Toward an Optimal Procedure for PC-ANN Model Building: Prediction of the Carcinogenic Activity of a Large Set of Drugs

Bahram Hemmateenejad,*[,‡] Mohammad A. Safarpour,[‡,†] Ramin Miri,[‡] and Nasim Nesari[‡]

Medicinal & Natural Products Chemistry Research Center, Shiraz University of Medical Sciences,
Shiraz, Iran, and Chemistry Department, School of Basic Science, Persian Gulf University, Boushehr, Iran

Received July 27, 2004

The performances of the three novel QSAR algorithms, principal component-artificial neural network modeling method combining with three factor selection procedures named eigenvalue ranking, correlation ranking, and genetic algorithm (ER-PC-ANN, CR-PC-ANN, PC-GA-ANN, respectively), are compared by application of these model to the prediction of the carcinogenic activity of a large set of drugs (735 drugs) belonging to a diverse type of compounds. A total number of 1350 theoretical descriptors are calculated for each molecule. The matrix of calculated descriptors (with 735 × 1350 dimension) is subjected to PCA. 95% of the variances in the matrix are explained by the first 137 principal components (PC's). From the pool of 137 PC's, the factor selection methods (ER, CR, and GA) are employed to select the best set of PC's for PC-ANN modeling. In the ER-PC-ANN, the PC's are successively entered into the ANN based on their decreasing eigenvalue. In the CR-PC-ANN, the ANN is first employed to model the nonlinear relationship between each one of the PC's and the carcinogen activity separately. Then, the PC's are ranked based on their decreasing correlating ability and entered to the input layer of the network one after another. Finally, a search algorithm (i.e. genetic algorithm) is used to find the best set of PC's. Both the external and cross-validation methods are used to validate the performances of the resulting models. One is able to see that the results obtained by the PC-GA-ANN and CR-PC-ANN procedures are superior to those resulted from the EV-PC-ANN. Comparison of the results reveals that the results produced by the PC-GA-ANN algorithm are better than those produced by CR-PC-ANN. However, the difference is not significant.

1. INTRODUCTION

Quantitative structure−activity relationships (QSAR), mathematical equations relating chemical structure to their biological activity, are a major factor in contemporary drug design.[1] At present, one cannot talk about drug design without mentioning QSAR.[2] The first assays represented by Hansch et al. (1962), which demonstrate that the biological activity of a molecule can be quantitatively linked to some molecular parameters, introduced QSAR to medicinal chemistry.[3] The QSAR approach attempts to find consistent relationships between the variations in the values of molecular properties (molecular descriptors) and the biological activity. The results of QSAR studies not only are used to predict the biological activity of new compounds but also by selecting chemically relevant descriptors, they give some information about the mechanism of drug action.

Like the other data mining approaches, QSAR is performed in successive steps including data preparation, data reduction, model building, and model evaluation. Clearly, the quality of performance of each step affects the quality of the resulting QSAR. In QSAR studies, a regression model of the form $\mathbf{y} = f(\mathbf{X}) + \mathrm{e}$ is used to describe a set of predictor variables ($\mathbf{X}$) with a predicted variable ($\mathbf{y}$). The relationship between the dependent and independent variables is described by a function $f(\mathbf{X})$, which may be linear or nonlinear. In the linear model, a regression equation of the type $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathrm{e}$ may be used, where the regression coefficient vector is calculated by different algorithms such as multiple linear regression (MLR) and principal component analysis (PCA)-based regression methods, which include principal component regression (PCR) and partial least squares (PLS).[4]

Because of the complexity of the relationships existing between the activity of the molecules and the structures, nonlinear modeling methods are often used to model the structure−activity relationships. To deal with nonlinear behaviors, different algorithms have been proposed, and among them artificial neural networks have found much popularity in the QSAR studies.[5] Artificial neural networks (ANN) are nonparametric nonlinear modeling techniques that have attracted increasing interest in recent years.[6−8] Nonlinear multivariate maps use a nonlinear transformation of the input variable space to project inputs onto the designated attribute values in output space. The strength of modeling with layered, feed-forward artificial neural networks lies in the flexibility of the distributed soft model defined by the weight of the network. Both linear and nonlinear mapping functions may be modeled by suitably configuring the network. Multilayer feed-forward neural network trained with back-propagation learning algorithm becomes increasingly popular techniques.[9−12] The flexibility of ANN for discovering more complex relationships has made this method find wide application in QSAR studies, which was recently reviewed by Schneider and Wrede.[6]

* Corresponding author phone: 98-711-230-3872; fax: 98-711-233-2225; e-mail: hemmatb@sums.ac.ir.

[†] Persian Gulf University.

[‡] Shiraz University of Medical Sciences.

10.1021/ci049766z CCC: $30.25
Published on Web 12/09/2004

Molecular descriptors define the variation in the molecular structure and physicochemical properties of molecules by a single number. A wide variety of molecular invariants including topological, constitutional, electronic, empirical, and chemical descriptors has been reported for use in QSAR analysis.[13] The use of numerous descriptors that are indicative of molecular structure and topology has become more common in QSAR. These types of descriptors, potentially numbering in the thousands, are easily calculated from molecular structures. For example, the Dragon software, designed by the Milano Chemometrics Research Group,[14] can calculate more than 1300 descriptors for a molecule in a few seconds.

However, as the number of descriptors (variables) increases, the model becomes complicated, and its interpretation is difficult if many variables are used in modeling. Therefore, the application of these techniques usually requires variable selection for building well-fitted models. Thus, to obtain robust and accurate models, the ANN models, like other modeling methods, should be trained by a subset of descriptors instead of all generated descriptors.[15−17] There exist two ways of reducing the descriptors space. The first one is to select the features with respect to their generalization ability, which is called *feature selection* (FS).[18−23] The other alternative is to extract features by building linear and nonlinear combinations of a lower dimension of the input features, which is called *feature extraction* (FE).[24−26] The former give models that are simple to interpret, however, in the presence of a large number of descriptors, the ANN modeling becomes complex and time-consuming. In contrast, the latter extracts the information contents of the original descriptors into new variables by simple algorithms such as PCA.[27] The potential usefulness of the FE is that the information from a large number of descriptors is extracted to a few numbers of new variables by using simple algorithms. This decreases the model complexity and computation time. Nowadays, the combined feature selection-extraction or feature extraction-selection (FS-FE and FE-FS, respectively) approaches have also been reported.[28−33] In the FS-FE method, the most relevant set of descriptors is first selected by different algorithms such as genetic algorithm or simulated annealing, and subsequently, a dimension reduction method such as PCA is used to extract the information contents of the selected variables into a space with a lower dimension. In the latter combined approach, the variables (descriptors) are first subjected to FS (by PCA), and then a selection procedure is used to select the most relevant set of the extracted features (PC's).

The main problem, which arises from all of the PCA-based algorithms, is how many and which PC's constitute a good subset for predictive purposes. This is due to the fact the information contents of some extracted features (PC's) may not be in the same direction of the activity data. Therefore, different methods have been addressed to select the significant PC's for calibration purposes.[30−35] The simplest and most common one is a top-down variable selection where the factors are ranked in the order of decreasing eigenvalues (eigenvalue ranking, ER). The factors with the highest eigenvalue are considered as the most significant ones, and, subsequently, the factors are introduced into the calibration model until no further improvement of the calibration model is obtained. However, the magnitude of an eigenvalue is not necessarily a measure of its significance for the calibration. This procedure is currently used in both the PCR and PCA-ANN models. In another method, called correlation ranking (CR), the factors are first ranked according to their correlation coefficient with the activity and then are selected by the procedure discussed for eigenvalue ranking. The most satisfying results are often achieved by this method.[32,35] Very recently, search algorithms such as genetic algorithm (GA) have been applied for the selection of variables in PCR.[33,34] Previously, we proposed a GA procedure for factor selection in the PC-ANN model (PC-GA-ANN algorithm) and found that this procedure produced superior results in comparison with the ER procedure.[36−38] The PC-GA-ANN algorithm, similar to the other GA-based ANN models, is a complex procedure. Therefore, we proposed the correlation ranking procedure, which is simpler than PC-GA-ANN, for the factor selection in the PC-ANN model.[39] The algorithm was employed on the QSAR study of two biological activity data sets including the carcinogenic activity of 60 organic solvents and blood-brain barrier partitioning of 115 diverse organic compounds. In this research, we will report the results of our extensive work on the factor selection in PC-ANN modeling to compare the performances of the three existed factor selection methods by using a big data set, containing the carcinogenic activity of 735 drugs and 1350 calculated descriptors.

The successful development of a new drug depends on a number of criteria that have to be met. For example, in addition to intrinsic activity, the drug must be able to reach to its target and should not produce toxic effects. The significant failure rate of a drug candidate in the late stage development derives the need for predictive tools that eliminate inappropriate compounds before investing substantial time and money in testing.[40] The use of ADME/Tox (**A**bsorption, **D**istribution, **M**etabolism, **E**xcretion, and **T**oxicity) properties is becoming increasingly important in drug discovery and development.[40−45] One of the fundamental criteria in drug design is the drug toxicity. Prediction of the genotoxicity of drugs, pesticides, and natural products is an important area of research in contemporary toxicology.[46] Carcinogenicity of drugs becomes a problem when a drug is used for long periods. This may be specifically so when the drug is to be used for prolonged local application. The chemical structure of a drug may also make one suspect its carcinogenic potency. From existing knowledge of prolonged toxicity studies in animals may reveal some changes in cell nuclei which may raise suspicion. Under such circumstances, studies have to be undertaken to check the carcinogenic potential of new drugs.[47] Short-term tests (STT) have been developed to assess the potential carcinogenic hazard of chemicals to humans. QSAR is a convenient and successful strategy for prediction of the carcinogenic activity of a drug.[48,49] The carcinogenic activity of the 735 drugs belonging to a wide variety of drug families is calculated by using the designed PC-ANN modeling methods. It is obtained that CR-PC-ANN and PC-GA-ANN can predict the carcinogenic potency of the drugs, accurately.

2. MATERIALS AND METHODS

2.1. Activity Data and Descriptors. Galvez gathered the carcinogenesis activity in the DF scale (DF_{carc}) of an

Table 2. Type of Descriptors Used in the Study

descriptor type	molecular descriptors	no. of descriptors
constitutional	molecular weight, no. of atoms, no. of non-H atoms, no. of bonds, no. of heteroatom, 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	47
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	255
molecular walk counts	molecular walk counts of order 1−10, self-returning of order 1−10	21
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	21
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	293
molecular profile indices	Randic molecular profile no. 1−20, Randic shape profile no. 1−20	41
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,	218
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
radial distribution function	unweighted radial distribution function descriptors, radial distribution function descriptors weighted by van der Waals volumes, radial distribution function descriptors weighted by atomic masses, radial distribution function descriptors weighted by atomic electronegativity, radial distribution function descriptors weighted by atomic polarizability	150
functional group and atom-centered group	numbers of different types of carbons, number of allenes groups, number of esters (aliphatic or aromatic), number of amides, number of different functional groups, number of CH3R, number of CR4, number of CR2 × 2, number of different halgenes attached to different type of carbons, number of PX3, number of PR3 and ...	241
empirical and chemical descriptors	unsaturation index, hydrophilic factor, aromatic ratio, LogP, polarizability, density, molar refractivity, parachor, surface polar surface area	6

extensive set of organic compounds from the Merck index based on the annual report of carcinogenesis.[50,51] A total of 735 molecules belonging to diverse types of drugs are selected from this list and used in this study (Table 1, Supporting Information).

A wide variety of descriptors have been reported in the literature for use in the QSAR analysis.[13] There is a recently increased use of theoretical descriptors in QSAR studies. In this study, a total of 1355 molecular descriptors belonging to 12 different types of theoretical descriptors are calculated for each molecule. The chemical structure of the molecules is drawn into a computer by the Hyperchem (Ver. 7, Hypercube Inc.) software and saved by the HIN extension. No geometry optimization is operated. Then, the DRAGON software[14] is used to calculate the descriptors. A brief description of the types of descriptors calculated by DRAGON is represented in Table 2.

2.2. Principal Component Analysis (PCA). The calculated descriptors are first analyzed for the existence of the constant or near constant variables, and then those detected are removed. In addition, to decrease the redundancy existing in the descriptors data matrix, the correlation of descriptors with each other and with the activity ($\mathrm{DF_{carc}}$) of the molecules are examined, and collinear descriptors (i.e. $r > 0.9$) are detected. Among the collinear descriptors, one that has the highest correlation with activity is retained, and the others are removed from the data matrix. Then, the remaining descriptors are collected in an ($n \times m$) data matrix (**D**), where n and m are the number of compounds and the number of descriptors, respectively. Before statistical analysis, the descriptors are scaled to zero mean and unit variance (autoscaling procedure). Among the 735 drugs used in this study, 400 of them are randomly selected as calibration (or training) samples. The remainders are used as a validation set (200 samples) and a *test set* (100 samples) to evaluate the performances of the resulting models in the calibration step and to examine the performances of the resulting models, respectively.

The training data matrix of descriptors ($\mathbf{D}_\mathrm{t}$) is subjected to principal component analysis (PCA) using the singular value decomposition procedure (SVD)[27]

$$\mathbf{D}_\mathrm{t} = \mathbf{U}_\mathrm{t}\mathbf{S}_\mathrm{t}\mathbf{V}_\mathrm{t}^{\mathrm{T}} \tag{1}$$

where $\mathbf{U}_\mathrm{t}$ and $\mathbf{V}_\mathrm{t}$ are the orthonormal matrices spanning the respective row and column spaces of the data matrix ($\mathbf{D}_\mathrm{t}$). $\mathbf{S}_\mathrm{t}$ 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 validation and *test* sets were calculated by the following equation:

$$\mathbf{U}_{v/p} = \mathbf{D}_{v/p}\mathbf{S}_{t/p}^{-1}\,\mathbf{V}_{t/p} \qquad (2)$$

The PC's, which can explain more than 95% of variances in the original descriptors data matrix, are selected and used as the input variables of the ANN models based on the following procedures.

2.3. Neural Networks. A feed-forward neural network with back-propagation of an error algorithm was constructed to model the structure−activity relationship.[5−12] Our network has one input layer, one hidden layer, and one output layer. The input vectors are the set of PC's, selected by three different procedures. The number of nodes in the input layer is dependent on the number of PCs introduced in the network. A bias unit with a constant activation of unity is connected to each unit in the hidden and output layers. The ANN models are confined to a single hidden layer because the network with more than one hidden layer is harder to train. The number of nodes in the hidden layer is optimized through a learning procedure. There is only one node in the output layer. For each descriptor subset, the best topology of the ANN's was searched by using the training and validation data sets. The validation set is used to monitor the overall performances of the trained network. Once the best topology of the network is obtained and the convergence criterion is reached, a leave-20-out cross-validation procedure[20,52] is also employed to more validate the performances of the resulted networks. The root-mean-square errors of the training, validation, and cross-validation ($\mathrm{RMSE_T}$, $\mathrm{RMSE_V}$, and $\mathrm{RMSE_{C-V}}$, respectively) and the corresponding correlation coefficients (R^2_{T}, R^2_{V}, and $R^2_{\mathrm{C-V}}$, respectively) have been monitored during the training of the networks.

2.4. Eigenvalue Ranking Procedure (EV-PC-ANN Model). In this procedure, the extracted PC's are ranked by their decreasing eigenvalues and entered to the ANN one after another. Once each new PC is entered, during the training procedure the network architecture and parameters are optimized to obtain better performances.

2.5. Correlation Ranking Procedure (CR-PC-ANN Model). Since in the ANN a specific hard model is not assumed between the input and output variables, the determination of the correlation between these two types of variables is a difficult task. Here, different ANN models are built for each PC separately, so that each network has a single variable (one PC) in its input layer. By the procedure discussed in section 2.3, the networks are trained to model the nonlinear relationship between an individual PC and the $\mathrm{DF_{cars}}$. When the training of the network is stopped, the resulting ANN model is used to predict the activity ($\mathrm{DF_{cars}}$) of the *test* set samples, the square of the correlation coefficient between the predicted and actual activities (R^2_{p}, i.e., the amount of the variances in the $\mathrm{DF_{cars}}$, which can be explain by each PC) is calculated for each PC, and this quantity is used as a measure of the nonlinear correlation ability of each PC. Then, the PCs are ranked in the order of decreasing correlation, and a procedure similar to what was discussed for the EV-PC-ANN method is used.

2.6. Genetic Algorithm (PC-GA-ANN). A genetic algorithm is a problem solving method that uses generic rules such as reproduction, crossover, and mutation to build pseudoorganisms that are then selected on the basis of a fitness criterion to survive and pass information on to the next generation. The GA used here is the same as we used previously.[21−23,36−38] The GA uses a binary bit string representation as the coding technique for a given problem; the presence or absence of a PC in a chromosome is coded by 1 or 0.[53−57] A string is composed of several genes that represent a specific characteristic that should be studied. In the present case, a string is composed of 137 genes, representing the presence or absence of a PC. By encoding various PC's with bit strings, called chromosomes, the initial population is created randomly. The population size is varied between 50 and 250 for different GA runs. Besides, the number of genes with the values of 1 is kept relatively low to have a small subset of descriptors, i.e., the probability of generating 0 for a gene is set greater (at least 65%).

Using the initial population, ANN models are built, and their fitness (predictivity of the model) is computed by the leave-20-out cross-validation procedure based on the root-mean-square errors ($\mathrm{RMSE_{C-V}}$) value. The inverse of $\mathrm{RMSE_{C-V}}$ is considered as a fitness function. The chromosomes with the least numbers of selected descriptors and the highest fitness are marked as informative chromosomes. These chromosomes are kept for natural selecting and crossovering steps to survive in the next generation preferentially.

The operators used in this research are crossover and mutation. In the crossovering procedure, new chromosomes are generated from a pair of randomly selected chromosomes. Many methods have been proposed for the crossovering technique;[58] here, the uniform crossover technique is applied to 10 pairs of chromosomes in each iteration of generation. In the mutation procedure, the binary bit pattern in each chromosome is changed with a small probability. The probability of the application of these operators is varied linearly with generation renewal (0−10% for mutation and 60−90% for crossover). For a typical run, the evolution of the generation is terminated when 90% of the generations reached the same fitness.

3. RESULTS AND DISCUSSION

To evaluate the performances of the three feature extraction-selection methods for use in ANN, a large data set comprising 735 drug and 1355 descriptors is used. The drugs and their corresponding carcinogenic activity are listed in Table 1. As seen, the drugs belong to a wide variety of drugs family. The carcinogenic activities vary between 7.76 to −12.20 with an average and standard deviation equal to −1.78 and 3.21, respectively. To show the distribution of the activities, their histogram-plot is shown in Figure 1. The histogram, which shows a relatively normal distribution of the activity data, indicates that the carcinogenic activity of 85.6% of the drugs is between 2.27 and −5.17. Besides, it is obvious from the histogram that 27% of drugs have a high carcinogenic activity, i.e., their carcinogenic activity ($\mathrm{DF_{cars}}$) is greater than zero. Among these, 6 of them including calcifediol, dihydrotachysterol, estradiol, nandrolone, stanozolol, and timiperone are highly carcinogen.

Earlier QSAR studies have revealed that not only 2-dimensional theoretical descriptors but also 3-dimensional

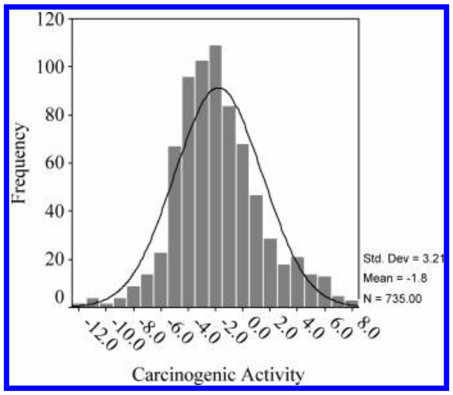

Figure 1. Histogram of the distribution of the carcinogenic activity of the 735 drugs used in this study. The solid curve is the fitting of the activity data to the normal distribution.

properties of molecules such as their electronic features affect the carcinogenesis and mutagenesis activity of some organic compounds.[59−61] In this research, however, only the formers are used because 3-dimensional geometry optimization and calculation of the electronic properties for 735 molecules are tedious. The brief description of the molecular descriptors used in this study is given in Table 2. The descriptors, belonging to 12 different types of molecular descriptors, are simple descriptor and are calculated based on their 2-dimensional structure, rapidly.

The results of the application of PCA on the descriptors data matrix are listed in Table 3. In this table, the logarithm of eigenvalue (logEV) corresponding to each PC, the percent of variances explained by each PC (%V), and the cumulative percent of variances (C%V) are reported. As seen, 95% of variances in the original descriptors are explained by the first 137 extracted PC's. These PC's are used in the feature analyses. It should be noted that at this stage, each PC contains some information about the predictor variables. For example, the PC1 has the maximum information about the descriptors; PC2 has less information and so on. If all of the calculated descriptors have informative variance about the predicted variable (DF_{cars}), it should be concluded that the all of the extracted PC's also contained some information about the carcinogenic activity. However, the descriptors data matrix includes both the informative and uninformative descriptors (i.e. variable selection is not performed before applying the PCA on the data matrix), thus, none of the entire PC's has useful information about the biological activity, and, consequently, three factor selection algorithms are employed to find the best set of PC's.

Table 3. Results of the PCA of the Descriptors Data Matrix

PC no.	logEV	%V	C%V	PC no.	logEV	%V	C%V
1	5.45	28.18	28.18	70	3.23	0.17	87.83
2	4.88	7.62	35.80	71	3.23	0.17	88.00
3	4.86	7.35	43.15	72	3.22	0.17	88.16
4	4.67	4.75	47.90	73	3.21	0.16	88.33
5	4.58	3.79	51.69	74	3.20	0.16	88.49
6	4.44	2.75	54.44	75	3.19	0.16	88.64
7	4.31	2.03	56.47	76	3.19	0.15	88.80
8	4.27	1.87	58.34	77	3.18	0.15	88.95
9	4.19	1.57	59.91	78	3.17	0.15	89.10
10	4.16	1.44	61.34	79	3.16	0.15	89.25
11	4.12	1.32	62.66	80	3.16	0.14	89.39
12	4.08	1.20	63.86	81	3.15	0.14	89.53
13	4.07	1.17	65.03	82	3.14	0.14	89.67
14	4.03	1.07	66.11	83	3.14	0.14	89.81
15	3.99	0.99	67.09	84	3.12	0.13	89.94
16	3.97	0.95	68.04	85	3.12	0.13	90.07
17	3.96	0.92	68.96	86	3.12	0.13	90.20
18	3.92	0.84	69.80	87	3.11	0.13	90.33
19	3.89	0.78	70.58	88	3.10	0.13	90.46
20	3.85	0.71	71.29	89	3.10	0.13	90.59
21	3.84	0.70	71.99	90	3.09	0.12	90.71
22	3.81	0.65	72.63	91	3.09	0.12	90.83
23	3.79	0.62	73.25	92	3.08	0.12	90.95
24	3.77	0.59	73.84	93	3.07	0.12	91.07
25	3.75	0.56	74.41	94	3.06	0.12	91.19
26	3.73	0.54	74.94	95	3.06	0.12	91.30
27	3.71	0.51	75.45	96	3.06	0.11	91.42
28	3.69	0.50	75.95	97	3.05	0.11	91.53
29	3.68	0.49	76.43	98	3.04	0.11	91.64
30	3.67	0.47	76.91	99	3.04	0.11	91.75
31	3.65	0.45	77.36	100	3.03	0.11	91.86
32	3.64	0.44	77.80	101	3.03	0.11	91.96
33	3.63	0.43	78.24	102	3.02	0.10	92.07
34	3.60	0.40	78.64	103	3.01	0.10	92.17
35	3.59	0.39	79.03	104	3.00	0.10	92.27
36	3.58	0.38	79.41	105	3.00	0.10	92.37
37	3.57	0.38	79.79	106	2.99	0.10	92.47
38	3.56	0.36	80.15	107	2.99	0.10	92.57
39	3.54	0.35	80.50	108	2.98	0.09	92.66
40	3.53	0.34	80.84	109	2.97	0.09	92.76
41	3.51	0.32	81.16	110	2.97	0.09	92.85
42	3.49	0.31	81.47	111	2.97	0.09	92.94
43	3.48	0.31	81.78	112	2.96	0.09	93.04
44	3.46	0.29	82.07	113	2.95	0.09	93.13
45	3.45	0.28	82.35	114	2.95	0.09	93.22
46	3.44	0.28	82.63	115	2.95	0.09	93.31
47	3.44	0.27	82.90	116	2.94	0.09	93.39
48	3.43	0.27	83.17	117	2.93	0.09	93.48
49	3.42	0.27	83.44	118	2.92	0.08	93.56
50	3.39	0.25	83.69	119	2.92	0.08	93.64
51	3.39	0.25	83.93	120	2.91	0.08	93.73
52	3.39	0.25	84.18	121	2.91	0.08	93.81
53	3.38	0.24	84.42	122	2.90	0.08	93.89
54	3.37	0.24	84.66	123	2.90	0.08	93.97
55	3.37	0.23	84.89	124	2.89	0.08	94.05
56	3.36	0.23	85.12	125	2.89	0.08	94.12
57	3.34	0.22	85.34	126	2.88	0.08	94.20
58	3.34	0.22	85.56	127	2.88	0.08	94.27
59	3.32	0.21	85.77	128	2.87	0.07	94.35
60	3.31	0.21	85.97	129	2.87	0.07	94.42
61	3.31	0.20	86.18	130	2.87	0.07	94.50
62	3.29	0.20	86.37	131	2.86	0.07	94.57
63	3.29	0.20	86.57	132	2.85	0.07	94.64
64	3.27	0.19	86.76	133	2.84	0.07	94.71
65	3.26	0.18	86.94	134	2.84	0.07	94.78
66	3.26	0.18	87.12	135	2.84	0.07	94.85
67	3.25	0.18	87.30	136	2.83	0.07	94.92
68	3.25	0.18	87.48	137	2.82	0.07	94.99
69	3.24	0.18	87.66				

3.1. ER-PC-ANN. In this procedure, the extracted PC's are consecutively entered to the ANN model based on their decreasing eigenvalue (as shown in the Table 3). In each step, the ANN structure and parameters are optimized to give better performances. The plot of the root-mean-square error of cross-validation and the root-mean-square error for the test set against the number of entered PC's is shown in Figure 2A. Obviously, the RMS errors are decreased, with high fluctuations, as more PC's are added to the network. In the presence of 34 entered PC's, the least RMS values are obtained. The fluctuations observed in the plots of Figure 2A reveal that some PC's (even with high information content about the descriptor) do not have useful information about the carcinogenic activity data, and therefore the addition of these PC's increases the errors. The ANN parameters and statistical quantities of the best resulted ER-PC-ANN model are listed in Table 4. Although the statistical quality of the resulted model is good, because of the

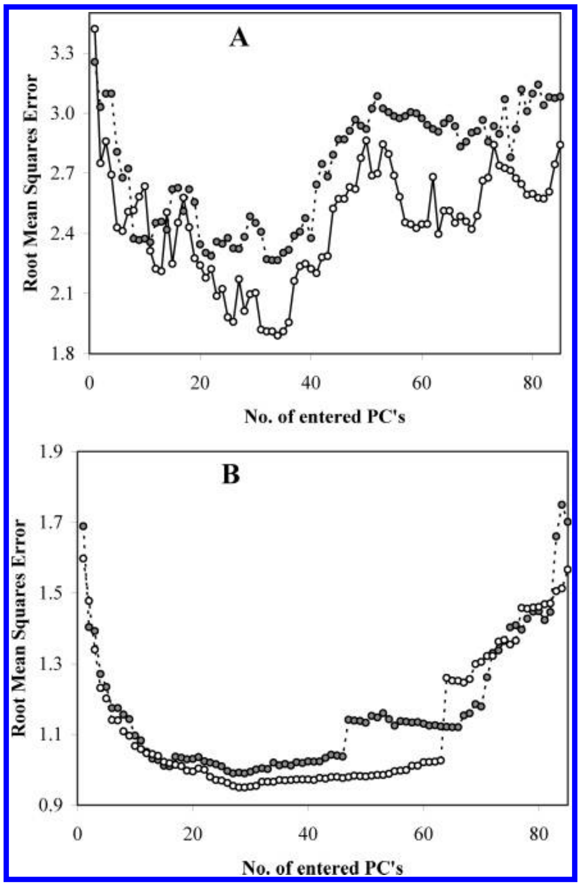

Figure 2. The variation of the root-mean-squared errors as a function of the entered PC's to the PC-ANN model for the (A) eigenvalue ranking and (B) correlation ranking. The filled and open markers refer to the $\mathrm{RMSE}_{\mathrm{C-V}}$ and $\mathrm{RMSE}_{\mathrm{T}}$, respectively.

significant differences between the statistic quantity of the test set and the other sets, it seems that an overfitted model was obtained. This may be due to the incorporation of the uninformative PC's to the model. The values of $\mathrm{DF}_{\mathrm{cars}}$ predicted by the ER-PC-ANN model are included in Table 1 and plotted against the corresponding experimental values in Figure 3. The plots show high scattering of the data around a straight line.

3.2. CR-PC-ANN. As discussed in the previous section, some PC's with a high eigenvalue do not have useful information about the activity data, and, therefore, incorporating these PC's to the model decreases the overall performances of the ANN. To show which PC contains more information about the carcinogenic activity of drugs, a correlation ranking procedure is used. For this purpose, the nonlinear relationship between each one of the PC's and $\mathrm{DF}_{\mathrm{cars}}$ is modeled by the ANN according to the procedure discussed in section 2.5. The resulting correlation coefficient for each PC is plotted in Figure 4. It is obvious from this plot that the PC's with a higher correlation ability are not essentially those which have a higher eigenvalue. On the basis of their decreasing correlation coefficient, the order of PC's is PC7 > PC11 > PC1 > PC44 > PC24 > PC4 > PC19 > PC 39 > PC14 > PC2 > PC32 > PC9 > PC43 > .... The PC's are entered step by step to the ANN model in this order, and in each step the ANN are trained and its structure is optimized to reach to the best performances. The RMS errors obtained by the leave-20-out cross-validation and those obtained for the test samples in the presence of different added PC's are represented in Figure 2B. A gradual decreasing in the errors is observed when the PC's are successively introduced to the ANN model up to 29 PC's. No improvement is obtained when the network is trained with more than 29 PC's. Instead, the RMS errors are increased as more than 29 PC's are used in the input layer of the network. The structure and statistical quantities of the best CR-PC-ANN model are listed in Table 4. Clearly, the PC-ANN model obtained by the correlation ranking procedure possesses high statistical qualities with low RMS errors. The network uses 29 nodes in its input layer (29 PC's) and 6 nodes in its hidden layer. In comparison with the ER-PC-ANN algorithm, the CR-PC-ANN algorithm presents more accurate results. Meanwhile, the closeness of the statistical quantities of the calibration and validation sets indicates the generalization ability of the resulted models. The carcinogenic activities calculated by the resulted CR-PC-ANN model are listed in Table 1 and are plotted against the experimental activities in Figure 5. As seen, the data are distributed around a straight line with low scattering even for the test samples. This reveals the high prediction ability of the resulted model.

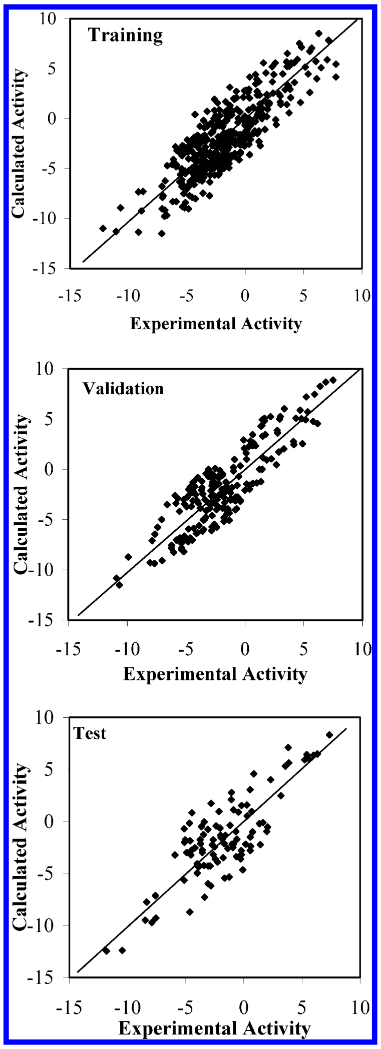

Figure 3. Plot of the activity calculated by the ER-PC-ANN model against the experimental activity for all three types of data set. The solid lines are an ideal fit with the respective intercept and slope equal to zero and one.

3.3. PC-GA-ANN. The PC-GA-ANN algorithm for factor selection is more complex than the other methods. To study the effect of network parameters on its performances, some

Table 4. Neuron Structure and Statistical Quantities for the Different PC-ANN Algorithms

model	N_I	N_H	TF_H*	$RMSE_{C-V}$	$RMSE_V$	$RMSE_T$	R^2_{C-V}	R^2_V	R^2_T
ER-PC-ANN	34	7	S	1.89	1.99	2.27	0.714	0.754	0.689
(selected PC's)	(PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9, PC10, PC11, PC12, PC13, PC14, PC15, PC16, PC17, PC18, PC9, PC20, PC21, PC22, PC23, PC24, PC25, PC26, PC27, PC28, PC29, PC30, PC31, PC32, PC33, PC34)								
CR-PC-ANN	29	6	T	0.95	1.04	0.99	0.911	0.913	0.925
(selected PC's)	(PC7, PC11, PC1, PC24, PC44, PC4, PC19, PC39, PC14, PC2, PC32, PC9, PC43, PC65, PC52, PC59, PC74, PC12, PC129, PC33, PC76, PC112, PC29, PC121, PC92, PC18, PC16, PC96, PC10)								
PC-GA-ANN1	30	7	S	0.879	0.977	1.03	0.922	0.923	0.921
(selected PC's)	(PC7, PC11, PC1, PC24, PC44, PC4, PC39, PC14, PC2, PC9, PC43, PC65, PC52, PC59, PC74, PC12, PC129, PC33, PC76, PC112, PC29, PC121, PC92, PC18, PC16, PC96, PC10, PC6, PC22, PC8)								
PC-GA-ANN2	28	8	S	0.894	0.985	1.05	0.920	0.922	0.918
(selected PC's)	(PC7, PC11, PC1, PC24, PC44, PC4, PC39, PC14, PC2, PC32, PC9, PC43, PC65, PC52, PC59, PC74, PC12, PC33, PC76, PC112,PC121, PC92, PC18, PC16, PC96, PC10, PC6, PC41)								
PC-GA-ANN3	31	5	T	0.880	0.971	1.07	0.921	0.924	0.912
(selected PC's)	(PC7, PC11, PC1, PC24, PC44, PC19, PC39, PC14, PC2, PC32, PC9, PC43, PC52, PC59, PC74, PC12, PC129, PC33, PC76, PC112, PC29, PC18, PC96, PC10, PC5, PC23, PC31, PC13, PC106, PC9)								

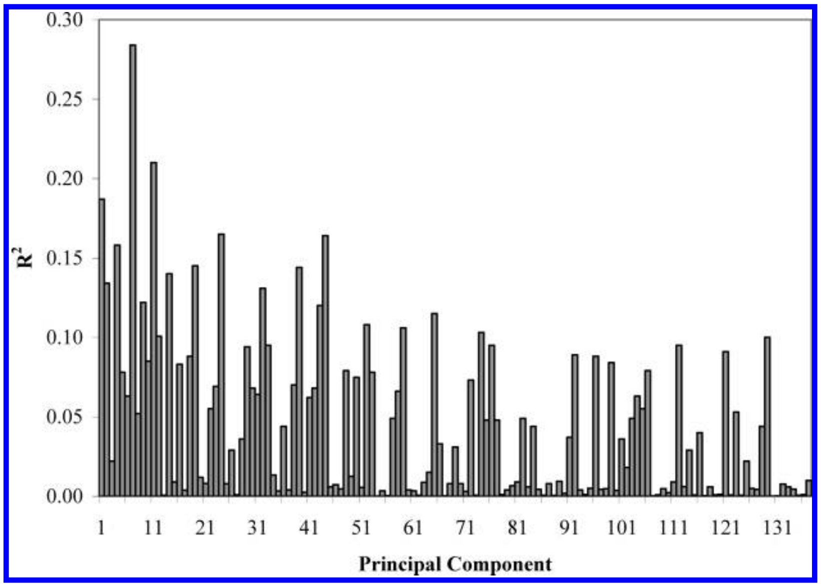

Figure 4. Nonlinear correlation between each one of the PC's and the DF_{cars} obtained by ANN modeling.

networks with different parameters and configurations are built, and GA is used to select the most relevant set of PCs for each network, separately. Thus, in a typical GA run, an ANN model with the same size, the same training and *test* sets, and the same training parameters is used. For each chromosome of the GAs, the neural network with the specified structure is trained to reach to a minimum $RMSE_{C-V}$. Besides, to obtain the best fitted model, the GA has been run many times with different initial set of GA populations. Therefore, a population of good models has been obtained. The structure and statistical quantities of three models, produced better results, are listed in Table 4. It should be noted that one of the final PC-GA-ANN models is the same as what was obtained by the CR-PC-ANN model. It is obvious from the data reported in Table 4 that the PC-ANN models obtained by the GA factor selection method have generated superior results than those of two other factor selection methods. However, the results are very close to the results of CR-PC-ANN model. There is a fair agreement between the statistical quantities of the calibration, validation and test samples, which confirms the lack of overfitting problem in the resulted models. In the tables is also listed the PC's selected by each one of the resulting models. Clearly, the PC's selected by the ER-PC-ANN are very different from those selected by the other methods. While, those selected by the CR-PC-AN and PC-GA-ANN are very similar. These models differ only in a small amount of PC's. The carcinogenic activities calculated by the PC-GA-ANN1, which gave better results, are listed in Table 1 and are plotted in Figure 6 against the experimental ones. Comparison of the plots shown in Figures 5 and 6 shows similar pattern of the scattering of the data around a straight line.

3.4. Testing the Adequacy of the CR and GA Factor Selection Procedures. The results given in the previous sections indicate that selecting a subset of PC's (by methods such as correlation ranking or genetic algorithm), which minimizes the RMS errors, produces better models relative to the models whose PC's are selected by a top-down eigenvalue ranking procedure. On the other hand, some readers may say that GA and CR introduce some low variances PC's, which contain noise, to the models. In other words, correlations to random noise may be found in these methods, while an ER procedure minimizes the chance of using noisy PC's in PC-ANN modeling. However, the descriptors used in this study all are theoretical descriptors, which are calculated for each molecule exactly without any uncertainty. The first 137 PC's used in this study are attributed to the 95% of variances in the descriptors data matrix. Besides, the correlations are obtained not only by using calibration data but also by using test samples.

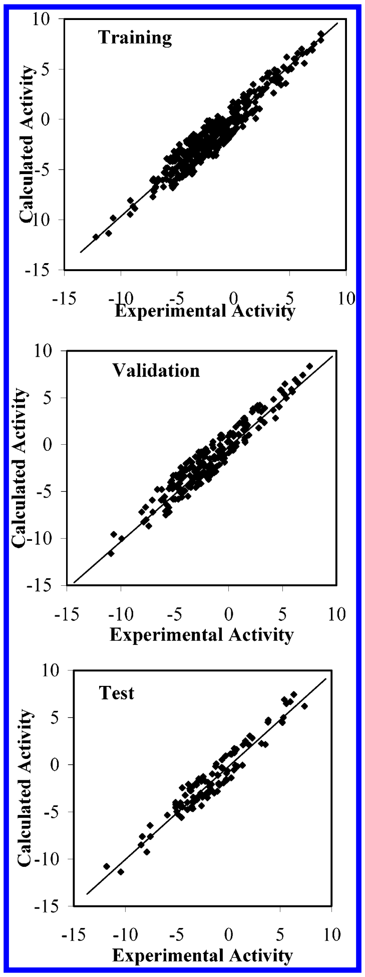

Figure 5. Plot of the activity calculated by the CR-PC-ANN model against the experimental activity for all three types of data set. The solid lines are an ideal fit with the respective intercept and slope equal to zero and one.

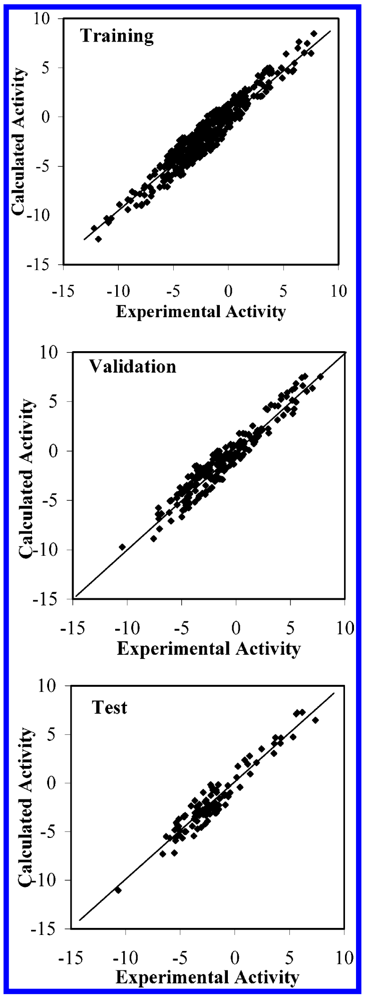

Figure 6. Plot of the activity calculated by the PC-GA-ANN model against the experimental activity for all three types of data set. The solid lines are an ideal fit with the respective intercept and slope equal to zero and one.

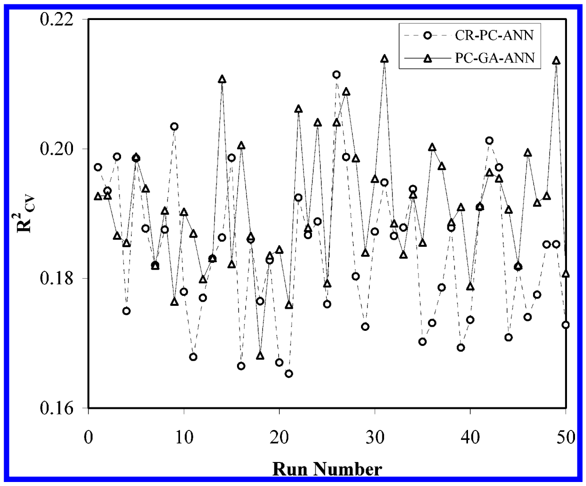

Figure 7. The cross-validated square of correlation coefficient obtained by the chance correlation test at 50 different runs for CR-PC-ANN and PC-GA-ANN models.

Nevertheless, three tests are employed to check the suitability of the CR and GA-based PC-ANN models and to show that the models have not been obtained by chance.

One can assume that the model obtained by the eigenvalue ranking is the optimal model, which minimizes the use of noisy PC's, and in correlation ranking, the low variances PC's (or noisy PC's) are incorporated to the model by chance. To prove or reject this assumption, a chance correlation test is employed. The dependent variable (DF_{carc}) is randomized, and attempts are made to obtain ANN models between the randomized DF_{cars} and 137 selected PC's by the CR and GA methods. This procedure is repeated 50 times for each factor selection method, separately. As shown in Figure 7, low correlation coefficient (even lower than that obtained by ER-PC-ANN model) is obtained for the randomized variables. This indicates that the CR- and GA-based PC-ANN models are not obtained by chance.

In the other test, correlation of each one of the PC's above the PC137 (i.e. PC138−PC237) with the carcinogenic activity data is determined by separate ANN models. The resulting squares of correlation coefficients for all of these extra PC's are lower than 0.05. Table 4 shows that PC10 is the least correlated PC used by the CR-PC-ANN model. The nonlinear squared correlation coefficient of this PC with DF_{carc} is 0.08 (Figure 4). Thus the correlation of the extra PC's is lower than that of the threshold correlation. Hence, this confirms that the CR-PC-ANN and PC-GA-ANN models are not obtained by chance.

In the final test, CR and GA are used to select PC's only from the set of PC's selected by ER (PC1 to PC34). It is found that the CR and GA select the same PCs (i.e. PC7, PC11, PC1, PC4, PC19, PC14, PC2, PC32, PC9, PC12, PC33, PC29, PC18, PC16, and PC10). The performance of the resulting model ($R^2_{C-V} = 0.759$ and $RMSE_{C-V} = 1.71$) is higher than that was found by the ER-PC-ANN model ($R^2_{C-V} = 0.714$ and $RMSE_{C-V} = 1.89$). However, it is not as high as those of the CR-PC-ANN ($R^2_{C-V} = 0.911$ and $RMSE_{C-V} = 0.95$) and PC-GA-ANN models ($R^2_{C-V} = 0.922$ and $RMSE_{C-V} = 0.88$). These results together with the results of the two other tests indicate that the PC-ANN models optimized by CR and GA are reasonable models, and the correlations are not obtained by chance.

4. CONCLUSION

The performances of the three different factor selection procedures including eigenvalue ranking, correlation ranking, and genetic algorithm are evaluated. To do so, the PC-ANN modeling method is applied to predict the carcinogenic activity of 735 drugs using 1350 theoretically derived descriptors. It is found that both genetic algorithm and correlation ranking resulted in more generalized models and can predict the activity of the drugs more accurately. However, the convenient model is not obtained by the eigenvalue procedure. The PC-ANN model obtained by the correlation ranking procedure produces results that are comparable with those obtained by the genetic algorithm. Moreover, the correlation ranking procedure is much simpler than GA.

ACKNOWLEDGMENT

The financial support of this project by the Shiraz University of Medical Sciences Research Council (SUMSRC) is acknowledged.

Supporting Information Available: Drug names and experimental and calculated activities (Table 1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hansch, C.; Leo, A. In *Exploring QSAR: Fundamental and application in chemistry and biology*; Heller S. R., Eds; American Chemical Society: Washington, DC, 1995.

(2) Krogsgaard-Larsen, P.; Liljefors, T.; Madsen, U. *Textbook of drug design and discovery*; Taylor & Francis: London, 2002.

(3) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M. The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammet constants and partition coefficients. *J. Am. Chem. Soc.* **1963**, *85*, 2817−2824.

(4) Polanski, J.; Gieleciak, R. The comparative molecular surface analysis (CoMSA) with modified uninformative variable elimination-PLS (UVE-PLS) Method: Application to the steroids binding the aromatatase enzyme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 656−666.

(5) Zupan, J.; Gasteiger, J. *Neural networks in chemistry and drug design*; Wiley-VCH: Germany, 1999.

(6) Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biolog.* **1998**, *70*, 175−222.

(7) Boger, Z. Selection of quasi-optimal inputs in chemometrics modeling by artificial neural network analysis. *Anal. Chim. Acta* **2003**, *490*, 31−40.

(8) So, S. S.; Karplus, M. Genetic neural networks for quantitative structure−activity relationships: improvement and application of Benzodiazepine affinity for Benzodiazepine/GABA A receptor. *J. Med. Chem.* **1996**, *39*, 5246−5256.

(9) Shamsipur, M.; Hemmateenejad, B.; Akhond, M. Multicomponent acid−base titration by principal component-artificial neural network calibration. *Anal. Chim. Acta* **2002**, *461*, 147−153.

(10) Hossain, A. S.; Yu, X.; Johnson, R. D. Application of neural network computing in pharmaceutical product development. *Pharm. Res.* **1991**, *8*, 1248−1252.

(11) Despagne, F.; Massart, D. L. Neural networks in multivariate calibration. *Analyst* **1998**, *123*, 157R-178R.

(12) Svozil, D.; Kvasnicka, V.; Pospichal, J. Introduction to multilayer feed-forward neural networks. *Chemom. Intell. Lab. Syst.* **1997**, *39*, 43−62.

(13) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors. In methods and principles in medicinal Chemistry*; Mannhold, R., Kubinyi, H., Timmerman, H, Eds.; Wiley-VCH: Weinheim, 2000.

(14) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. http://www.disat.unimib.it/chm/Dragon.htm.

(15) Tetko, I. V.; Vila, A. E. P.; Livingstone, D. J.; Luil, A. I. Neural network studies 2. variable selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794−803.

(16) Waller, C. L.; Bradley, M. P.; Development and validation of a novel variable selection technique with application to multidimensional quantitative structure−activity relationship studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 345−355.

(17) Turner, J. V.; Cutler, D. J.; Spence, I.; Maddalena D. J. Selective descriptor pruning for QSAR/QSPR studies using neural networks. *J. Comput. Chem.* **2003**, *24*, 891−897.

(18) Wegner, J. K.; Fröhlich, H.; Zell, A. Feature selection for descriptor based classification models. 1. Theory and GA-SEC algorithm. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 921−930.

(19) Yasri, A.; Hartsough, D. Toward an optimal procedure for variable selection and QSAR model building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218−1227.

(20) Baumann, K.; Albert, H.; Korff, M. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. part I. search algorithm, theory and simulations. *J. Chemom.* **2002**, *16*, 339−350.

(21) 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 genetic algorithm for variable selection in MLR and PLS methods. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 91−99.

(22) Hemmateenejad, B.; Safarpour, M. A.; Taghavi, F. Application of ab initio theory for the prediction of acidity constants of 1-hydroxy-9,-10-anthraquinone derivatives using genetic neural network. *J. Mol. Struct. (THEOCHEM)* **2003**, *635*, 183−190.

(23) Safarpour, M. A.; Hemmateenejad, B.; Miri, R.; Jamali, M. Quantum chemical-QSAR study of some newly synthesized 1,4-dihydropyridiune calcium channel blockers. *QSAR Comb. Sci.* **2003**, *22*, 997−1005.

(24) 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.

(25) Indahl, U. G.; Naes, T. Evaluation of alternative spectral feature extraction methods of textural images for multivariate modeling. *J. Chemom.* **1998**, *12*, 261−278.

(26) 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.

(27) Malinowski, E. R. *Factor analysis in chemistry*; Wiley-Interscience: New York, 2002.

(28) Xue, L.; Bajorath, J. Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principal Component Analysis Identified by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 801−809.

(29) Gohlke, H.; Dullweber, F.; Kamm, W.; März, J.; Kissel, T.; Klebe, G. Prediction of Human Intestinal Absorption using a combined ‘Simulated Annealing/Back-propagation Neural Network’ Approach. *Rational Approaches Drug Des.* **2001**, 261−270.

(30) 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.

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

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

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

(34) Barros, A. S.; Rutledge, D. N. Genetic algorithm applied to the selection of principal components. *Chemom. Intell. Lab. Syst.* **1998**, *40*, 65−81.

(35) Verdu-Andres, J.; Massart, D. L. Comparison of Prediction-and Correlation-Based Methods to Select the Best Subset of Principal Components for Principal Component Regression and Detect Outlying Objects. *Appl. Spectrosc.* **1998**, *52*, 1425−1434.

(36) Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M. Genetic algorithm applied to the selection of factors in principle component-

artificial neural networks: application to QSAR studu of calcium channel antagonist activity of 1,4-dihydropyridines (nifedipine analogous*). J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1328−1334.

(37) Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Taghavi, F. Application of ab initio theory to QSAR study of 1,4-dihydropyridine-based calcium channel blockers using GA-MLR and PC-GA-ANN procedures. *J. Comput. Chem.* **2004**, *25*, 1495−1503.

(38) Hemmateenejad, B.; Shamsipur, M. Quantitative structure-electrochemistry relationship study of some organic compounds using PC-ANN and PCR. *Internet Electron. J. Mol. Des*. **2004**, *3*, 316−334.

(39) Hemmateenejad, B. Correlation ranking procedure for factor selection in PC-ANN modeling and application to ADMETox evaluation. *Chemom. Intell. Lab. Syst.* **2004**, in press.

(40) Klopman, G.; Stefan, L. R.; Saiakhov, R. D. ADME evaluation: 2. A computer model for the prediction of intestinal absorption in humans. *Eur. J. Pharm. Sci*. **2002**, *17*, 253−63.

(41) Stoner, C. L.; Gifford, E.; Stankovic, C.; Lepsy, C. S.; Brodfuehrer, J.; Vara Prasad, J. V. N.; Surendran, N. Implementation of an ADME enabling selection and visualization tool for drug discovery. *J. Pharm. Sci.* **2004**, *93*, 1131−1141.

(42) Yu, H.; Adedoyin, A. ADME-Tox in drug discovery: integration of experimental and computational technologies. *Drug. Discovery Today* **2003**, *8*, 852−861.

(43) Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery. 2. Prediction of Partition Coefficient by Atom-Additive Approach Based on Atom-Weighted Solvent Accessible Surface Areas. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1058−1067.

(44) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 266−275.

(45) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. *J. Mol. Mod*. **2002**, *8*, 337−349.

(46) Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test. 1: Mutagenicity of Aromatic and Heteroaromatic Amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.* **1992**, *19*, 37−52.

(47) Sheth, U. K. Symposium on Adverse Drug Reactions including Teratogenic and Carcinogenic effects of drugs and Chemicals. *Ind. J. Pharmacol*. **1972**, *4*, 32−34.

(48) Basak, S. C.; Grunwald, G. D. Predicting Mutagenecity of Chemicals Using Topological and Quantum Chemical Parameters: A Similarity Based Study. *Chemosphere* **1995**, *31*, 2529−2546.

(49) Toropov, A. A.; Toropov, A. P. Prediction of Heteroaromatic Amine Mutagenecity by Means of Correlation Weighting of Atom Orbital Graphs of Local Invariants. *J. Mol. Struct. (THEOCHEM)* **2001**, *538*, 287−293.

(50) Budavari, S.; O'Neil, M. J.; Heckelman, P. E. *The Merck Index*; Merck & Co., Inc.: Rahway, NJ, U.S.A., 1989.

(51) www.uv.es/~galvez/tablevi.pdf.

(52) Gramatica, P.; Papa, E. QSAR Modeling of Bioconcentration Factor by theoretical molecular descriptors. *QSAR Comb. Sci.* **2003**, *22*, 374−385.

(53) 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.

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

(55) Browan, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A graph-based genetic algorithm and its application to the multiobjective evaluation of median molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079−1087.

(56) Xiang, Y. H.; Jiang, H. Y.; Cai, W. S.; Shao, X. G. An efficient method based on lattice construction and the genetic algorithm for optimization of large Lennard-Jones clusters. *J. Phys. Chem. A* **2004**, *108*, 3586−3592.

(57) Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J. Prediction of ^{1}H NMR chemical shifts using neural networks. *Anal. Chem.* **2002**, *74*, 80−90.

(58) Hibbert, D. B. Genetic algorithm in chemistry. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277−293.

(59) Benigni, R.; Passerini, L. Carcinogenecity of the aromatic amines: from structure−activity relationships to mechanisms of action and risk assessment. *Mutat. Res.* **2002**, *511*, 191−206.

(60) Lewis, D. F. V.; Parke, D. V. The genotoxicity of benzanthracenes: a quantitative structy-activity study. *Mutat. Res.* **1995**, *328*, 207−214.

(61) Chung, K. T.; Kirkovsky, L.; Kirkovsky, A.; Purcell, W. Review of mutagenecity of monocyclic aromatic amines: quantitative structure−activity relationships. *Mutat. Res.* **1997**, *387*, 1−16.

CI049766Z