Internet Electronic Journal of Molecular Design **2003**, *2*, 000–000
BioChem Press http://www.biochempress.com

Quantitative Structure-Electrochemistry Relationship Study of Some Organic Compounds Using PC-ANN and PCR

Bahram Hemmateenejad,[1] Mojtaba Shamsipur[2,*]

[1] Medicinal & Natural Products Chemistry Research Center, Shiraz University of Medical Science, Shiraz, Iran
[2] Department of Chemistry, Razi University, Kermanshah, Iran

Received xxx; Preprint published xxx; Accepted xxx ; Published xxx

Internet Electron. J. Mol. Des. **2003**, *1*, 000–000

Abstract

Motivation. A QSPR analysis has been conducted on the half-wave reduction potential ($E_{1/2}$) of a diverse set of organic compounds by means of principal component regression (PCR) and principal component-artificial neural network (PC-ANN) modeling method. Genetic algorithm was employed as a factor selection procedure for both modeling methods. The results were compared with two other factor selection methods namely eigen-value ranking (EV) and correlation ranking (CR) procedures.

Method. By using the Dragon software more than 1000 structural descriptors were calculated for each molecule. The descriptor data matrix was subjected to principal component analysis and the most significant principal components (PC) were extracted. Multiple linear regression and artificial neural network were employed for the respective linear and nonlinear modeling between the extracted principal components and $E_{1/2}$. First, the principal components were ranked by decreasing eigen-values and entered successively to each modeling method separately. In addition, the factors were ranked by their corresponding correlation (linear correlation for PCR and nonlinear correlation for PC-ANN models) with the half-wave potentials and entered to the models. Finally, genetic algorithm (GA) was also employed to select the best set of factors for both models.

Results. The 96% of variances in the descriptor data matrix could be explained by 30 first extracted PCs. Among these, 10, 6 and 10 PCs were selected by EV, CR and GA, respectively, for PCR , while for the ANN model, 7 PCs were selected by all of the factor selection procedures. The ANN model with EV, CR and GA factor selection procedures could explain 78.4%, 94.3% and 96% of variances in the $E_{1/2}$ data, respectively. While, the respective values obtained from different PCR procedures were 52.9%, 58.2% and 74.4%.

Conclusions. The results of this project showed that factor selection by correlation ranking and genetic algorithm gives superior results relative to those obtained by eigen value ranking. This confirms that the magnitude of the eigen value of a PC is not necessarily a measure of its significance in calibration. Moreover, it was found that for PCR method, the results obtained by GA has a major difference with those by EV and CR procedures, while, the GA and CR factor selection methods give results close to each other.

Keywords. half-wave potential; QSPR; genetic algorithm; principal component; neural network; correlation ranking; eigen-value ranking.

Abbreviations and notations

ANN, artificial neural network	PC-ANN, principal component-artificial neural network
GA, genetic algorithm	CR, correlation ranking
PLS, partial least squares	QSAR, quantitative structure-activity relationships
PCR, princiupal component regression	QSPR, quantitative structure-property relationships
EV, eigen-value ranking	$E_{1/2}$, half-wave reduction potential

Dedicated on the occasion of the 65th birthday to Professor Nenad Trinajstić.

* Correspondence author; phone: +98-831-4223307; fax: +98-831-4228439; E-mail: mshamsipur@yahoo.com

1 INTRODUCTION

With the development of experimental chemistry, a great amount of new compounds are synthesized every year. However, a large part of these compounds are not tested for fundamental or relevant thermodynamic and physicochemical properties or biological activities, which still remain unknown due to unavailability or no easily handling (toxic, odorous, etc.). A procedure able to predict, within a reasonable error margin, the physicochemical properties and biological activities of untested compounds is required to evaluate these molecular features in a fast and inexpensive way [1]. In recent years, numerous QSAR/QSPR models have been introduced for calculating the physicochemical properties with various numerical descriptors of chemical structures. These relationships derive correlations between the similarities of individual compounds and their biological activity/chemical property [2-4].

In QSAR/QSPR studies, a regression model of the form $\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{e}$ may be used to describe a set of predictor variables (**X**) with a predicted variable (**y**) by means of a regression vector (**b**). However, the colinearity, which often existed between independent variables, creates a severe problem in certain types of mathematical treatment such as matrix inversion [5]. A better predictive model can be obtained by ortogonalization of the variables by means of principal component analysis (PCA) and the consequent method is called principal component regression (PCR) [6-8]. In order to reduce the dimensionality of the independent variable space, a limited number of principal components (PCs) is used and therefore a major question will arise after the PCA is how many and which PCs constitute a good subset for predictive purposes? Hence, the selection of significant and informative PCs is the main problem in almost all PCA-based calibration methods [9-13]. Different methods have been addressed to select the significant PCs for calibration purposes. The simplest and most common one is a top-down variable selection where the factors are ranked in the order of decreasing eigen-values. The factor with highest eigen-value is considered as the most significant one 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 eigen-value is not necessarily a measure of its significance for the calibration (see Ref. 12 and references therein). In another method, called correlation ranking, the factors are ranked by their correlation coefficient with the property to be correlated (i.e., a dependent variable) and selected by the procedure discussed for eigen-value ranking [13]. Better results are often achieved by this method. Very recently, search algorithms such as genetic algorithm (GA) have been applied for the selection of variables in PCR A GA is a stochastic method to solve optimization problems defined by a fitness criterion applying the evolution hypothesis of Darwin and different genetic functions, i.e. crossover and mutation [14-16].

Artificial neural networks (ANN) are nonparametric nonlinear modeling techniques that have attracted increasing interest in recent years [17-19]. 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 suitable configuring of the network. Multilayer feed-forward neural networks trained with a back-propagation learning algorithm have become increasingly popular techniques [20-22]. The flexibility of ANN for discovering more complex relationships lead this method to find a wide application in QSAR/QSPR studies, as recently reviewed by Schneider and Wrede [23].

A principal component-artificial neural network (PC-ANN) system, which combines the PCA with ANN, is another PCA-based calibration technique for nonlinear modeling between the PCs and dependent variables [24, 25]. The problem of PC selection in PC-ANN is more serious than PCR because of the unknown and complex relationships between PCs and dependent variables. A routine method for the selection of factors in PC-ANN is the eigen-value ranking. In our previous work, we proposed a new PC-ANN algorithm called PC-GA-ANN and found that the selection of PCs by GA for PC-ANN gives better results than the eigen-value ranking method [26]. Here, we aimed to compare three different PC selection methods (i.e., eigen-value ranking, correlation ranking and genetic algorithm selection) for PC-ANN and PCR methods. The data set we used was the half-wave potential of 72 different organic compounds.

Half-wave potential ($E_{1/2}$) is an important electrochemical property of organic compounds. This property, which is a characteristic constant for a reversible oxidation-reduction system, can be useful for predicting other electrochemical properties of organic compounds [27]. There are some different electrochemical methods which permit determination of the half-wave potential of wide variety of organic, inorganic and organometallic compounds [28]. A successful strategy for prediction of the reduction potentials is construction of the QSPR models. Tompe et al. have reported a quantitative structure-electrochemistry relationship study on the half-wave potential of α,β-unsaturated ketones in nongaseous acetonitrile solution [29]. They found a linear relationship between the electronic substituent constants and $E_{1/2}$. Li and coworkers used some topological indices to correlate with the half-wave potential of different classes of organic compounds, separately [30]. However, they could not extent their model to all of the organic compounds they used. In this paper, we employed PCR and PC-ANN models to conduct a QSPR study on the data set of Li et al. [30] using theoretical descriptors.

2 MATERIALS AND METHODS

2.1 Chemical Data and Descriptors

The half-wave reduction potentials of 68 organic compounds were collected from a paper by Li et al. [30]. The $E_{1/2}$ (in mV) of these compounds are included in Table 1.

Molecular descriptors define the molecular structure and physicochemical properties of molecules by a single number. Wide variety of descriptors have been reported in the literature for use in the QSAR analyses [31-36]. There is a recently increased use of theoretical descriptors in QSAR studies. In this work, about 1200 descriptors including constitutional descriptors [31], topological indices [32, 33], topological charge indices [35], geometrical descriptors [32], molecular walk counts [34], Burden's eigen-value descriptors [35], autocorrelation descriptors [36], and physicochemical parameters and liquid properties [31] were generated for each compound. The molecular structures were drawn by the HyperChem Software [37] and saved by "hin" extension. No geometry optimization was used. The descriptors were calculated for each molecule using Dragon software [38]. The types of descriptors and number of descriptors in each group are summarized In Table 2.

2.2 Principal Component Analysis

In order to decrease the redundancy existed in the descriptors data matrix, the correlation of descriptors with each other and with the $E_{1/2}$ of the molecules was examined and collinear descriptors (i.e. r>0.9) were detected. Among the collinear descriptors, one with the lowest correlation with the half-wave potential was removed from the data matrix. In addition, the descriptors were analyzed for the existence of constant or near constant variables and those found were removed. The remaining descriptors were gathered in a new data matrix (**D**). The data set was classified into calibration ($\mathbf{D}_c$) and prediction ($\mathbf{D}_p$) sets, randomly (the number of molecules used in the calibration and prediction sets was 45 and 23, respectively). The same classification was done on the potential data. The descriptors were autoscaled to zero means and unit variance before performing PCA or any other modeling.

The calibration data matrix was subjected to PCA using the singular value decomposition procedure (SVD) [39]:

$$\mathbf{D}_c = \mathbf{U}_c \, \mathbf{S}_t \, \mathbf{V}_t^T \qquad (1)$$

where $\mathbf{U}_c$ and $\mathbf{V}_c$ are the orthonormal matrices spanned the respective row and column spaces of the data matrix ($\mathbf{D}_c$). $\mathbf{S}_c$ is a diagonal matrix whose elements are the square root of the eigen-values. The superscript “T” denote the transpose of the matrix. The eigen-vectors included in $\mathbf{U}_c$ are named as principal components (PC). The PCs of the prediction set were calculated by the equation:

$$\mathbf{U}_p = \mathbf{D}_p \; \mathbf{S}_t^{-1} \, \mathbf{V}_t \qquad (2)$$

The first 30 PCs were found to process more than 95% of variances in the original descriptors data matrix. The extracted PCs were used as the predictor variables (input) for PCR and neural network model.

2.3 Genetic Algorithm

A genetic algorithm is a problem solving method that uses generic rules such as reproduction, crossover and mutation to build pseudo organisms that are then selected, based a fitness criteria, to

survive and pass information on to the next generation. The GA used here was the same as we used previously [20,26,40]. The GA used a binary bit string representation as the coding technique for a given problem; the presence or absence of a descriptor or its second power in a chromosome is coded by 1 or 0 [14-16]. A string is composed of several genes that represent a specific characteristic to be studied. In the present case, a string is composed of 30 genes, representing the presence or absence of a PC. The GA performs its optimization by variation and selection via the evaluation of the fitness function. The fitness function was the inverse of the PRESS, which was calculated from the cross-validation procedure. The operators used here were crossover and mutation. The probability for the application of these operators was varied linearly with the generation renewal.

2.4 Principal Component Regressions

Three types PCR analysis were employed including eigen-value ranking based PCR (EV-PCR), correlation raking based-PCR (CR-PCR) and GA-based PC selection PCR (GA-PCR). In the EV-PCR procedure, the PCs were entered to the PCR model consecutively based on their decreasing eigen-value. In each step, leave-four-out cross validation was used to estimate the performances of the model by calculating the PRESS. Meanwhile, the $E_{1/2}$ of the prediction set compounds was estimated in each step and the relative error of prediction (REP) was calculated. The optimum number of factors was obtained by minimum PRESS and REP. The procedure for the CR-PCR method was similar to that discussed for the eigen-value method, except that the stepwise entrance of the PCs was based on their decreasing correlation with the $E_{1/2}$. In the GA-PCR method, the selected PCs (genes) at each string were used to build the PCR model and calculating the regression vector, **b**. This vector was then used for the calculation of the $E_{1/2}$ of the leave out compounds in the cross validation procedure. After this, the PRESS was calculated for the cross validated compounds (i.e. $PRESS_{CV}$). Therefore, each string had an associated $PRESS_{CV}$ value that measures its fitness.

2.5 Artificial Neural Network Modeling

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 set of PCs which selected by three different procedures namely eigen-value ranking, correlation ranking and GA-based selection. Number of nodes in the input layer was depended on the number of PCs introduced in the network. A bias unit with a constant activation of unity was connected to each unit in the hidden and output layers. The ANN models confined to a single hidden layer, because the network with more than one hidden layer would be harder to train. The number of nodes in the hidden layer was optimized through learning procedure. There was only one node in the output layer. The training and prediction data sets were used to optimize the network performance. To ensure that the over-fitting and under fitting 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 (i.e., RMSET and RMSEP, respectively) was

used to evaluate the performance of each neuron. This fitness function, recently proposed by Depczynski et al. [10], was used in this paper:

$$\eta = \{[(m_c - n - 1)\ RMSEC^2 + m_p\ RMSEP^2]/(m_c + m_p - n - 1)\}^{1/2} \quad (3)$$

where m_c 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 no improvement was observed in η.

2.5.1 Eigen-value ranking ANN

In the eigen-value ranking ANN (EV-ANN), the PCs were successively introduced into the network based on their decreasing eigen-value, and in each step, the number of nodes in the hidden layer was optimized. The best network structure was selected based on minimum fitness function, η.

2.5.2 Correlation ranking ANN

Since in the ANN a specific hard model is assumed between the input and output variables, the determination of the correlation between these two types of variables is a difficult task. Here, 30 different ANN models were 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 2.5, the networks were trained to model the nonlinear relationship between an individual PC and the $E_{1/2}$ of the calibration samples. The nonlinear correlation for each PC was obtained by plotting the $E_{1/2}$ predicted by its corresponding ANN model against the experimental $E_{1/2}$. After that, the PCs were ranked in the order of decreasing correlation. For the correlation ranking ANN (CR-ANN) a procedure similar to that discussed for the EV-ANN method was used.

2.5.3 PC-GA-ANN model

Previously, we described the principal of PC-GA-ANN modeling method [26]. Here, we represent only a short discussion about this model. The PCs selected on each string (genes with value 1) were used as input for the ANN. Each ANN model was trained to obtain the optimized relationship between the selected PCs and $E_{1/2}$. Several network configurations were tested, each with a different number of hidden layer elements. For each configuration, the fitness function (η) was calculated from the calibration and prediction data. For each chromosome of the GA, the training was stopped after observing no improvement in the fitness.

2.6 Software

All calculations were run on a Pentium IV personal computer with windows XP operating system. The HyperChem software was used for drawing the molecular structures [37]. The Descriptors were calculated by the Dragon software [38]. All the necessary programs for PCA, PCR, GA, ANN and other statistical analysis were written in MATLAB (ver. 6.5, MathWork Inc.).

3 RESULTS AND DISCUSSION

In Table 1 are represented the chemical structures of the compounds used in this study. As it is shown, a wide variety of organics including aliphatic, aromatic, halogenated, nitro, keto and acidic compounds are used. The experimental have-wave potentials are also included in Table 1. More than 1200 theoretical descriptors were calculated for each molecule. After the elimination of the constants and one of the collinear ones, 1150 descriptors were remained. The results of the application of the PCA on the descriptors data matrix by SVD are given in Table 2. In this Table, the eigen-values, the percent of variances explained by each eigen-value and the cumulative percent of variances are represented. The real error in the reproduction of the original data matrix using the abstracted scores and loading are also included in Table 2. As it is seen, the first 30 PCs can explain 95.33% of the variances in the original descriptors data matrix. Therefore, we restricted the next studies to these 30 PCs.

3.1 PCR Modeling

The results of the EV-PCR are summarized in Table 3. In this Table, the coefficients of determination for the cross validation, prediction and calibration are represented by R^2_{CV}, R^2_P and R^2_C, respectively. In addition, the $PRESS_{CV}$, $PRESS_P$ and standard error of calibration (SEC) are also included in Table 3. Meanwhile, F denotes the Fisher's F-ratio. The plot of PRESSCV and PRESSP against the number of factors are shown in Figure 1A. The results confirm that 10 first PCs are needed for the PCR modeling. The R^2 of cross-validation, prediction and calibration are 0.529, 0.683, and 0.731, respectively, which means at least 52.9% of variances in the reduction potentials are explained by the first 10 PCs of the descriptors data matrix. In Table 1, the predicted $E_{1/2}$ values obtained by EV-PCR using 10 first PCs are also included.

In the last column of Table 2, the correlation coefficient between each one of the PCs and $E_{1/2}$ is included. The eigen-values denote the amount of variances in the independent variables (descriptors data matrix), which can be explained by the corresponding eigen-vector (i.e. PC). In the other hand, the correlation coefficient is a part of the variance in the dependent variables ($E_{1/2}$) which is explained by a PC. As it is seen from Table 2, the changes in the correlation coefficients are not in the same trend as the eigen-values. This means that none of the extracted PCs has information content about the dependent variables. PCs with higher correlation coefficients have greater information about the variation in the $E_{1/2}$. The order of decreasing of the correlation coefficient of PCs with E1/2 is: PC1 > PC4 > PC18 > PC11 > PC23 > PC7 > PC8 > PC26 > PC22 > PC24>PC9>…. This trend shows that PC2 and PC3 which have high information content about the descriptors do not have useful information about $E_{1/2}$. Therefore, it seems reasonable to select the PCs based on their information contents about dependent variables (called correlation ranking) instead of their information contents about independent variables (called eigen-value ranking).

In Table 4, the results obtained by CR-PCR method are represented. First of all, it should be noted the statistical quality of the results obtained here are better than those found by the EV-PCR method. The plots of PRESS against the number of PC entered are shown in Figure 1B. As it is seen, the values of PRESS for the cross-validation and prediction are decreased by increasing number of PCs up to 6, and after that, the PRESS values are gradually increased. The R^2 values for the cross-validation, prediction and calibration are 0.589, 0.744, and 0.781, respectively, which means that the six PCs selected by correlation ranking procedure can explain at least 58.9% variances in $E_{1/2}$. A comparison of the F-values obtained from the two PCR methods revealed that the models obtained by the CR-PCR have greater F-values than those obtained by the EV-PCR method, indicating that the CR-PCR method gives statistically more significant models than the EV-PCR method. The predicted values of $E_{1/2}$ by the means of the best CR-PCR (model number 6 of Table 4) are included in Table 1.

In the GA-PCR method, different GAs with different set of initial population and different number of populations in each generation were run. Almost all of the models give relatively the same results. In Table 4, the two best models are represented. As it is obvious, the quality of the results obtained by GA-PCR method is superior to that found by the two other PCR models. A comparison of the PCs selected by different procedures reveals that the PCs selected by GA are close to those selected by the correlation ranking. Five PCs (i.e., PC1, PC4, PC18, PC11 and PC23) are selected by both GA and CR procedures. The models number 12 and 13 in Table 4 demonstrate that more than 73% and 68% of variances in the $E_{1/2}$ data can be explained by 10 and 8 PCs, selected by GA, respectively. In Figure 2A are shown the plots of predicted $E_{1/2}$ obtained by using the GA-PCR (model number 12 of Tables 4) against the experimental $E_{1/2}$, and the predicted $E_{1/2}$ values are included, in Table 1.

3.2 ANN Modeling

A three-layered feed-forward ANN model with back-propagation learning algorithm [41] was employed in this work. We have already used this algorithm for different QSAR studies [20, 26, 40] and some multi-component analysis [21, 42]. Since the ANN modeling by using GA for feature selection is a complex and time consuming procedure, the ANN model was confined to a single hidden layer and a sigmoid transfer function, as a more versatile transfer function, was used in this layer. The number of nodes in the hidden layer was optimized through the learning procedure.

The results of the eigen-value ranking ANN model are given in Table 5. In this method, the eigen-values were entered step by step to the network based on their decreasing eigen-values. In each step, the ANN architecture (i.e., the number of nodes in the hidden layer; n_H) and parameters (i.e. learning rate and momentum) were optimized to reach the lowest fitness (Eq. 3). The performance of the resulted models was evaluated by the fitness function (η), which was calculated based on the root mean square error of both calibration and prediction data. A plot of fitness as a

function of number of PCs entered is shown in Figure 3. The results indicate that an ANN with 7 PCs as input variables resulted in the optimum network model. This model has 2 nodes in its hidden layer and the percent of variances which can be explained by this model is 79.9% and 82.9% for the prediction and calibration, respectively. The predicted values of $E_{1/2}$ resulted from this ANN model are shown in Table 6. The major difference between the EV-PCR and EV-ANN is that the latter used lower number of PCs, while it explains more variance in the $E_{1/2}$ than the former. This is due to the nonlinear relationship between the PCs and $E_{1/2}$.

In order to apply the correlation ranking-neural network (CR-ANN) model, the ANN was used to model the nonlinear relationship between each one of the PCs and the $E_{1/2}$ data. Therefore, separate ANN models were optimized for each PC and the PCs were ranked by the variances in $E_{1/2}$, which could be explained by each of them. The results are summarized in Table 7. As it is obvious from this Table, the correlation coefficient obtained for most PCs by ANN is higher than that found by the linear model (i.e. PCR), which it is an indication for a nonlinear relationship between the extracted PCs and the $E_{1/2}$ data. The correlation coefficients reported in Table 7 were used to perform the CR-ANN. The EV-ANN procedure was repeated for the CR-ANN except that the PCs were entered into the ANN model based on their decreasing correlation coefficient. The results are summarized in Tables 8. The modeling ability of the ANN was increased by introducing more PCs in the model, up to 7 PCs. No improvement in the modeling power of the ANN was observed when more PCs were introduced to the network. A plot of the fitness function as a function of the number of PCs entered is shown in Figure 3. Thus, the resulted optimum CR-ANN model contained seven PCs (i.e., PC4, PC1, PC18, PC3, PC25, PC5, and PC7). This model which used 3 nodes in its hidden layer could explain more than 94% of variances in the half-wave potential data with a fitness function equal to 0.078. The predicted $E_{1/2}$ values obtained by this model are also included in Table 6. Referring to the data shown in Tables 5 and 8, which obtained from the respective EV-ANN and CR-ANN, reveals that both models used the same number of PCs as the input (i.e. 7 PCs) with 3 common PCs (i.e., PCs 1, 3 and 4); however, the latter could model the structure-electrochemistry relationship better, than the former one.

In order to enhance the modeling ability of ANN, the genetic algorithm was also used for selection of PCs. In the GA procedure, the population of the first generation was selected randomly. The number of genes with the value 1 was kept relatively low to have a small subset of descriptors in the ANN modeling method, i.e., the probability of generating 0 for a gene was set greater than the value 1. The operators used here were crossover and mutation. The probability for the application of these operators was varied linearly with the generation renewal (0–1% for mutation and 60–90% for crossover). The population size was varied between 50 and 250 for different GA runs. For a typical run, the evolution of the generation was stopped when 90% of the generations took the same fitness (Eq. 3). The best GA-ANN model is summarized in the last row of Table 8. As it is observed, this model with 7 PCs has a fitness function equal to 0.066, which is lower than

that obtained by the other two ANN models and could explain more than 96% of variances in the $E_{1/2}$ data. The PC1, PC3, PC4, PC18, PC14, PC26 and PC21 selected by GA are used as the input variables in ANN. These selected PCs are closer to those selected by CR-ANN, relative to the selected PCs by the EA-ANN model. In addition, the results of the CR- and GA-based factor selection method in ANN modeling method are closer to each other than the PCR method. The predicted $E_{1/2}$ data obtained by the GA-ANN are given in Table 6, and a plot of the predicted potentials by GA-ANN against the experimental potentials is shown in Figure 2B.

4 CONCLUSIONS

A quantitative-electrochemistry relationship analysis has been conducted on the half-wave potential of 69 different organic compounds by using the principal component regression and principal component-artificial neural network modeling methods, with application of thee different factor selection procedures GA, CR and EV. The genetic algorithm-based factor selection gave results superior to those found by the eigen-value ranking and the correlation ranking procedures. Meanwhile, the results of correlation ranking and genetic algorithm selection were more close to each other, especially for ANN modeling method. Thus, it can be concluded that the factor selection for ANN by correlation ranking is more straightforward than genetic algorithm due to the complexity of the principal-component-genetic algorithm-artificial neural network model.

5 REFERENCES

[1] H. Kubinyi, QSAR and 3D QSAR in Drug Design Part 1: Methodology, *Drug. Devlop. Theory* **1997**, *2*, 457-467.
[2] D. Hadjipavlou-Litina, Review, Reevaluation ,and New Results in Quantitative Structure–Activity Studies of Anticonvulsants*, Med. Res. Rev.* **1998**, *18*, 91–119.
[3] P. Gramatica, E. Papa, QSAR Modeling of Bioconcentration Factor by Theoretical Molecular Descriptors, *QSAR Comb. Sci.* **2003**, *22*, 374-385.
[4] C. Hansch, A. Kurup, Chem-Bioinformatics and QSAR: A Review of QSAR Lacking Positive Hydrophobic Terms, R. Garg, H. Gao, *Chem. Rev.* **2001**, *101*, 619-672.
[5] D.C. Montgomery, E.A. Peck, *Introduction to Linear Regression Analysis*, Wiley, New York, 1982.
[6] I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, 1986.
[7] J. H. Kalivas and P. M. Lang, *Mathematical Analysis of Spectral Orthogonality*, Marcel Dekker, New York, 1994.
[8] G. Puchwein, Selection of Calibration Samples for Near-Infrared Spectrometry by Factor Analysis of Spectra, *Anal. Chem.* **1988**, *60*, 569-573.
[9] Y. L. Xie and J. H. Kalivas, Evaluation of Principal Component Selection Methods to Form a Global Prediction Model by Principal Component Regression, *Anal. Chim. Acta* **1997**, *348*, 19-27.
[10] U. Depczynski, V. J. Frost, K. Molt, Genetic Algorithm Applied to the Selection of Factors in Principal Component Regression, *Anal. Chim. Acta* **2000**, *420*, 217-227.
[11] A. S. Barros and D. N. Rutledge, Genetic Algorithm Applied to the Selection of Principal Components, *Chemomet. Intell. Lab. Syst.* **1998**, *40*, 65-81.
[12] J. M. Sutter, J. H. Kalivas and P. M. Lang, Which Principal Components to Utilize for Principal Component Regression, *J. Chemometr.* **1992**, *6*, 217-225.
[13] J. Sun, A Correlation Principal Component Regression Analysis of NIR Data, *J. Chemometr.* **1995**, *9*, 21-29.
[14] D. Jouanrimbaud, D. L. Massart, R. Leardi and O. E. deNoord, Genetic Algorithms as a Tool for Wavelength

Selection in Multivariate Calibration. *Anal. Chem.* **1995**, *67*, 4295-4301.
[15] Lucasius, C. B.; Beckers, M. L. M.; Kateman, G. Genetic algorithms in wavelength selection: a comparative study. *Anal. Chim. Acta* **1994**, *286*, 135-148.
[16] W. Cai, B. Xia, X. Shao, Q. Guo, B. Maigret, Z. Pan, Molecular Interactions of α-Cyclodextrin Inclusion Complexes Using a Genetic Algorithm**,** *J. Mol. Struct. (Theochem.)* **2001**, *535*, 115-119.
[17] K. Petritis, L.J. Kangas, P.L. Ferguson, G.A. Anderson, L. Pasa-Tolic, M.S. Lipton, K.J. Auberry, E.F. Strittmatter, Y. Shen, R. Zhao, R.D. Smith, Use of Artificial Neural Networks for the Accurate Prediction of Peptide Liquid Chromatography Elution Times in Proteome Analyses, *Anal. Chem*. **2003**, *75*, 1039-1048.
[18] L. Douali, D. Villemin and D. Cherqaoui, Neural Networks: Accurate Nonlinear QSAR Model for HEPT Derivatives, *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1200-1207
[19] J. Polanski, J. Gasteiger, M. Wagener, J. Sadowski, The Comparison of Molecular Surfaces by Neural Networks and its Applications to Quantitative Structure Activity Studies, *Quant. Struct. Act. Relat.* **1998**, *17*, 27-36.
[20] B. Hemmateenejad, M. A. Safarpour, F. Taghavi, Application of ab initio Theory for the Prediction of Acidity Constants of Some 1-Hydroxy-9,10-Anthraquinone Derivatives Using Genetic Neural Network, *J. Mol. Struct. (Theochem.)* **2003**, *635*, 183-190.
[21] M. Shamsipur, B. Hemmateenejad and M. Akhond, Multicomponent Acid-Base Titration by Principal Component-Artificial Neural Network Calibration, *Anal. Chim. Acta* **2002**, *461*, 147-153.
[22] D. T. Manallack, B. G. Tehan, E. Gancia, B. D. Hudson, M. G. Ford, D. J. Livingstone, D. C. Whitley and W. R. Pitt, A Consensus Neural Network-Based Technique for Discriminating Soluble and poorly Soluble Compounds, *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 674-679.
[23] Schneider, P. Wrede, Artificial Neural Networks for Computer-Based Molecular Design, *Prog. Biophys. Mol. Biol*. **1998**, *70*, 175-222.
[24] P. J. Gemperline, J. R. Long, and G. Gregoriou, Nonlinear multivariate calibration using principal components regression and artificial neural networks. *Anal. Chem.* **1991**, *63*, 2313-2317.
[25] R. Vendrame, R. S. Braga, Y. Takahata and D. S. Galvao, Structure-Activity Relationship Studies of Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons Using Calculated Molecular Descriptors with Principal Component Analysis and Neural Network Methods, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1094-1104.
[26] B. Hemmateenejad, M. Akhond, R. Miri, M. Shamsipur, 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), *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1328-1334.
[27] R. L. McNaughton, A. A. Tipton, N. D. Rubie, R. R. Conry and M. L. Kirk, Electronic Structure Studies of Oxomolybdenum Tetrathiolate Complexes: Origin of Reduction Potential Differences and Relationship to Cysteine-Molybdenum Bonding in Sulfite Oxidase, *Inorg. Chem*. **2000**, *39*, 5697-5706.
[28] S. Niu, X. B. Wang, J. A. Nichols, L. S. Wang and T. Ichiye, Combined Quantum Chemistry and Photoelectron Spectroscopy Study of the Electronic Structure and Reduction Potentials of Rubredoxin Redox Site Analogues, *J. Phys. Chem. A*. **2003**, *107*, 2898-2907.
[29] P. Tompe, Gy. Clementis, I. Petnehazy, Zs. M. Jaszay, L. Toke, Quantitative Structure-Electrochemistry Relationships of α, β-Unsaturated Ketones, *Anal. Chim. Acta* **1995***, 305*, 295-303.
[30] H. Li, L. Xu, Q. Su, Structure-Property Relationship Between Half-Wave Potentials of Organic Compounds and Their Topology, *Anal. Chim. Acta* **1995**, *316*, 39-45.
[31] R. Todeschini, and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH: Weinheim, Germany, 2000.
[32] L. B. Kier and L. H. Hall, *Molecular Connecti*V*ity in Structure-Activity Analysis;* RSP-Wiley: Chichester, UK, 1986.
[33] E. V. Kostantinora, Exploring Functional Group Transformations on CASREACT. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 54-58.
[34] G. Rucker and C. Rucker, Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683-688.
[35] J. Galvez, R. Garcia, M. T. Salabert and R. Soler, Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci*. **1994**, 34, 520-525.
[36] P. Broto, G. Moreau and C. Vandicke, 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*, 66-71.
[37] Hypercube Inc., http://www.hyper.com.
[38] R. Todeschini, *Milano Chemometrics and QSAR group*; http://www.disat.unimib.it/vhm.
[39] I. N. Koprinarov, A. P. Hitchcock, C. T. McCrory and R. F. Childs, Quantitative Mapping of Structured

Polymeric Systems Using Singular Value Decomposition Analysis of Soft X-ray Images, *J. Phys. Chem. B.* **2002**, *106*, 5358-5364.

[40] B. Hemmateenejad, R. Miri, M. Akhond, M. Shamsipur, 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, *Chemometr. Intell. Lab. Syst*. **2002**, *64*, 91-99.

[41] D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533-539.

[42] M. Shamsipur, B. Hemmateenejad, M. Akhond, Simultaneous Determination of Promethazine, Chlorpromazine, and Perphenazine by Multivariate Calibration Methods and Derivative Spectrophotometry, *J. AOAC Int.* **2002**, *85*, 555-562.

Table 1. The experimental half wave potentials and the predicted values by different PCR models

No.	*Compound*	*Exp. (V)*	*Pred. (V)*		
			EV-PCR	CR-PCR	GA-PCR
1	Anthraquinone	-0.54	-0.389	-0.944	-0.641
2	Benzoquinone	0.15	-1.155	-0.558	-0.062
3	2,3-Dimethyl naphtoquinone	-0.22	-0.113	-0.590	-0.610
4	Duroquinone	-0.09	-0.047	-0.575	-0.119
5	2-Methyl-1,4-naphtoquinone	-0.17	-0.378	-0.652	-0.554
6	Toluquinone	0.09	-0.822	-0.682	-0.092
7	Azobenzene	-0.33	-1.137	-0.468	-0.485
8	Benzenediazonium chloride	-0.67	-1.372	-1.168	-1.137
9	m-Dinitobenzene	-0.26	-0.149	-0.318	0.0109
10	Methyl o-nitrobenzoate	-0.25	-0.013	-0.099	-0.641
11	Methyl m-nitrobenzoate	-0.24	-0.012	-0.188	-0.162
12	Methyl p-nitrobenzoate	-0.20	-0.029	-0.232	-0.263
13	p-Nitoaniline	-0.36	-0.623	-0.522	-0.695
14	o-Nitroaniline	-0.29	-0.517	-0.125	-0.383
15	p-Nitroanisole	-0.35	-0.374	-0.212	0.105
16	m-Nitrobenzaldehyde	-0.28	-0.505	-0.660	-0.507
17	o-Nitobenzoic acid	-0.23	-0.335	-0.166	0.0734
18	m-Nitobenzoic acid	-0.20	-0.279	-0.506	-0.479
19	p-Nitobenzoic acid	-0.17	-0.252	-0.691	-0.534
20	o-Nitrophenol	-0.23	-0.679	-0.611	-0.236
21	m-Nitrophenol	-0.37	-0.686	-0.544	-0.629
22	p-Nitrophenol	-0.35	-0.644	-0.520	-0.683
23	o-Nitrotoluene	-0.26	-0.615	-0.404	-0.658
24	m-Nitrotoluene	-0.22	-0.607	-0.404	-0.654
25	p-Nitrotoluene	-0.24	-0.570	-0.386	-0.532
26	Acetaldehyde	-1.89	-1.439	-1.400	-1.622
27	Acrolein	-1.36	-1.488	-1.388	-1.564
28	Benzaldehyde	-0.94	-1.304	-1.312	-1.583
29	Crotonaldehyde	-0.92	-1.484	-1.403	-1.020
30	Formaldehyde	-1.59	-1.476	-1.693	-2.092
31	Furfural	-1.06	-1.409	-1.413	-1.046
32	Glyoxal	-1.41	-1.353	-1.708	-1.246
33	p-Hydroxybenzalhedyde	-1.16	-1.038	-1.011	-1.116
34	o-Methoxybenzaldehyde	-1.03	-0.740	-0.593	-0.773

35	p-Methoxybenzaldehyde	-1.07	-0.766	-0.648	-0.598
36	Methyl glyoxal	-0.83	-1.083	-0.974	-1.095
37	Salicylaldehyde	-1.02	-1.044	-1.028	-1.097
38	Acridine	-0.80	-0.885	-0.607	-0.332
39	8-Hydroxyquinoline	-1.39	-0.971	-1.024	-0.867
40	Nicotinamide	-1.56	-0.985	-1.020	-1.476
41	Pyridine	-1.49	-1.581	-1.338	-1.975
42	Quinaldinic acid	-0.86	-0.651	-0.884	-0.436
43	Quinoline	-1.23	-1.258	-1.240	-0.959
44	Quinoline-8-carboxylic acid	-1.11	-0.592	-0.861	-0.597
45	Saccharin	-1.77	-0.662	-1.227	-1.868
46	Acrilonitrile	-1.94	-1.577	-1.509	-1.542
47	Ascorbic acid	-0.17	-0.628	-0.019	-0.101
48	Bromoacetic acid	-0.54	-1.401	-1.107	-0.180
49	α-Bromopropionic acid	-0.39	-1.284	-0.762	-0.322
50	Crotonic acid	-1.94	-1.131	-1.087	-2.064
51	Dibromoacetic acid	-0.03	-0.472	-0.068	0.0459
52	Diethyl fumarate	-0.84	-0.560	-0.593	-0.625
53	Diethyl maleate	-0.95	-0.622	-0.797	-0.887
54	Ethyl dichloroacetate	-0.86	-0.827	-1.088	-1.110
55	Fumaric acid	-1.60	-0.914	-1.304	-1.288
56	Maleic acid	-1.36	-0.914	-1.304	-1.288
57	Methylacrylonitrile	-2.07	-1.276	-1.359	-1.997
58	Pyruvic acid	-0.86	-0.897	-0.724	-0.932
59	Trichloroacetic acid	-0.84	-0.637	-1.355	-1.072
60	Allyl chloride	-1.91	-1.661	-1.539	-1.617
61	Allyl bromide	-1.29	-1.844	-1.750	-1.784
62	Benzal chloride	-1.81	-1.242	-1.096	-2.220
63	Benzotrichloride	-0.68	-1.113	-1.277	-1.234
64	Benzyl chloride	-1.94	-1.302	-1.107	-2.047
65	Bromobenzene	-2.32	-1.838	-1.729	-2.366
66	n-Butyl bromide	-2.27	-1.900	-1.914	-2.453
67	p-Dibromobenzene	-0.78	-2.001	-1.652	-2.311
68	m-Dichlorobenzene	-0.30	-1.402	-1.728	-2.480
69	Nitromethane	-0.83	-0.389	-0.944	-0.641

Table 2. The results of application of PCA on the descriptors data matrix

PC No.	*Log of Eigen-value*	*% of variance explained*	*Cumulative percent of variances*	*Real error*	*Correlation coefficient*
PC1	4.4527	36.6760	36.6760	0.7958	0.526
PC2	3.9158	10.6548	47.3309	0.7731	-0.032
PC3	3.7667	7.5584	54.8893	0.681	-0.011
PC4	3.6181	5.3682	60.2574	0.6450	0.407
PC5	3.5465	4.5521	64.8095	0.6118	-0.040
PC6	3.4434	3.5898	68.3994	0.5844	-0.196
PC7	3.3656	3.0012	71.4006	0.5605	0.121
PC8	3.2863	2.5006	73.9012	0.5398	0.123
PC9	3.2335	2.2141	76.1153	0.5208	0.065
PC10	3.1657	1.8940	78.0093	0.5040	-0.108
PC11	3.1181	1.6976	79.7069	0.4884	0.201
PC12	3.0932	1.6029	81.3097	0.4729	0.01
PC13	3.0152	1.3393	82.6490	0.4597	-0.106
PC14	3.0068	1.3138	83.9629	0.4461	0.097
PC15	2.9532	1.1612	85.1240	0.4337	0.007
PC16	2.9131	1.0587	86.1827	0.4219	-0.131
PC17	2.9056	1.0408	87.2235	0.4097	0.048
PC18	2.8564	0.9293	88.1528	0.3984	0.210
PC19	2.8275	0.8693	89.0221	0.3874	-0.049
PC20	2.8055	0.8265	89.8486	0.3764	-0.073
PC21	2.7562	0.7377	90.5863	0.3663	-0.150
PC22	2.7298	0.6943	91.2806	0.3564	0.068
PC23	2.6999	0.6480	91.9286	0.3467	0.172
PC24	2.6564	0.5864	92.5149	0.3376	0.067
PC25	2.6108	0.5278	93.0428	0.3292	-0.266
PC26	2.6043	0.5200	93.5628	0.3205	0.092
PC27	2.5767	0.4880	94.0507	0.3118	-0.158
PC28	2.5441	0.4527	94.5034	0.3034	-0.140
PC29	2.5226	0.4309	94.9343	0.2950	-0.052
PC30	2.4884	0.3982	95.3325	0.2869	0.104
PC31	2.4534	0.3674	95.6999	0.2790	0.043
PC32	2.4255	0.3446	96.0445	0.2713	-0.040
PC33	2.4192	0.3396	96.3840	0.2631	-0.071

Table 3. Results of EV-PCR in the presence of different entered PCs

PC entered	R^2_{CV}	$PRESS_{CV}$	R^2_P	$PRESS_P$	R^2_C	*SEC*	*F*
PC1	0.277	25.72	0.284	24.02	0.351	0.592	25.31
PC1+PC2	0.278	24.82	0.294	23.33	0.397	0.596	12.52
PC1+PC2+PC3	0.341	24.24	0.368	23.02	0.440	0.601	8.22
PC1+PC2+PC3+ PC4	0.444	21.186	0.484	20.08	0.511	0.532	12.55
PC1+PC2+PC3+ PC4+PC5	0.445	20.0229	0.520	19.44	0.588	0.535	9.94
PC1+PC2+PC3+PC4+PC5+PC6	0.484	19.735	0.532	19.11	0.601	0.521	9.51
PC1+PC2+PC3+PC4+PC5+PC6+ PC7	0.499	19.665	0.589	17.89	0.642	0.517	8.52
PC1+PC2+PC3+PC4+PC5+ PC6+PC7 + PC8	0.513	19.537	0.631	17.43	0.692	0.514	7.77
PC1+PC2+PC3+ PC4+PC5+PC6+ PC7+PC8+PC9	0.518	19.459	0.639	16.77	0.709	0.516	6.91
PC1+PC2+PC3+PC4+PC5+PC6+ PC7+PC8+PC9+PC10	0.529	18.349	0.683	16.07	0.731	0.514	6.41
PC1+PC2+PC3+PC4+PC5+PC6+ PC7+PC8+PC9+PC10+PC11	0.570	18.532	0.701	15.98	0.782	0.496	6.73
PC1+PC2+PC3+PC4+PC5+PC6+ PC7+PC8+PC9+PC10+PC11+PC12	0.570	18.534	0.700	15.97	0.790	0.500	6.01
PC1+PC2+PC3+PC4+PC5+PC6+ PC7+PC8+PC9+PC10+PC11+PC12+ PC13	0.581	18.291	0.704	15.94	0.802	0.498	5.76
PC1+PC2+PC3+PC4+PC5+PC6+ PC7+PC8+PC9+PC10+PC11+PC12+ PC13+PC14	0.590	18.310	0.710	15.95	0.808	0.498	5.45

Table 4. Results of CR-PCR and GA-PCR in the presence of different entered PCs

No.	*PC entered*	R^2_{CV}	$PRESS_{CV}$	R^2_P	$PRESS_P$	R^2_C	*SEC*	*F*
1	PC1[a]	0.277	25.72	0.284	24.02	0.351	0.592	25.31
2	PC1+PC4 [a]	0.442	19.36	0.511	19.05	0.397	0.545	25.78
3	PC1+PC4+PC18 [a]	0.487	18.61	0.552	18.43	0.579	0.507	20.22
4	PC1+PC4+PC18+PC11 [a]	0.527	18.11	0.608	18.02	0.637	0.490	17.54
5	PC1+PC4+PC18+PC11+PC23 [a]	0.556	17.55	0.668	17.12	0.709	0.478	15.55
6	PC1+PC4+PC18+PC11+PC23+ PC8 [a]	0.582	17.21	0.744	17.01	0.781	0.474	13.56
7	PC1+PC4+PC18+PC11+PC23+PC8 +PC7 [a]	0.589	17.23	0.752	17.89	0.787	0.470	12.41
8	PC1+PC4+PC18+PC11+PC23+PC8 +PC7+PC26 [a]	0.595	18.09	0.758	17.84	0.790	0.469	10.82
9	PC1+PC4+PC18+PC11+PC23+PC8 +PC7+PC26+ PC22 [a]	0.599	18.64	0.760	17.86	0.795	0.470	9.64
10	PC1+PC4+PC18+PC11+PC23+PC8 +PC7+PC26+ PC22 [a]	0.599	18.64	0.760	17.86	0.795	0.470	9.64
11	PC1+PC4+PC18+PC11+PC23+PC8 +PC7+PC26+ PC22+PC24 [a]	0.604	19.35	0.763	18.02	0.799	0.472	8.68
12	PC1+PC4+PC6+PC11+PC18+PC21 +PC23 +PC25+PC27+PC28[b]	0.733	11.81	0.857	12.09	0.893	0.387	15.60
13	PC1+PC4+PC6+PC11+PC18+PC21 +PC25+PC28[b]	0.684	13.19	0.818	13.65	0.799	0.414	15.93

[a] The results obtained by CR-PCR
[b] The results obtained by GA-PCR

Table 5. Results of EV-ANN in the presence of different entered PCs

PC entered	n_H	$RMSE_p$	R^2_P	$RMSE_c$	R^2_C	η
PC1	2	0.321	0.341	0.301	0.388	0.307
PC1+PC2	2	0.270	0.389	0.239	0.422	0.251
PC1+PC2+PC3	4	0.230	0.461	0.204	0.481	0.211
PC1+PC2+PC3+ PC4	3	0.192	0.598	0.158	0.610	0.171
PC1+PC2+PC3+ PC4+PC5	2	0.159	0.675	0.121	0.678	0.136
PC1+PC2+PC3+ PC4+PC5+PC6	4	0.147	0.722	0.120	0.741	0.131
PC1+PC2+PC3+ PC4+PC5+PC6 + PC7	4	0.138	0.784	0.120	0.823	0.127
PC1+PC2+PC3+ PC4 + PC5+PC6 + PC7+PC8	2	0.132	0.799	0.105	0.829	0.116
PC1+PC2+PC3+PC4+PC5+PC6+PC7+ PC8 + PC9	3	0.135	0.804	0.100	0.849	0.115
PC1+PC2+PC3+PC4+PC5+PC6+PC7 +PC8+PC9+PC10	5	0.130	0.814	0.104	0.863	0.115
PC1+PC2+PC3+PC4+PC5+PC6+PC7 +PC8+PC9+PC10+PC11	4	0.128	0.825	0.103	0.884	0.114
PC1+PC2+PC3+PC4+PC5+PC6+PC7 +PC8+PC9+PC10+PC11+PC12	5	0.141	0.830	0.091	0.891	0.113
PC1+PC2+PC3+PC4+PC5+PC6+PC7 +PC8+PC9+PC10+PC11+PC12+ PC13	4	0.137	0.832	0.078	0.894	0.109
PC1+PC2+PC3+ PC4+PC5+PC6+PC7 +PC8+PC9+PC10+PC11+PC12+PC13 +PC14	4	0.136	0.832	0.089	0.890	0.112

Table 6. The predicted values of $E_{1/2}$ by different ANN models

No. [a]	***Exp.***	***Pred. $E_{1/2}$***			***No.*** [a]	***Exp.***	***Pred. $E_{1/2}$***		
	$E_{1/2}$	EV	CR	GA		***$E_{1/2}$***	EV	CR	GA
1	-0.54	-0.641	-0.621	-0.548	35	-1.07	-0.857	-0.985	-1.209
2	0.146	0.040	0.137	0.139	36	-0.83	-1.096	-0.845	-0.979
3	-0.216	-0.460	-0.387	-0.371	37	-1.02	-1.200	-0.980	-1.110
4	-0.093	-0.309	-0.264	-0.169	38	-0.8	-0.538	-0.946	-0.909
5	-0.17	-0.420	-0.365	-0.186	39	-1.39	-1.019	-1.281	-1.282
6	0.09	0.055	0.059	0.055	40	-1.56	-1.477	-1.470	-1.661
7	-0.33	-0.486	-0.480	-0.445	41	-1.49	-1.656	-1.652	-1.443
8	-0.67	-0.761	-0.868	-0.789	42	-0.86	-0.650	-0.817	-0.890
9	-0.26	0.011	-0.253	-0.250	43	-1.23	-0.960	-1.214	-1.219
10	-0.25	-0.642	-0.476	-0.440	44	-1.11	-1.158	-1.281	-1.186
11	-0.24	-0.169	-0.178	-0.415	45	-1.77	-1.958	-1.844	-1.835
12	-0.2	-0.261	-0.230	-0.261	46	-1.94	-2.001	-2.004	-1.963
13	-0.36	-0.320	-0.326	-0.342	47	-0.17	-0.052	-0.152	-0.160
14	-0.29	-0.384	-0.465	-0.312	48	-0.54	-0.532	-0.543	-0.594
15	-0.35	-0.370	-0.247	-0.521	49	-0.39	-0.323	-0.325	-0.387
16	-0.28	-0.236	-0.320	-0.320	50	-1.94	-2.204	-1.919	-1.925
17	-0.23	-0.100	-0.186	-0.197	51	-0.03	-0.281	-0.204	-0.113
18	-0.2	-0.480	-0.362	-0.381	52	-0.84	-0.625	-1.047	-0.980
19	-0.17	-0.242	-0.191	-0.188	53	-0.95	-0.888	-0.889	-0.991
20	-0.23	-0.236	-0.235	-0.236	54	-0.86	-0.957	-0.934	-0.957
21	-0.37	-0.630	-0.527	-0.448	55	-1.6	-1.289	-1.511	-1.678
22	-0.35	-0.684	-0.437	-0.413	56	-1.36	-1.289	-1.293	-1.303
23	-0.26	-0.504	-0.428	-0.225	57	-2.07	-1.990	-2.209	-2.064
24	-0.22	-0.219	-0.219	-0.223	58	-0.86	-1.041	-1.041	-0.905
25	-0.24	-0.532	-0.412	-0.420	59	-0.84	-1.073	-1.002	-0.940
26	-1.89	-1.623	-1.796	-1.941	60	-1.91	-1.618	-1.891	-1.900
27	-1.36	-1.628	-1.354	-1.350	61	-1.29	-1.494	-1.421	-1.312
28	-0.94	-1.287	-1.091	-1.073	62	-1.81	-2.002	-1.980	-1.812
29	-0.92	-1.080	-1.080	-1.016	63	-0.68	-1.019	-0.672	-0.711
30	-1.59	-1.561	-1.568	-1.577	64	-1.94	-1.924	-1.927	-2.042
31	-1.06	-1.046	-1.047	-1.055	65	-2.32	-2.366	-2.341	-2.304
32	-1.41	-1.247	-1.252	-1.522	66	-2.27	-2.454	-2.181	-2.293
33	-1.16	-1.117	-1.326	-1.209	67	-2.10	-2.312	-2.075	-2.159
34	-1.03	-0.774	-1.136	-1.119	68	-2.48	-2.481	-2.480	-2.461

[a] The numbers refer to the compounds shown in Table 1.

Table 7. The results of ANN for modeling between each one of PCs and $E_{1/2}$

PC	n_H	***R***	***PC***	n_H	***r***
PC1	2	0.549	PC17	1	0.328
PC2	2	0.232	PC18	4	0.488
PC3	3	0.421	PC19	3	0.069
PC4	2	0.642	PC20	2	0.331
PC5	4	0.180	PC21	5	0.154
PC6	1	0.208	PC22	2	0.083
PC7	3	0.280	PC23	1	0.199
PC8	2	0.138	PC24	3	0.152
PC9	2	0.094	PC25	3	0.404
PC10	2	0.105	PC26	4	0.095
PC11	3	0.237	PC27	4	0.154
PC12	2	0.030	PC28	2	0.149
PC13	3	0.114	PC29	3	0.163
PC14	3	0.276	PC30	1	0.110
PC15	4	0.066	PC31	4	0.048
PC16	2	0.144	PC32	5	0.056

Table 8. Results of CR-ANN and GA-ANN in the presence of different entered PCs

PC entered	n_H	$RMSE_p$	R^2_P	$RMSE_c$	R^2_C	η
PC4[a]	2	0.275	0.427	0.297	0.466	0.289
PC4+PC1[a]	3	0.237	0.513	0.221	0.559	0.226
PC4+PC1+PC18[a]	2	0.204	0.627	0.192	0.696	0.196
PC4+PC1+PC18+PC3[a]	4	0.131	0.772	0.102	0.803	0.113
PC4+PC1+PC18+PC3+PC25[a]	3	0.103	0.864	0.094	0.875	0.097
PC4+PC1+PC18+PC3+PC25+PC20[a]	4	0.089	0.898	0.082	0.907	0.084
PC4+PC1+PC18+PC3+PC25+PC20 +PC17[a]	3	0.081	0.943	0.077	0.972	0.078
PC4+PC1+PC18+PC3+PC25+PC20 +PC17+PC7[a]	2	0.081	0.945	0.075	0.978	0.077
PC4+PC1+PC18+PC3+PC25+PC20 +PC17+PC7+PC20[a]	3	0.082	0.948	0.074	0.980	0.077
PC4+PC1+PC18+PC3+PC25+PC20 +PC17+PC7+PC20+PC14[a]	4	0.083	0.952	0.072	0.984	0.076
PC4+PC1+PC18+PC3+PC25+PC20 +PC17+PC7+PC20+PC14+PC11[a]	2	0.081	0.955	0.071	0.984	0.077
PC4+PC1+PC18+PC3+PC25+PC20+ PC17+PC7+PC20+PC14+PC11+PC2[a]	4	0.084	0.954	0.072	0.983	0.076
PC1+PC3+PC4+PC18+PC14+PC26+ PC20[b]	3	0.068	0.967	0.065	0.971	0.066

[a] The results obtained by CR-ANN
[b] The results obtained by GA-ANN

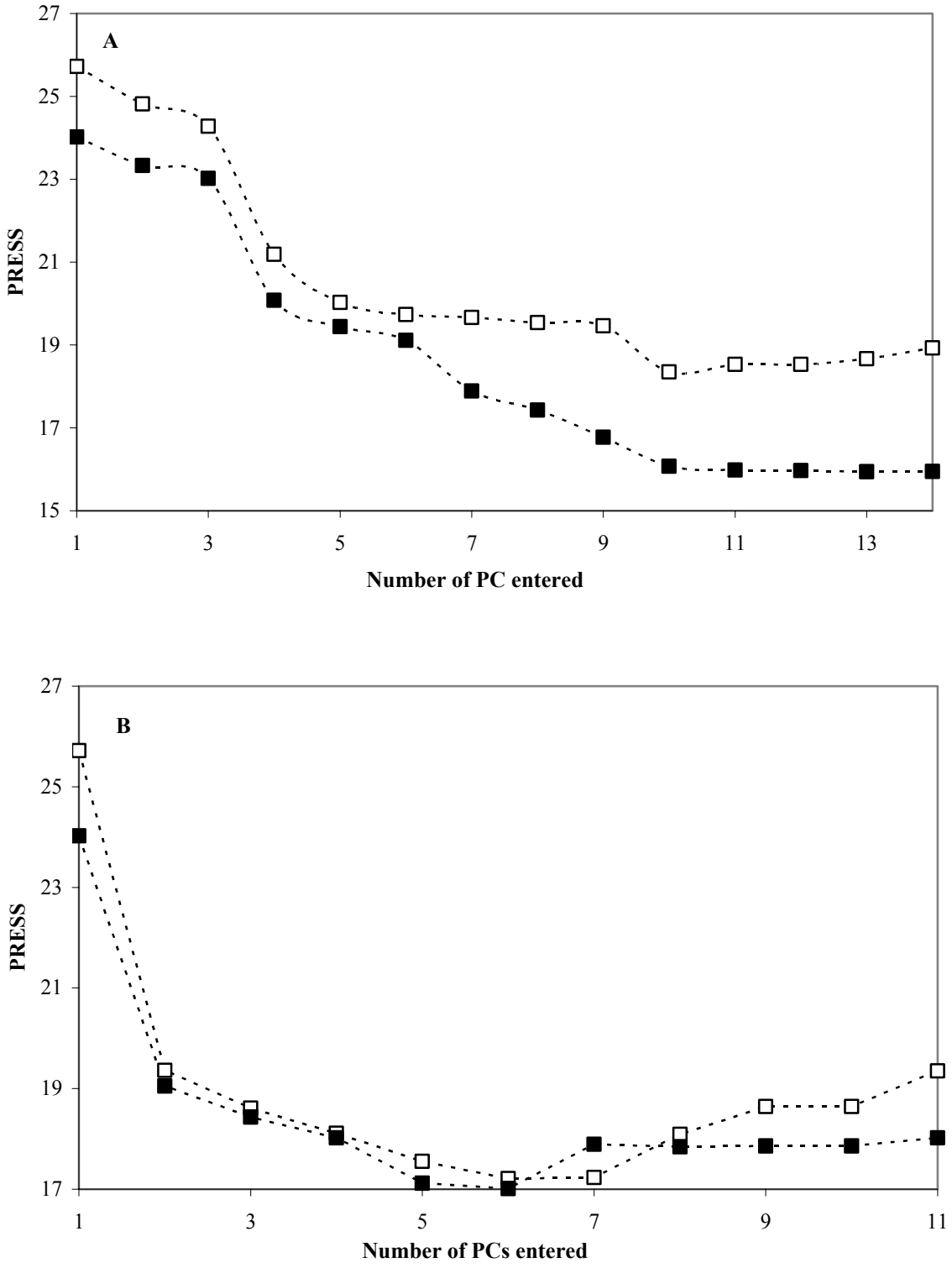

Figure 1. Plot of the PRESS for cross-validation (open markers) and prediction (filled markers) as a function of the number of PC entered: A)EV-PCR and B) CR-PCR.

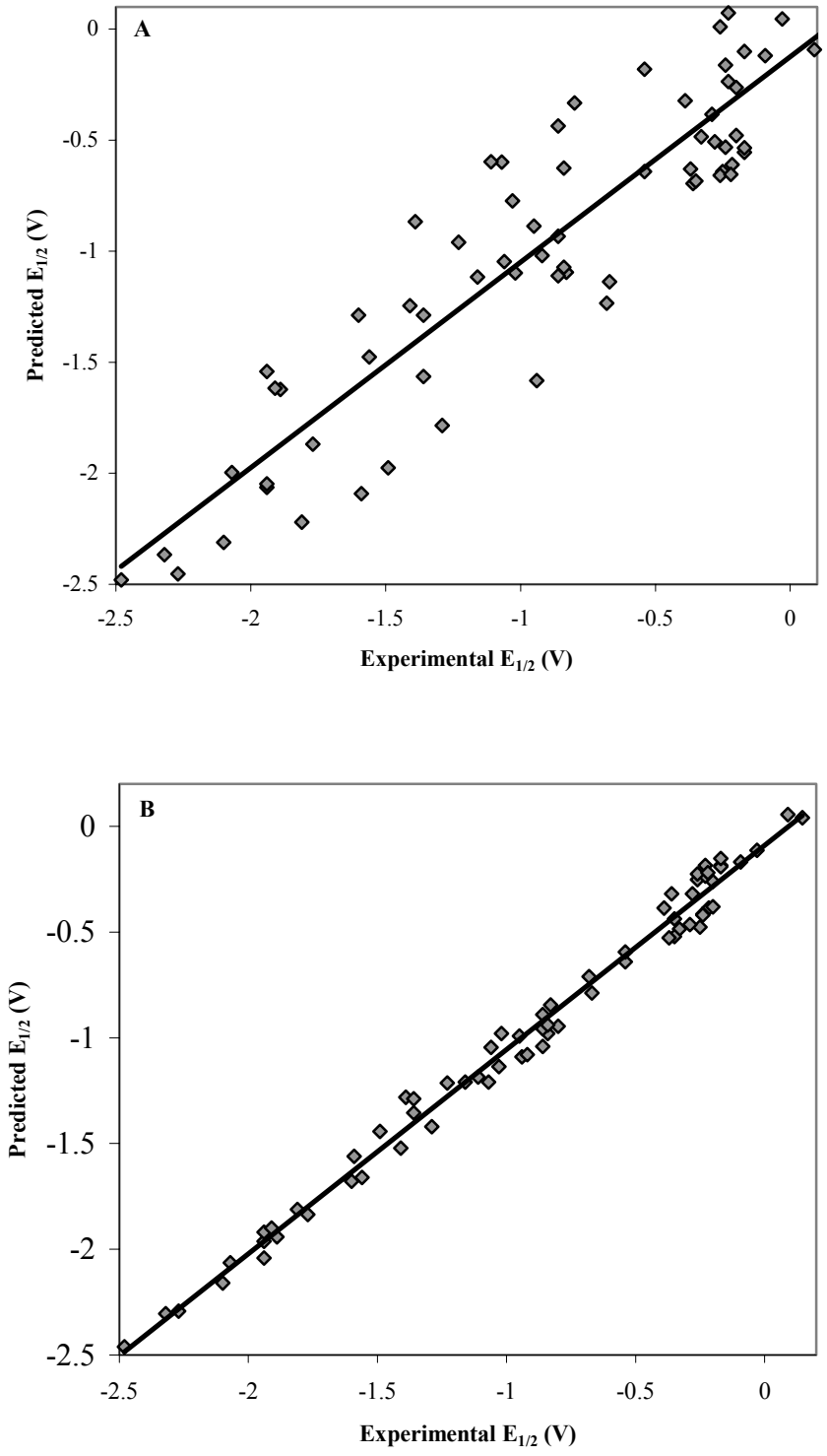

Figure 2. Plot of the predicted potential against the experimental potential for A) GA-PCR and B) GA-ANN

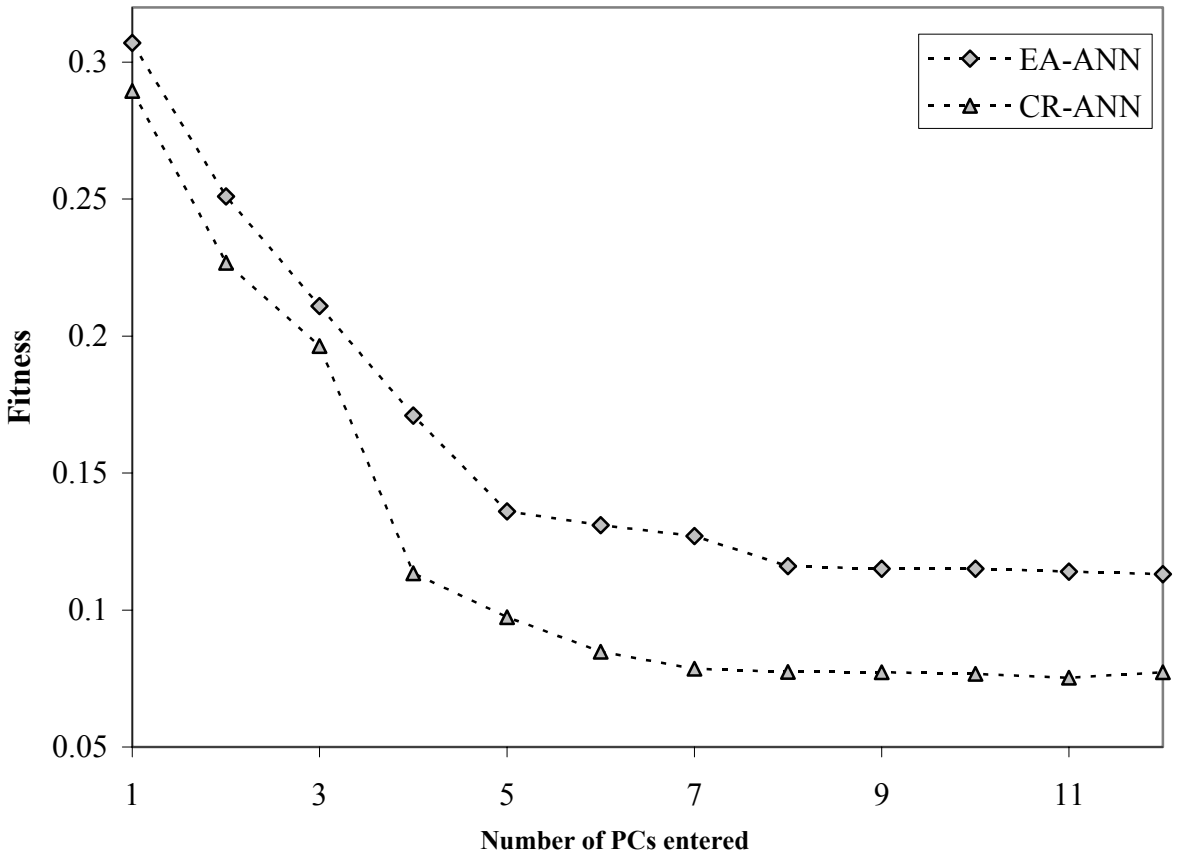

Figure 3. Plot of the variation of fitness as a function of the number of PC entered in the EA-ANN and CR-ANN models.