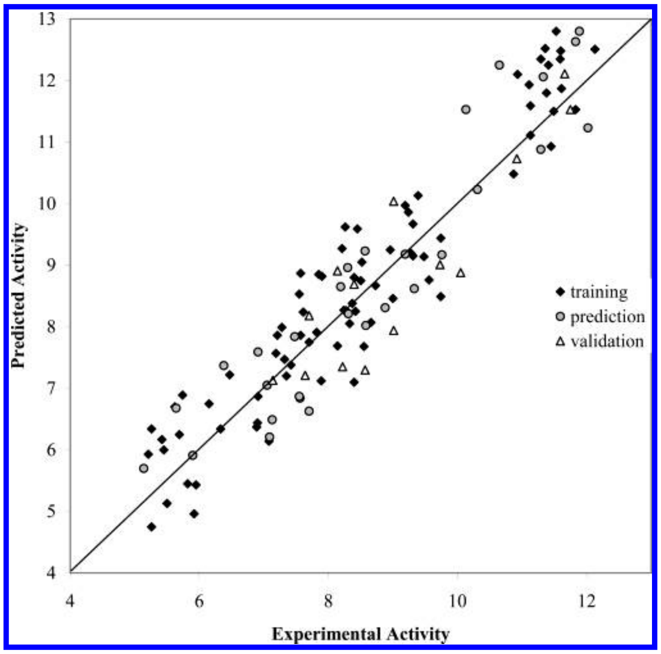

Figure 2. Plot of activity of DHPs predicted by PC-ANN against their corresponding experimental values.

3.2. PC-GA-ANN. As noted previously, not all of the information content of each PC is informative for QSAR

Table 5. Results of PCA on Each Descriptor Group

no.	descriptor type	no. of PCs	variance explained (%)
1	constitutional	3	98.57
2	topological	3	98.05
3	molecular walk counts	2	99.52
4	Burden's descriptors	10	97.41
5	charge topological	2	98.90
6	autocorrelation	3	96.94
7	molecular profile	1	99.31
8	3-dimensional and geometrical	1	99.59
9	VHIM	2	99.82
10	chemical	3	97.76

Table 6. PCs Selected by GA as Input of ANN Models

model	selected PCs
model 1	PC_{11}, PC_{13}, PC_{21}, PC_{32}, PC_{41}, PC_{42}, PC_{45}, PC_{51}, PC_{52}, PC_{62}, PC_{63}, PC_{81}, PC_{92}, PC_{102}
model 2	PC_{12}, PC_{21}, PC_{22}, PC_{32}, PC_{42}, PC_{51}, PC_{52}, PC_{63}, PC_{71}, PC_{81}, PC_{92}, PC_{102}

modeling. However, by decomposition of each PC to other PCs, there will be a chance to separate informative and noninformative parts of the PCs. Thus, in this section, the descriptors data matrix was divided into 10 submatrices. Each submatrix contained an individual group of descriptors (i.e. topological, constitutional, chemical, and so on). The PCA was then performed on each submatrix separately, and the PCs of each descriptor group were selected. The results of performing PCA on each descriptor group are represented in Table 5. In this table, the number of factors used for each group and the corresponding percent variations covered by these factors are indicated. As is obvious, in most cases, the number of PCs is 1−3 and only one group (i.e., Burden's eigenvalue) uses 10 PCs. Here, the 685 descriptors have been reduced to 30 principal components. Although these PCs are different from those obtained in the previous section, their information content about the original data matrix is the same. It can be considered that the latter is a linear combination of the formers. In the other words, the information contents of each PC calculated in the previous section is decomposed to some different PCs in this section.

The PCs extracted in this section were then used as the input of ANN model. To obtain a robust and accurate model and consider only the informative PCs, the ANN was trained with the subset of PCs instead of all calculated PCs. In this way, the neural network could train the set of PCs selected by GA. To find models with lower errors, the GA-ANN algorithm was run many times. Two models with low training and testing root-mean-square errors (i.e., low fitness) were finally obtained. The first model has 12 selected inputs and the second model uses 10 inputs. The selected PCs for both models are shown in Table 6. In this table, PC_{ij} means the *j*th PC of the *i*th descriptor group as indicated in Table 5 (e.g., PC_{23} means that the third PC of descriptor group number 2 is selected). As shown, the selected PCs are not based on the eigenvalue ranking. For example, the first and third PCs of the first descriptors groups are selected, and the second one is not considered in the model. In addition, the second and third PCs of descriptor group number 6 and only the second PC of descriptor group number 9 are used.

For both models, the fitness function (η) was decreased by increasing the number of nodes in the hidden layer. This trend was observed up to four nodes for model 1 and five nodes for model 2. Further increase in the number of hidden layer's nodes resulted in no considerable improvement in the fitness for both models. The optimized PC-GA-ANN model was used to predict the activity of molecules of validation set. The predicted values of Log(1/IC_{50}) for training, validation, and prediction sets by model 2 is shown in Table 1. The statistical parameters of the two models studied are also included in Table 4. The modeling errors of both PC-GA-ANN models are low. The relative standard errors of prediction (RSEP) are 4.61% and 5.27% for model 1 and model 2, respectively. Despite the fact that model 1 uses more PCs than model 2, the predictive ability of model 2 is better that of model 1. A plot of the predicted activities against the experimentally derived values is shown in Figure 3 for model 2.

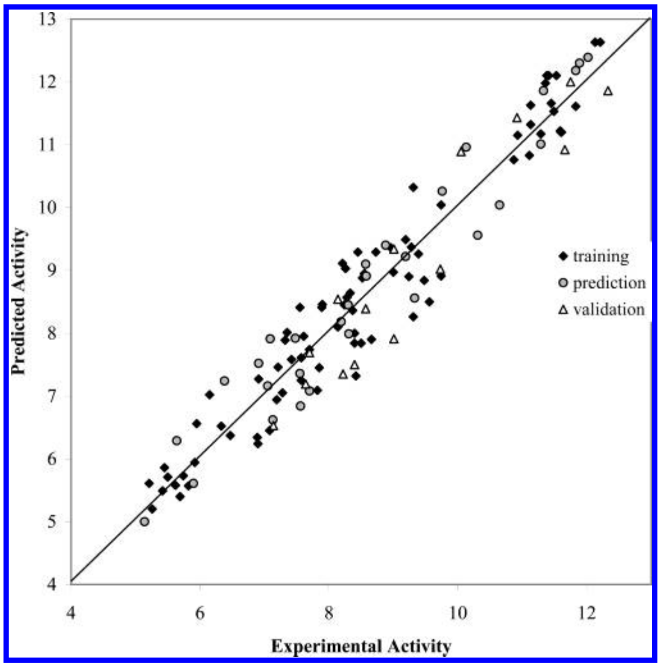

Figure 3. Plot of activity of DHPs predicted by PC-GA-ANN against their corresponding experimental values.

A comparison between the PC-ANN and the PC-GA-ANN models (Table 4) reveals that the latter models possess improved modeling ability toward calcium channel antagonist activity of dihydropyridine derivatives over the former one. The RSE of PC-ANN (6.85) is much more than that for the PC-GA-ANN models. In addition, a comparison between Figures 2 and 3 indicates the higher scattering of the data in Figure 2 (resulted from the PC-ANN) relative to Figure 3, which was derived from the PC-GA-ANN model. Thus, we proposed that the selection of PCs by genetic algorithm technique is more efficient than the eigenvalue ranking method.

Our further study for the use of GA for selection of the best set of PCs generated in the PC-ANN section failed because the best model selected by eigenvalue ranking and removing any of the eight selected PCs from the data set caused a large increase in ANN modeling error. This observation indicated that all of the eight PCs selected in the previous section have some information contents about the calcium channel antagonist activity. However, as shown in this section, by dividing the eight PCs into 30 PCs only the informative contents of the former PCs were used in the ANN model and a better performance was obtained.

4. CONCLUSION

We presented here the introduction of genetic algorithm as a new factor selection method into the PC-ANN technique. The algorithm was applied to the calcium channel antagonist activity of a set of recently synthesized 1,4-dihydropyridine derivatives (124 compounds) that are well-known as calcium channel blockers. To do this, 10 groups of theoretical descriptors (837 descriptors) were generated for each molecule. The PCA was run on each group, and the first more significant principal components (with nonzero eigenvalue) of each group were identified. These PCs were used as input of the neural network model. The best set of PCs which give the best-fitted models were selected by genetic algorithm. Two models with low prediction errors were obtained. The result of this algorithm was compared with eigenvalue ranking factor selection method. The results indicate that the former models give lower modeling errors relative to the routine PC-ANN model.

REFERENCES AND NOTES

(1) Wolowyk, M. W.; Knaus, E. E. *Calcium Channel Modulators in Heart and Smooth Muscle: Basic Mechanisms and Pharmacological Aspects*; Abraham, S., Amital, G., Eds.; VCH: 1990.

(2) Fossheim, R. Crystal structure of the dihydropyridine calcium antagonist felodipine. Dihydropyridine binding prerequisites assessed from crystallographic data. *J. Med. Chem*. **1986**, *29*, 305.

(3) Scholz, G. H.; Vieweg, S.; Uhlig, M.; Thormann, M.; Klossek, P.; Goldmann, S.; Hofmann, H. J. Inhibition of Thyroid Hormone Uptake by Calcium Antagonists of the Dihydropyridine Class. *J. Med. Chem.* **1997**, *40*, 1530.

(4) Norrington, F. E.; Hyde, R. M.; Williams, S. G.; Wooton, R. Physicochemical-activity relations in practice. 1. Rational and self-consistent data bank. *J. Med. Chem*. **1975**, *18*, 604.

(5) Mahmoudian, M.; Richards, G. W. QSAR study of dihydropyridine-type calcium antagonists to their receptor on ileal smooth muscle preparations. *J. Pharm. Pharmacol.* **1986**, *38*, 372.

(6) Rovnyak, G.; Anderson, N.; Gougoutas, J.; Hedberg, A.; Kimball, S. D.; Malley, M.; Moreland, S.; Porubcan, M.; Pudzianowski, A. Active conformation of 1,4-dihydropyridine calcium entry blockers. Effect of size of 2-aryl substituent on rotameric equilibria and receptor binding. *J. Med. Chem.* **1991**, *34*, 2521.

(7) Cobrun, R. A.; Weirzba, M.; Suto, M. J.; Solo, A. J.; Triggle, A. M.; Triggle, D. J. 1,4-Dihydropyridine antagonist activities at the calcium channel: a quantitative structure−activity relationship approach. *J. Med. Chem.* **1988**, *31*, 2103.

(8) Costa, M. C. A.; Gaudio, A. S.; Takahata, Y. A comparative study of principal component and linear multiple regression analysis in SAR and QSAR applied to 1,4-dihydropyridine calcium channel antagonists (nifedipine analogues). *J. Mol. Struct. (THEOCHEM)* **1997**, *394*, 291.

(9) Gaudio, A. S.; Korolkovas, A.; Takahata, Y. Quantitative Structure−Activity Relationships for 1,4-Dihydropyridine Calcium Channel Antagonists (Nifedipine Analogues): A Quantum/Classical Approach. *J. Pharm. Sci.* **1994**, *83*, 1110.

(10) Viswanadhan, V. N.; Mueller, G. A.; Basak, S. C.; Weinstein, J. N. Comparison of a Neural Net-Based QSAR Algorithm (PCANN) with Hologram- and Multiple Linear Regression-Based QSAR Approaches: Application to 1,4-Dihydropyridine-Based Calcium Channel Antagonists *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 505.

(11) Hemmateenejad, B.; Miri, R.; Akhond, M.; Shamsipur, M. QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. An application of geneticalgorithm for variable selection in MLR and PLS mehods. *Chemometr. Intell. Lab. Syst.* **2002**, *64*, 91.

(12) Khanna, T. *Fundations of Neural Networks;* Addison-Wesley: New York, 1991.

(13) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533.

(14) Manallak, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.* **1994**, *37*, 3758.

(15) Peterson, K. L. Quantitative Structure−Activity Relationships in Carboquinones and Benzodiazepines Using Counter-Propagation Neural Networks. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 896.

(16) Shamsipur, M.; Hemmateenejad, B.; Akhond, M. multicomponent acid−base titration using principal component-artificial neural network calibration. *Anal. Chim. Acta* **2002**, *461*, 147.

(17) Agantonovic-Kustrin, S.; Tucker, I. G.; Zecevic, M.; Zivanovic, L. Prediction of drug transfer into human milk from theoretically derived descriptors. *J. Anal. Chim. Acta* **2000**, *418*, 181.

(18) Gemperline, P. J.; Long, J. R.; V. Gregoriou, G. Nonlinear multivariate calibration using principal components regression and artificial neural networks. *Anal. Chem.* **1991**, *63*, 2313.

(19) Malinowski, E. R. Determination of the number of factors and the experimental error in a data matrix. *Anal. Chem.* **1977**, *49*, 612.

(20) Xie, Y. L.; Kalivas, J. H. Evaluation of principal component selection methods to form a global prediction model by principal component regression. *Anal. Chim. Acta* **1997**, *348*, 19.

(21) Sutter, J. M.; Kalivas, J. H. Which principal components to utilize for principal component regression. *J. Chemometr.* **1992**, *6*, 217.

(22) Sun, J. A correlation principal component regression analysis of NIR data. *J. Chemometr.* **1995**, *9*, 21.

(23) Depczynski, U.; Frost, V. J.; Molt, K. Genetic algorithms applied to the selection of factors in principal component regression. *Anal. Chim. Acta* **2000**, *420*, 217.

(24) Alsberg, B. K.; Marchand-Geneste, N.; King, R. D. A new 3D molecular structure representation using quantum topology with application to structure−property relationships. *Chemom. Intell. Lab. Sys.* **2000**, *54*, 75.

(25) Jouanrimbaud, D.; Massart, D. L.; Leardi, R.; deNoord, O. E. Genetic Algorithms as a Tool for Wavelength Selection in Multivariate Calibration. *Anal. Chem.* **1995**, *67*, 4295.

(26) Lucasius, C. B.; Beckers, M. L. M.; Kateman, G. Genetic algorithms in wavelength selection: a comparative study. *Anal. Chim. Acta* **1994**, *286*, 135.

(27) Shafiee, A.; Miri, R.; Dehpour, A. R.; Solimani, F. Synthesis and calcium channel antagonist activity of nifedipine analogues containing nitroimidazolyl substituent in guinea-pig ileal smooth muscle. *Pharm. Sci.* **1996**, *2*, 541.

(28) Miri, R.; Dehpour, A. R.; Azimi, M.; Shafiee, A. Synthesis and smooth muscle calcium channel antagonist effect of alkyl, aminoalkyl 1,4-dihydro-2,6-dimethyl-4-nitroimidazole-3,5 pyridine dicarboxylates. *Daru, J. Fac. Pharm., Tehran Univ. Med. Sci.* **2001**, *9*, 40.

(29) Miri, R.; Niknahad, H.; Vesal, Gh.; Shafiee, A. Synthesis and calcium channel antagonist activities of 3-nitrooxyalkyl, 5-alkyl 1,4-dihydro-2,6-dimethyl-4-(1-methyl-5-nitro-2-imidazolyl)-3,5-pyridinedicarboxylates. *IL FARMACO* **2002**, *57*, 123.

(30) Miri. R.; Howlett, S. E.; Knaus, E. E. Synthesis and calcium channel modulating effects of isopropyl 1,4- dihydro- 2,6- dimethyl-3-nitro-4-(thienyl)-5-pyridinecarboxylates. *Arch. Pharm. Pharm. Med. Chem.* **1997**, *330*, 290.

(31) Miri. R.; McEwen, C. A.; Knaus, E. E. Synthesis and calcium channel modulating effects of modified Hantzsch nitrooxyalkyl 1,4-dihydro-2,6-dimethyl-3-nitro-4-(pyridinyl or 2-trifluoromethylphenyl)-5-pyridinecarboxylates. *Drug Dev. Res.* **2000**, *51*, 225.

(32) Miri, R.; Niknahad, H.; Vazin, A.; Azarpira, A.; Shafiee, A. Synthesis and smooth muscle calcium channel antagonist effects of new derivatives of 1,4-dihydro pyridine containing nitroimidazolyl substituent. *Daru, J. Fac. Pharm., Tehran Univ. Med. Sci.* **2002**, *10*, 130.

(33) Shafiee, A.; Dehpour, A. R.; Hadizadeh, F.; Azimi, M. Syntheses and calcium channel antagonist activity of nifedipine analogues with methylsulfonylimidazolyl substituent. *Pharm. Acta Helv.* **1998**, *73*, 75.

(34) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(35) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis;* RSP-Wiley: Chichester, UK, 1986.

(36) Kostantinora, E. V. Exploring Functional Group Transformations on CASREACT. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 54.

(37) Rucker, G.; Rucker, C. Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683.

(38) Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 520.

(39) Broto, P.; Moreau, G.; Vandicke, C. Molecular structures: perception, autocorrelation descriptor and QSAR studies. System of atomic contributions for the calculation of the *n*-octanol/water coefficients. *Eur. J. Med. Chem.* **1984**, *19*, 66.

(40) ACD Lab Inc., http://www.acdlabs.com.

(41) Todeschini, R. *Milano Chemometrics and QSAR group*; http://www.disat.unimib.it/vhm.

(42) Hypercube Inc., http://www.hyper.com.

CI025661P