QSAR Modeling of 1-(3,3-Diphenylpropyl)-Piperidinyl Amides as CCR5 Modulators Using Multivariate Adaptive Regression Spline and Bayesian Regularized Genetic Neural Networks

ARTICLE *in* QSAR & COMBINATORIAL SCIENCE · SEPTEMBER 2009

Impact Factor: 1.55 · DOI: 10.1002/qsar.200860136

Mehdi Jalali-Heravi

Sharif University of Technology

63 PUBLICATIONS **1,017** CITATIONS

SEE PROFILE

Ahmad Mani-Varnosfaderani

29 PUBLICATIONS **83** CITATIONS

SEE PROFILE

QSAR Modeling of 1-(3,3-Diphenylpropyl)-Piperidinyl Amides as CCR5 Modulators Using Multivariate Adaptive Regression Spline and Bayesian Regularized Genetic Neural Networks

Mehdi Jalali-Heravi * and Ahmad Mani-Varnosfaderani

Department of Chemistry, Sharif University of Technology, P. O. Box 11155-9516, Tehran, Iran
*e-mail: jalali@sharif.edu; Telephone: +98-21-66165315; Fax: +98-21-66012983

Keywords: Structure-activity relationship, Multivariate adaptive regression spline (MARS), Bayesian regularized neural network (BRANN), Genetic algorithm, CCR5, Bioinformatics

Received: October 4, 2008; Accepted: March 20, 2009

DOI: 10.1002/qsar.200860136

Abstract

This study deals with developing a quantitative structure-activity relationship (QSAR) model for describing and predicting the inhibition activity of 1-(3,3-diphenylpropyl)-piperidinyl derivatives as CCR5 modulators. Applying the multiple linear regressions (MLR) and its inability in predicting the inhibition behavior showed that the interaction has no linear characteristics. To assess the nonlinear characteristics of the inhibition activity artificial neural networks (ANN) was used for data modeling. In order to select the variables needed for developing ANNs, three variable selection algorithms were used: Stepwise-MLR, genetic algorithm-partial least squares (GA-PLS), and Bayesian regularized genetic neural networks (BRGNNs). R^2 and root mean square error (*RMSE*) values for training (t) and leave-one-out (LOO) procedures revealed that BRGNNs is a robust algorithm for the variable selection and regression method simultaneously. Due to the 'black box' limitation of neural networks, multivariate adaptive regression spline (MARS) technique was used for modeling. A prominent advantage of MARS with respect to ANN is its ability in interpreting of the results of the model. Q^2_{LOO} and R_t^2 (0.982 and 0.947) reveal that MARS can describe and predict inhibition activity of these modulators and is as robust as ANN. Because the MARS model can explain the activity of molecules, it is a useful model for designing novel CCR5 inhibitors.

1 Introduction

Acquired immune deficiency syndrome (AIDS) has become a deadly global disease. According to the joint United Nations programmed on HIV/AIDS (UNAIDS), over 39 million people are living with HIV/AIDS and about 22 million cases have died from it until now [1]. Current antiretroviral therapies (ARTs) against AIDS are generally based on reverse transcriptase inhibitors and protease inhibitors. Such therapies can control the spread of the virus and can lead to improved quality of life in patients, but they cannot eliminate the virus from the body and can have undesirable side effects. Several investigations have recognized that one very promising possible alternative approach would be to develop novel therapeutics that can prevent the entry of HIV-1 into its target cells and, hence, block the first crucial step of the infection process [2]. Following the discovery that HIV infection initiated by fusion of the virus with the target cell through the binding of the viral gp120 protein with the CD4 receptor protein and its co-receptors CCR5 and CXCR4, there has been considerable interest in developing novel ligands that can modulate the co-receptor conformations and, hence, ultimately block virus-cell fusion [3]. Investigation of screening hits, led to the identification of *N*-alkyl-*N*-[1-(3,3-diphenylpropyl) piperidin-4-yl]-2-phenyl-acetamides as potent, selective ligands for the human CCR5 chemokine receptor [4].

The present paper deals with QSAR modeling of CCR5 binding affinity data of substituted [1-(3,3-diphenylpropyl) piperidinyl amides. The data reported by Cumming et al. have been used as the model dataset for the present QSAR study [4]. We used artificial neural network (ANN) and multivariate adaptive regression spline (MARS) for modeling pIC_{50} values of fifty derivatives of these compounds.

Neural networks are one of the most popular data mining approaches and are well known for their ability to model nonlinear functions. The researches have shown that a neural network with a sufficient number of parameters can model any continuous nonlinear function accurately [5]. Some authors also showed that neural networks are valuable in fitting models to data containing interactions [5]. Perhaps the main disadvantage of the neural networks is the inability of users to understand or explain them. Because the neural network is very complex function, there is no way to summarize the relationship between independent and dependent variables with function that can be interpreted by data analysts.

MARS has been successfully introduced into different fields of science. In chemistry MARS has been applied to QSAR studies and the modeling of retention times in HPLC [6–8]. This technique is a regression based technique which allows the analyst to use procedures to fit models to large complex databases. Because the technique is a regression based, its output is a linear function that is readily understood by analysts. Thus the technique does not suffer from the 'black box' limitation of the neural networks [5].

A common problem for QSAR studies is choosing an optimal set of molecular descriptors. To overcome this problem a powerful variable selection technique is needed. There are some published papers suggesting that genetic algorithms (GAs) are useful in data analysis and have also been applied as feature selection method in regression techniques [9–12]. The combination of genetic algorithms with multiple linear regression (MLR), (GA-MLR) and partial least square (PLS), (GA-PLS) is used in some cases in QSAR and QSPR studies [13, 14]. These optimization algorithms depend on an assumed linear relationship between the dependent variable and one or more descriptors, while there may be a nonlinear relationship between them. On the other hand some researchers have used genetic algorithms-neural networks as a nonlinear feature selection method [15, 16]. In this case ANN acts as a nonlinear regression method and GA plays the role of a method which selects the best set of input variables for ANN. The GA runs until the ANN error function such as *RMSE* reaches to its optimum value. In the case of GA-ANN, the ANN parameters together with GAs parameters should be optimized. This means that the optimization of the genetic algorithm-based neural network models is a complex procedure.

In this study we have used automated bayesian regularization algorithm as a complementary algorithm for levenberg-marquardt because using this automated algorithm makes combination of genetic algorithm with ANN easy. However the main aim of this work was making comparison between MARS and BRGNNs as feature selection and modeling technique for predicting and describing activity of considered compounds. The results of this work are promising and show that MARS can model the data as robust as ANN and it is also interpretable and can be understood easily.

2 Methods

2.1 Multivariate Adaptive Regression Spline

MARS, introduced by Friedman in 1991 [17], is a multivariate nonparametric regression technique which models complex relationships that are difficult, if not possible, for other modeling methods to reveal. In a sense, MARS is based on divide-and-conquer strategy partitioning the training data sets into separate regions, each of which gets its own regression equation. This makes MARS particularly suitable for problems with high input dimensions. MARS can be considered as generalization of classification and regression trees (CART) [18], and is able to overcome some limitations of CART [19]. The MARS regression model is constructed by fitting basis functions to distinct intervals of independent variables. Generally, piecewise polynomials, also called spline, have pieces smoothly connected together. In MARS terminology, the joining points of knots are called knots, nodes, or breakdown points. These will be denoted by small letter *t*. For a spline of degree *q* each segment is a polynomial function. MARS uses two-sided truncated power function as spline basis functions, described by the following equations:

$$[-(x-t)]_-^q = \begin{cases} (t-x)^q, & \text{If} \quad x<t \\ 0, & \text{Otherwise} \end{cases} \tag{1}$$

$$[+(x-t)]_+^q = \begin{cases} (x-t)^q, & \text{If} \quad x \geq t, \\ 0, & \text{Otherwise} \end{cases} \tag{2}$$

Where $q \geq 0$ is the power to which the splines are raised and which determines the degree of smoothness of the resultant function estimate.

When $q=1$, which is the case in this study, only simple linear splines are considered. The MARS model of dependent variable Y (here pIC_{50} values) with M basis functions can be written as:

$$\hat{y} = \hat{f}_M(x) = a_0 + \sum_{m=1}^{M} a_m B_m(x) \tag{3}$$

Where $\hat{y}$ is the dependent variable predicted by the MARS model, a_0 is a constant, $B_m(x)$ is the m-th basis function, which may be a single spline function or a product (interaction) of two or more spline basis functions, and a_m is the coefficients of m-th basis function.

Both the variables (here molecular descriptors) to be introduced into the model and the knot positions for each individual variable have to be optimized. For a data set X, containing n molecule and p descriptors, there are $N = n \times P$ pairs of spline basis functions, given by Equation 1 and 2, with knot location $x_{ij}(i=1, 2, ..., n; j=1, 2...p)$.

A two-step procedure is followed to construct the final model; two-at-a-time forward basis function selection and backward basis function deletion. This forward stepwise selection of basis functions leads to a very complex and overfitted model, such a model although fits the data well, but has poor predictive abilities for new objects. To overcome this problem, the redundant basis functions are removed one at a time using a backward stepwise procedure. To determine which basis function should be included in the model, MARS utilizes the generalized cross-validation (GCV) [19]. The GCV is mean squared residual error divided by a penalty dependent on the model complexity. The GCV is defined in the following way:

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_M(x_i)\right)^2}{\left(1 - {C(M)}/{n}\right)^2} \tag{4}$$

Where the $C(M)$ is the complexity penalty that increases with the number of basis functions in the model and which is defined as:

$$C(M) = C(M+1) + dM \tag{5}$$

Where M is number of basis functions in Equation 3, and the parameter d is a penalty for each basis function included into the model. Larger value of d leads to fewer basis functions and therefore smoother function estimates. In our study the parameter d were set to be 10, and the maximum interaction level of spline basis functions restricted to zero. For detailed descriptions about the selection of d parameter and determination of interactions in MARS model see reference [17].

2.1.1 The Importance of Variables in the MARS Model

Once the model is built, it is possible to estimate, on scale between 0 and 100, the relative importance of a variable in terms of its contribution to fit the model. To calculate the relative importance of a variable, we delete all terms containing the variable in question, refit the model, and then calculate the reduction in the fit. The most important variable is the one that, by its deletion, the reduction in the fit is maximum. Less important variables receive lower scores.

2.2 Artificial Neural Network

ANN is an information-processing paradigm inspired by the densely interconnected, parallel structure of brain processes information. Learning in biological system involves adjustment to the synaptic connections that exist between the neurons. Learning typically occurs by example through training, where the training algorithm iteratively adjusts the connection weights (synapses). These connection weights store the knowledge necessary to solve specific problems. Because artificial neural networks are not restricted to linear correlations, they can be used for nonlinear phenomena or curved manifold. Back propagation neural networks (BNNs) are most often used in analytical applications. The back propagation network receives a set of inputs, which is multiplied by each node and then a nonlinear transfer function is applied. The goal of training the network is to change the weights between the layers in direction to minimize the output errors. There are many algorithms for training multilayer perceptrons such as gradient descent, variable learning rate gradient descent, conjugate gradient descent, Newton algorithms, and faster quasi-Newton algorithms like levenberg-marquardt. In this study we used levenberg-marquardt algorithm for training the network. The detailed discussion about training algorithms for ANN can be found elsewhere [21].

2.2.1 Generalization Techniques in ANN

One of the problems that occur during the neural network training is called overfitting. The error on the training set is driven into small value, but when new data is presented to the network the error is large. There are two common methods for improving generalization in neural network and prevent overfitting: early stopping and regularization techniques.

2.2.2 Early Stopping Generalization

In this technique the available data is divided into two subsets. The first subset is the training set, which is used for updating the network weights and biases. The second set is testing set. The error on the testing set is monitored during the training process. Testing set error normally decrease during the initial phase of training, as does the training set error. However, when the network begins to overfit the data, the errors on the testing set begins to rise. When the test set error increases for a specified number of iterations, the training is stopped, and the weights and biases at the minimum of the testing set error are returned.

2.2.3 Regularization

This method involves modifying the performance function, which is usually chosen to be the mean sum of squares of the network errors (*mse*) on the training set.

$$mse = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 \tag{6}$$

It is possible to improve generalization if modifying the performance function by adding a term that consist the mean of the sum of squares of the network weights and biases. This performance function is named mean square error of regularization (*msereg*) and is defined by the following equation.

$$msereg = \alpha\ mse + \beta\ msw \tag{7}$$

Where α and β are performance ratios, and *msw* is the mean square of weights which can be calculated by using Equation 8.

$$mse = \frac{1}{n}\sum\nolimits_{j=1}^{n} w_j^2 \qquad (8)$$

By using this performance function the networks would have smaller weights and biases, and this force the network response to be smoother and less likely to overfit. The problem with regularization is that it is difficult to determine the optimum value for performance ratio parameter. One approach to determine optimum value of performance ratio is Bayesian framework of David MacKay [22, 23]. The detailed discussion of Bayesian regularization in combination with Levenberg-Marquardt training can be found in [23].

2.3 Feature Selection Methods

Feature selection refers to the problem of selecting input variables, otherwise called features that are relevant to predicting a target value for each instance in a dataset. These methods can either be used to rank all potentially relevant input variables or to build a good classifier, and each task may lead to a different methodological approach. Feature selection is a search problem, where each state in the search space corresponds to a subset of features.

Genetic algorithms have been applied as feature selection method in regression analysis. In these methods GA is being used for selection of the best solution according to evolutionary algorithms. One of the problems in ANN is inability of the network to select the best set of input variables among large set of possible solutions. In this study we used three feature selection methods for choosing input variables for ANN: Stepwise-MLR, GA-PLS and BRGNNs.

In our previous works we have used Stepwise-MLR and GA-PLS algorithm as feature selection methods for ANN [24, 25], therfore for the sake of brevity we restrain discuss about detailed methodology of these algorithms.

2.4 Bayesian Regularized Genetic Neural Networks (BRGNNs)

Bayesian regularization solves some of the well known problems of back propagated ANNs. Bayesian regularized artificial neural networks (BRANNs) have the potential to give models which are relatively independent of neural network architecture above a minimum architecture [26, 27]. The Bayesian regularization estimates the number of effective parameters which are lower than the number of weights. In this method the concerns about overfitting and overtraining are eliminated so that the definite and reproducible model is attained. In addition they are faster than standard neural networks. The joining of BRANNs with genetic algorithm feature selection is known as Bayesian regularized genetic neural networks (BRGNNs). In BRGNNs approach the error function of BRANNs is being used as objective function in genetic algorithm for optimization. In this case GA searches for input variables that lead to the less error in BRANNs.

3 Experimental

3.1 Data Set

The data set consists of 50 derivatives of 1-(3,3-diphenylpropyl)-piperidinyl amides and was taken from the article published by Cumming et al. [4]. The structure of these compounds is given in Figure 1. The inhibition activities expressed as pIC_{50} in terms of micromolar affinity for the investigated compounds and are given in Table 1. The values of pIC_{50} were used as dependent variable for developing the ANN and MARS models. In order to evaluate the generated models, leave-one-out cross-validation is used. In this algorithm, one compound is left in each step as prediction set and the model is developed using the remaining molecules as training set. For a further exhaustive testing the prediction power of the model, in addition to LOO-CV, leave-multiple-out cross-validation (LMO-CV) algorithm was also carried out. Here we have performed leave-7-out (L7O). A group of seven compounds was randomly selected from the data set and was left out. Then the pIC_{50} of this group was predicted by the model developed by using the remaining observations as the training set.

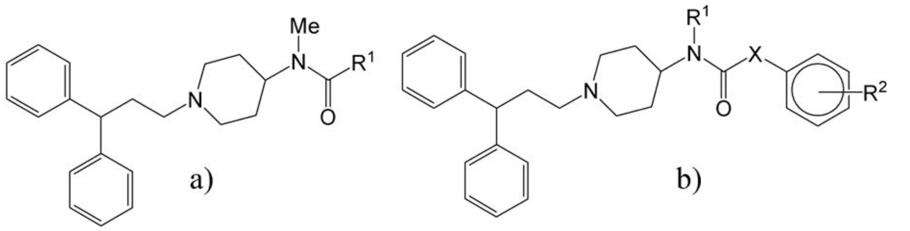
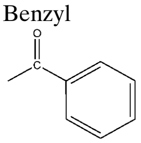
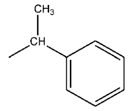


Figure 1. Structures of 1-(3, 3-diphenylpropyl)-piperidinyl amides as CCR5 inhibitors.

Table 1. Experimental and predicted pIC_{50} values for 1-(3,3-diphenylpropyl)-piperidinyl amides using MARS and BRGNNs.

Compound	Structure [a]	R^1	R^2	EXP. pIC_{50}	Pred. BRGNNs	Pred. MARS
1	1	H	–	5.387	5.327	5.775
2	1	4-Pyridinyl	–	5.215	5.168	5.014
3^{p}	1	4-phenyl	–	5.143	5.282	5.282
4	1	3-NO_2-phenyl	–	5.292	5.372	5.201
5	1	2-Thienyl	–	5.060	5.03	5.195
6	1	2-Furanyl	–	5.102	5.112	4.976
7^{p}	1	Cyclobutyl	–	5.131	5.115	5.094
8	1	Isobutyl	–	5.468	5.270	5.354
9	1	Neopentyl	–	5.260	5.374	5.402
10^{p}	1	Benzyl	–	6.092	6.181	6.005
11	1		–	5.229	5.457	5.681
12	1		–	5.168	4.893	5.182
13^{p}	2	Me	2-Cl	5.444	5.565	5.241
14	2	Me	3-Cl	5.568	5.853	5.696
15	2	Me	4-Cl	6.097	6.058	6.188
16	2	Me	3,4-di-Cl	6.108	5.842	5.796
17^{p}	2	Me	2,4-di-Cl	5.580	5.665	5.613
18	2	Me	2-F	5.721	5.565	5.671
19	2	Me	3-F	5.854	5.935	5.864
20	2	Me	4-F	6.180	6.136	6.411
21	2	Me	3,4-di-F	6.161	6.124	6.112
22^{p}	2	Me	3-OMe	6.168	6.293	5.992
23	2	Me	4-OMe	6.237	6.752	6.294
24	2	Me	3,4-di-OMe	6.187	6.248	5.848
25	2	Me	3,5-di-OMe	5.569	5.753	5.880
26^{p}	2	Me	2,4,5-tri-OMe	5.959	5.942	5.646
27	2	Me	4-Br	6.237	6.272	6.174
28	2	Me	4-Benzyloxy	5.456	5.380	5.523
29	2	Me	4-Phenyl	5.638	5.593	5.604
30	2	Me	4-CF_3	6.432	6.727	6.594
31^{p}	2	Me	4-OCF_3	6.538	6.233	6.629
32	2	Me	4-NHCOMe	6.168	6.180	6.218
33	2	Me	4-CN	7.222	6.934	7.024
34	2	Me	4-SO_2NH_2	7.041	7.023	6.808
35	2	Me	4-$SO_2N(Me)_2$	7.333	7.071	7.190
36^{p}	2	Me	4-SMe	6.252	6.167	6.117
37	2	Me	4-CO_2Me	6.201	6.246	6.458
38	2	Me	4-OH	6.328	6.138	6.310
39^{p}	2	Me	4-NO_2	6.824	6.878	6.620
40	2	Et	4-OCF_3	6.509	6.666	6.481
41	2	Et	4-CN	7.180	6.948	7.362
42	2	Et	4-SO_2NH_2	7.420	7.193	7.374
43^{p}	2	Et	4-SO_2Me	7.119	7.320	7.120
44	2	Et	4-NO_2	6.959	6.785	6.800
45	2	Cyclopropyl	4-SO_2NH_2	7.482	7.458	7.570
46	2	Cyclopropyl	4-SO_2Me	7.292	7.451	7.390
47	2	Cyclopropyl	4-NO_2	6.509	6.452	6.561
48	2	Allyl	4-OCF_3	6.456	6.467	6.686
49^{p}	2	Allyl	4-SO_2Me	7.432	7.688	7.329
50	2	Allyl	4-NO_2	6.745	6.758	6.672

[a] The structures are specified in Figure 1.

Table 2. Selected descriptors using stepwise-MLR technique. Statistics of model are: $R_t^2 = 0.885$, $R^2_{LOO} = 0.893$, $RMSE_{LOO'} = 0.246$, $RMSE_{t'} = 0.235$, $F = 85.46$.

Descriptor	Type of descriptor	Notation	Relative mean effect	Coefficient
Topological charge index of order 9	Galvez topological charge index	GGI9	36.78	0.781 ± (0.043)
Sum of topological distances between N…N	Topological descriptors	T(N...N)	4.22	012 ± (0.005)
Molecular multiple path count of order 9	Topological descriptors	piPC09	24.10	− 0.003 ± (0.000)
R maximal index	GETAWAY descriptors	RTv +	16.53	38.362 ± (7.199)
Lowest eigenvalue of Burden matrix	BCUT descriptors	BELv3	8.62	7.930 ± (2.161)
R maximal autocorelation of lag 5	GETAWAY descriptors	R5m +	9.75	− 9.866 ± (2.837)
Constant				12.328 ± (0.131)

3.2 Descriptor Generation

Molecular descriptors will probably play an increasing role in scientific growth. In fact, the availability of large number of theoretical descriptors, containing diverse source of chemical information, would be useful for a better understanding of the relationships between molecular structure and experimental evidence.

In this work, a total of 1497 descriptors of 0-, 1-, 2- and 3-dimensional, including constitutional, topological, geometric, molecular walk path counts, 2D autocorrelations, aromaticity indices, BCUT-descriptors, GALVEZ topological charge indices, Randic molecular indices, RDF, 3D-Morse, WHIM descriptors, GETAWAY, functional group counts, atom-centered fragments, charge, empirical and molecular properties were generated using Dragon(V.3.0) software. Detailed description of molecular descriptors can be found in reference 28. Descriptors which have more than 10% constant or zero values were eliminated and finally 1182 descriptors were remained for further investigations.

3.3 Linear Regression Analysis

A stepwise multiple linear regression model was built by using the calculated molecular descriptors. This method has been used for variable selection or model development in biological systems [29, 30]. $\mathrm{RMSE}_{\mathrm{LOO}}$ has been applied as fitness function for developing the stepwise-MLR model. It is clear that many MLR models will be resulted using stepwise multiple regression procedure; among them we have to choose the best one. It is common to consider four statistical parameters for this purpose. These parameters are the number of descriptors, correlation coefficient (R) and root mean square error ($RMSE$) for the training and validation procedures, and F statistic. A reliable MLR model is one that has high R^2 and F values, low $RMSE$ and number of descriptors. In addition to these, the model should have a high predictive ability. Consequently, among different models, the best model was chosen, whose specifications are presented in Table 2. As can be seen in this table, although this model has acceptable R_{CV}^2 (LOO) and R_t^2 values, but $RMSE_{CV}$ and $RMSE_t$ are inadequately high. As a result in the next section we have used artificial neural network for the investigation of nonlinear relationships.

3.4 Artificial Neural Networks

For generating a network as a regression model some parameters like number of nodes in hidden layer, type of input variables, type of transfer functions and optimum number of training iteration should be determined. In BRANNs, α and β are being automatically optimized to prevent overfitting and the optimum number of nodes in hidden layer can be determined using LOO-CV algorithm. The next important parameters which affect the response of the network are input variables and a method by which these variables are being selected. In this study three feature selection methods of Stepwise-MLR, GA-PLS and BRGNNs were applied for the construction of ANNs. In Stepwise-MLR-BRANNs approach, the Stepwise-MLR selected variables were used for the construction of BRANNs. In GA-PLS, the cross validation technique was used for evaluating the descriptors selected by GAs in each step. The data set is divided into q equal deletion groups (here q was set to be 5). Then each group was left out as a test set, and the PLS models were developed with remaining ($q-1$) groups as training set. This procedure is repeated until all objects have been predicted once. Then selected descriptors (here 6 descriptors) in each chromosome were evaluated by fitness function of PLS (here $RMSE_{LOO}$). Finally, in BRGNNs we used $RMSE_t$ of a 6-x-1 network as fitness function for optimization in genetic algorithms. In this framework, GA searches for the best inputs that lead to less $RMSE_t$ in BRANNs. This algorithm was repeated with various numbers of nodes (here x) in hidden layer and then $RMSE_{LOO}$ for each network is being determined. The network with the less $RMSE_{LOO}$ contains the best set of descriptors and optimum number of nodes in hidden layer for the modeling. Figure 2 represents $RMSE_t$ values for a 6-2-1 network during the optimization of BRANNs by GA. In BRGNNs approach some parameters such as number of nodes in ANN and population size, mutation parameters, crossover function, migration condi-

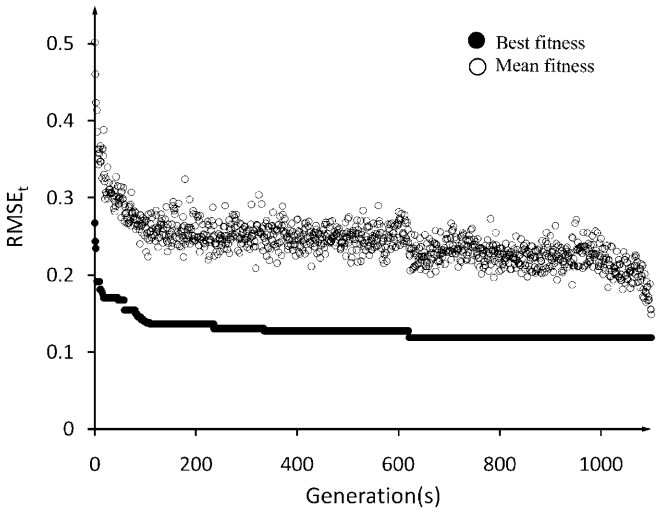


Figure 2. *RMSE*$_t$ values for the 6-2-1 network during the optimization by GA.

tion and hybrid local search methods in genetic algorithms should be optimized. Table 3 shows the specifications of the optimized BRGNNs.

According to the strategy of feature selection method, various types of variables were appeared for the construction of ANN. Table 4 represents the name and type of descriptors selected by GA-PLS and BRGNNs algorithms.

Table 3. Specifications of BRGNNs parameters.

BRGNNs parameters	Optimum condition
Number of generation	1100
Population size	50
Mutation function	Gaussian
Mutation scale	1.0
Mutation shrink	1.0
Migration condition	Forward
Cross over function	Scattered
Number of nodes in ANN	2
Hybrid function	Pattern search

3.5 The Procedures of MARS Modeling

A total of 1182 molecular descriptors were used as predictor variables to build the MARS models for the prediction of pIC_{50} values. The main steps of the MARS algorithm as applied here are as follows:

1. Selecting the maximal allowed basis functions for the model: we have used 30 basis functions for this step.
2. Starting with the simplest model, i.e. with the constant coefficient only.
3. Exploring the space of basis functions for each explanatory variable using two-at-a-time forward basis function selection and backward basis function deletion: At first, two left and right basis functions for each variable were generated and interred into model. Then GCV was calculated for this procedure, reduction in GCV causes entering these basis functions into the model and increasing in GCV causes to exclude them. Backward deletion step is a strategy for searching basis functions that deletion of them leads to reduction in GCV. By applying these processes on data matrixes the best set of basis functions were selected.
4. Selecting the best descriptors using MARS: With the strategy discussed in Section 2.1.1 the importance of variables can be calculated in MARS models. Since the space of calculating variables may be large, to speed of the MARS procedure one can use just the most important variables in the model. Figure 3 shows the importance of entered descriptors in primary MARS model which contains thirty basis functions. In the present work, the six most important variables were considered for further investigations.
5. Determining the optimum number of basis functions: In order to control the overfitting, optimum number of basis functions that are being generated with these six variables should be determined. For catching it, MARS algorithm was accomplished with maximum number of predetermined basis functions from 1 to 20 and for each step $RMSE_{LOO}$ was evaluated. By applying this method one can find the optimum number of basis functions for developing the best generalized model. The optimum

Table 4. Definition and notation of descriptors for BRGNNs and GA-PLS.

GA-PLS		BRGNNs	
Notation	Descriptor	Notation	Descriptor
GGI9	Galvez topological charge index of order 9	S1K	1-path Kier alpha index
MATS8e	Moran autocorrelation-lag 8 weighted by Sanderson electronegativity	CENT	Kier flexibility index
N-072	No. of RCO–N/N–X=X	Xindex	Centralization
SPP	Subpolarity parameter	PHI	Babalan X index
qnmax	Maximum negative charge	RDF055p	Radial distribution function weighted by polarizability
GATS7p	Geary autocorrelation – lag 7 weighted by atomic polarizability	Dp	D total accessibility index weighted by atomic polarizability

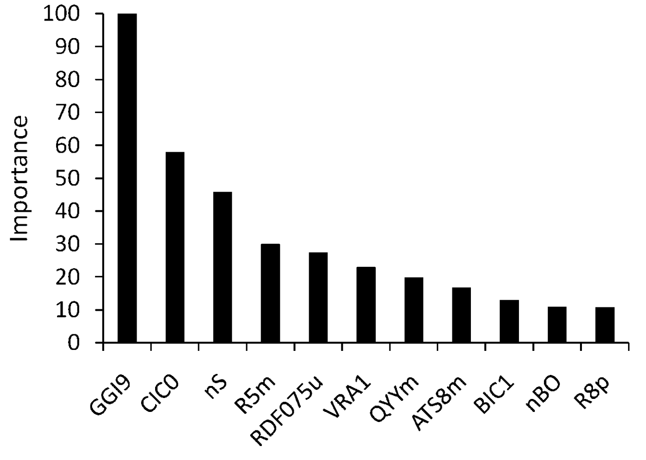


Figure 3. Importance of variables in the MARS model.

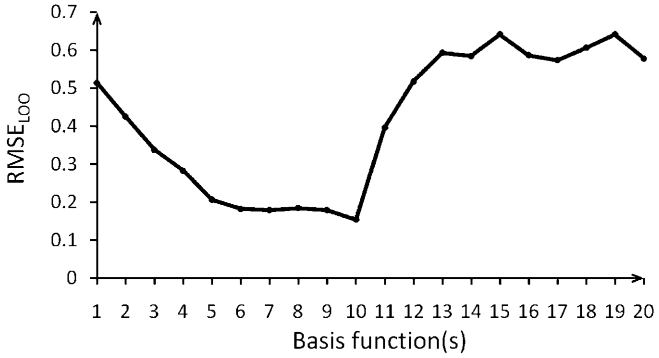


Figure 4. Determination of optimum number of basis functions by using RMSECV.

number of the basis functions was 10 in the present contribution. The plot is shown in Figure 4.

6. The 10 basis functions created here which leads to the less $RMSE_{\mathrm{LOO}}$ are the final MARS equations.

Before model building, the number of interactions between basis functions should be determined.

Therefore we performed MARS with various numbers of authorized interactions between the basis functions. However, the best statistical results were obtained when no interaction was entered into the model. MATLAB (version 7, Mathwork Inc.) [31] Toolboxes, (ANN toolbox, Genetic algorithm toolbox and ENTOOL toolbox), were used in this study for running regularized neural networks, Genetic algorithms and MARS.

4 Results and Discussion

4.1 Artificial Neural Networks

4.1.1 Discussion about Regularization and Early Stopping

In the case of early stopping procedure in ANN, the test set should be representative of all data points, therefore the choice of the test set is important. On the other hand, the best test set for obtaining the best training may vary when the input variables have been varied. But, Bayesian regularization does not require a test set, to be separated out of the training set; it uses all the data. This is an important advantage of this technique. The network uses all the data points for training and can have better prediction ability compared to early stopping technique. Table 5 represents statistical results for fifty times running of both early stopping and Bayesian regularization, for a 6-2-1 network with Stepwise-MLR descriptors as its input variables. It can be seen that Bayesian regularization is more reproducible with higher predictive ability compared to the early stopping generalization method.

The advantages such as better prediction ability, more reproducibility and adaptive generalization with variation of input variables to the network, makes Bayesian regularization-levenberg-marquardt (BR-LM) a suitable algorithm for the combination of the genetic algorithm with ANN.

4.1.2 Feature Selection Methods for ANN

Generally the performance of a model depends on the variables which are selected as its inputs. Table 6 represents the training and validation results for the three feature selection methods for selecting the best variables for the construction of a 6-2-1 network. The results revealed that BRGNNs is a robust algorithm for the feature selection and modeling, simultaneously. Stepwise-MLR and GA-PLS are based on assumed linear relationship between the descriptors and the activity of compounds, while BRGNNs considers nonlinear interactions and relationships in complex systems. The best and mean fitness values for training ($RMSE_t$) the 6-2-1 network in a specified generation, during optimization by GA, is shown in Figure 2 by solid and empty circles, respectively. This figure exhibits that after about 600 generations the best chromosome which con-

Table 5. Comparison between early stopping and Bayesian regularization algorithms for MLR-ANN.

	R_t^2 [a]	$RMSE_t$	SD_t [b]	$RMSE_{\mathrm{LOO}}$	SD_{RMSELOO}
Bayesian regularization	0.921	0.202	0.000	0.252	0.000
Early stopping	0.870	0.280	0.102	0.316	0.183

[a] The values are the averages of fifty times running of the algorithms. [b] Standard deviation for fifty times running of the algorithm.

Table 6. R^2 and *RMSE* values for training and validation sets of MARS and ANN models.

	Training		Validation			
	R_t^2	$RMSE_t$	Q^2_{LOO}	$RMSE_{LOO}$	Q^2_{L7O}	$RMSE_{L7O}$
MLR-ANN	0.912	0.210	0.861	0.268	0.902	0.233
GA-PLS-ANN	0.812	0.248	0.688	0.317	0.732	0.932
BRGNNs	0.981	0.112	0.946	0.167	0.956	0.177
MARS	0.982	0.106	0.948	0.158	0.948	0.184

tains the best solution is being selected and after about 1000 generations the algorithm has been converged.

4.2 The MARS Model

Because MARS algorithm contains some search methods like forward selection and backward-deletion, we used MARS as a feature selection method for developing the final model in this study. First, the best thirty basis functions among all possible $N = n \times P$ (n = number of molecules and P = number of descriptors) basis functions were selected. Then, the importance of variables in these basis functions was calculated and just the six most important descriptors were selected for building the final MARS model. In order to prevent the overfitting, the optimum number of basis functions was determined. These functions are built by using just the six most important descriptors. Table 7 presents the basis functions (B_m) and their coefficients (a_m) as final MARS model equations. Every basis function in MARS model is a regression equation for a definite region and several basis functions can be considered for each descriptor in a model. The MARS model obtained in this study contains ten basis functions built by the six entered descriptors. For each of variables CIC0 and VRA1, there are three basis functions while for each of GGI9, R5m, RDF075u and nS variables there is just one basis function. The detailed definition and theory for these descriptors can be found in Reference [28].

Table 7. The different basis functions (B_m), of the MARS model and their coefficients (a_m).

B_m	Definition [a]	a_m
B_0	1	2.305
B_1	$(0, \text{GGI9}-0.190)_+$	8.852
B_2	$(0, \text{nS}+0.00)_+$	1.015
B_3	$(0, \text{R5m}-0.296)_+$	−5.959
B_4	$(0, \text{RDF075u}-27.907)_+$	0.030
B_5	$(0, \text{VRA1}-306.918)_+$	−0.027
B_6	$(0, \text{VRA1}-339.513)_+$	8.852
B_7	$(0, \text{VRA1}-280.600)_+$	1.155
B_8	$(0, \text{CIC0}-4.624)_+$	−4.551
B_9	$(0, \text{CIC0}-4.687)_+$	0.030
B_{10}	$(0, \text{CIC0}-4.795)_+$	−0.027

[a] The subscript + indicates that the results of the function is 0 when the argument is not satisfied.

Several QSAR models can be found in the literature dealing with 1-(3, 3-diphenylpropyl)-piperidinyl amides and ureas [32, 33]. The main aspects of these models are summarized in Table 8. Contrary to the previous models, the present contribution is focused on the modeling of amides, rather than both of amides and ureas. On the other hand, the first twelve molecules in Table 1 have not been considered in the previous QSAR models. The statistical results obtained by MARS and BRGNNs for the fifty molecules of the present work are superior over those were reported previously. In the present work, the results of MARS are summarized in the form of ten basis functions, which provide the required knowledge for describing and predicting the activity of considered molecules.

4.3 Comparison Between Prediction Ability of MARS and BRGNNs

Both MARS and BRGNNs techniques use some strategies for preventing the overfitting during the construction of the model. In the procedure of MARS the GCV parameter is being optimized. This prevents increasing the number of basis functions and so prevents the overfitting. In the case of BRGNNs a compromise is made between the weights and *RMSE* by the Bayes theorem, which prevents occurrence of overfitting in the model. Therefore, in MARS and BRGNNs the GCV and Bayesian regularization (BR), respectively, prevent overfitting during the

Table 8. Comparison of the results of the present work with previous QSAR models. No. Comp.: Number of compounds; No. Des.: number of descriptors; PPR: project pursuit regression; SVM: support vector machine; nr: not reported.

Model	No. Comp.	Class of Comp.	No. Des.	$RMSE_{LOO}$	R_t^2	$RMSE_t$	Model summery	Ref.
Stepwise regression	79	Amides and Ureas	8	0.487	0.747	0.455 [a]	Linear-interpretable	[32]
FA-PLS	79	Amides and Ureas	13	0.437	0.797	nr	Linear-interpretable	[32]
PPR	79	Amides and Ureas	8	nr	0.837	0.345	Nonlinear-black box	[33]
SVM	79	Amides and Ureas	8	nr	0.867	0.308	Nonlinear-black box	[33]
Stepwise-MLR	50	Amides	6	0.246	0.885	0.235	Linear-interpretable	Present work
Stepwise-MLR-ANN	50	Amides	6	0.268	0.921	0.210	Black box	Present work
BRGNNs	50	Amides	6	0.167	0.981	0.112	Nonlinear-black box	Present work
MARS	50	Amides	6	0.158	0.982	0.106	Nonlinear-interpretable	Present work

[a] It is reported as standard error of estimate for training.

model building. After the development of the MARS model and optimization of BRANNs by using GA, in order to evaluate the robustness of the generated models, leave-one-out cross-validation (LOO-CV) method was applied. In LOO-CV, each object of the data is taken away one at a time and its pIC_{50} is predicted from the model developed using the remaining molecules as training set. The developed model by using the training molecules which has the minimum GCV (for training set) and minimum *msereg* (for training set) is used for predicting the activity of the left out molecule. It is noteworthy that the left-out molecule has no role in the determination of optimum GCV or optimum *msereg* of the developed models. This procedure is repeated until the activity of each molecule in the data set be predicted once by a model which the left-out molecule has no role in the determination of its parameters. Q^2, which is considered a measure of the model fit to cross-validation set, can be calculated as:

$$Q^2 = 1 - \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n} (y_{\text{exp}} - y_{\text{pred}})^2}{\sum_{i=1}^{n} (y_{\text{exp}} - \bar{y})^2} \tag{9}$$

Where PRESS is predictive sum of squares of the residuals and *SSY* is the sum of squares of the response variables corrected from the mean. The statistical parameters obtained by LOO-CV for the MARS and neural network-based methods are compared in Table 6. It can be seen from this table that statistical results for MARS and BRGNNs are better than other methods and also MARS can model the data as robust as ANN. Inspection of the results of this table reveals a higher R^2 and Q^2 values and lower RMSEs for the BRGNNs and MARS models. The calculated pIC_{50} values by MARS and BRGNNs using LOO-CV method are given in Table 1.

It is shown in the literature that ($Q^2 > 0.5$) can be considered as proof of the prediction ability of a model [34]. Several authors suggest that high value of Q^2 appears to be necessary but not sufficient for a model to have a good predictive power [35]. However, one may use a collection of compounds that were not being used for the building of the model as prediction set. We believe that applying only LOO-CV is not sufficient to evaluate the predictive ability of the model. Therefore for further evaluating the prediction ability of the models, L7O-CV has been also applied. The main feature of L7O-CV is similar to that of LOO-CV. In this case seven molecules were selected randomly from the data set and then a model was developed by using the remaining molecules to predict the activity of the left-out molecules. This procedure repeated until all molecules were examined as prediction set. The results of L7O-CV are presented in Table 6.

As a complementary method for evaluating the prediction ability of MARS and BRGGNs models, the data set was divided into training and prediction sets after sorting the pIC_{50} values. The training and prediction sets contain 38 and 12 molecules, respectively. The training set was used for model generation and the prediction set in which its molecules have no role in the model building was used for the evaluation of the predictive ability of BRGNNs and MARS models. Molecules labeled with notation «p» in Table 1 were included in the prediction set. Correlation (R^2) values for the prediction set were obtained to be 0.911 and 0.890 for MARS and BRGNNs, respectively. Also the RMSE of the prediction set were obtained to be 0.173 and 0.208, respectively for MARS and BRGNNs. Figure 6 represents the experimental versus the calculated values of activity using MARS for the training and prediction sets.

Statistical results for the calculation of the activity of 12 molecules as prediction set and L7O-CV reveal that both methods of MARS and BRGNNs can predict accurately the biological activity of the molecules. BRGNNs is a powerful method for searching the best variables in possible solutions according to nonlinear estimation of neural networks. However, MARS has been proposed as a viable competitor and alternative to neural networks that does not suffer from some of the limitations of the neural networks [36]. The obtained results reveal that MARS similar to BRGNNs can be used as a feature selection and regression method when analyzing complex structures, such as nonlinearities and interactions, which commonly can be found in a data set. However, unlike BRGNNs, MARS is not a 'black box', but produces models that can be explained and interpreted. Illustration of basis functions for CIC0 and VRA1 are presented in Figure 5. As can be seen in this figure, MARS can explain the nonlinear relationships between these variables and the activity of the molecules.

4.4 The Interpretation of the Results of the Models

BRGNNs is a neural network-based model and its interpretation is difficult because of its 'black box' limitation. Some methods such as square contribution value (SCV) and sensitivity analysis can be used for the determination of importance of a descriptor in three layer neural networks [37, 38]. Some other 'rule extraction' techniques have recently been designed for extracting the rules from the neural network [39, 40]. However, these methods are complicated and usually result in promiscuous rules which their understanding is difficult. On the other hand all of these techniques must be applied in addition to neural network, and the black box limitation of neural network remains an intrinsic property for ANN-based models. In contrast to neural network**,** MARS is not a 'black box' intrinsically, and produces some basis functions which clearly represent how a dependent variable is affected by independent variables. The basis functions produced by MARS can be considered as rules for describing the studied system. These rules are simpler than those obtained by the rule extraction methods from the neural network.

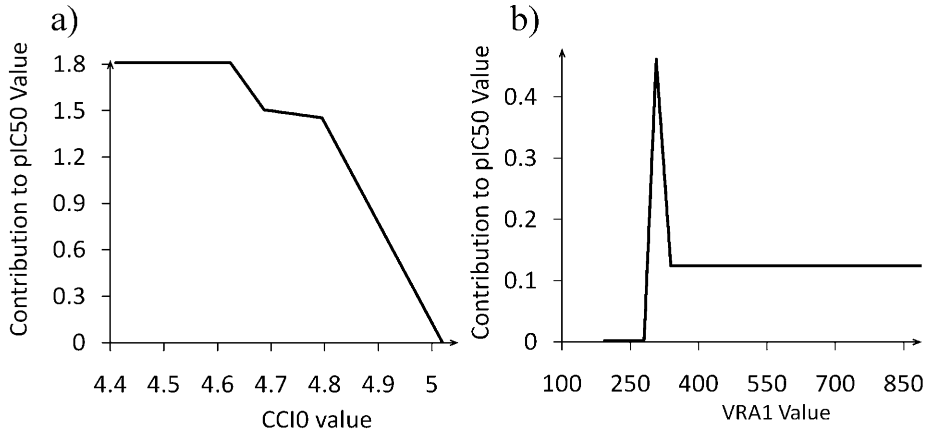


Figure 5. Illustration of basis functions for a) CIC0 and b) VRA1 variables.

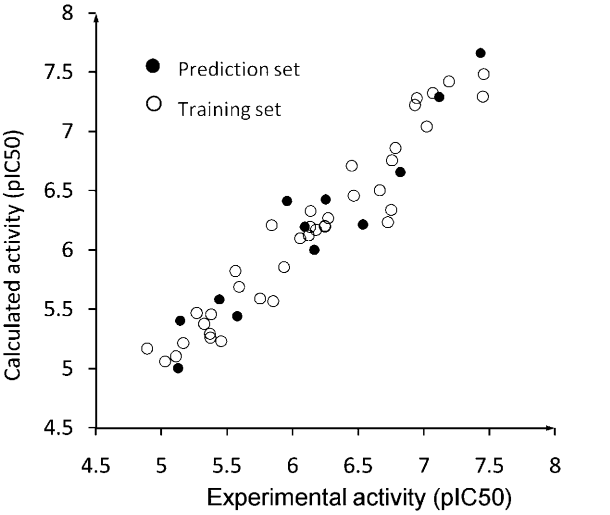


Figure 6. Plot of MARS calculated p*IC*50 values against the experimental ones for the training and validation sets.

Therefore, the results of MARS have been used in this work for interpreting the activity of the molecules. The results of MARS consisted of ten basis functions, built by six descriptors and can be considered as ten rules for describing the activity of the amide derivatives as CCR5 modulators.

The most important variable selected by MARS is Galvez topological charge index of order 9 (GGI9) which represents the total amount of charge transfer in the molecules [41]. This descriptor was also appeared in the Stepwise-MLR and GA-PLS variable selection methods and has the most relative mean effect and frequency of repetition in these methods, respectively. Therefore, this may be the most important parameter for describing the inhibition property of considered molecules. The basis function, in Table 7 shows the effect of this parameter on pIC50 values. The positive coefficient, α_1 and the knot point reveal that for GGI9 values greater than 0.19 the activity of the molecules increases. Since the addition of electron acceptor groups increases local positive charges in the benzyl group, substitutions like $-SO_2Me$, $-SO_2NH_2$ and $-NO_2$ lead to increase the value of GGI9 and therefore resulted in a better inhibitory activity.

The parameter of CIC0 is defined as complementary information content (neighborhood of order 0) and is a topological descriptor [28]. Complementary information content of a system is being calculated by the following equation:

$$CIC0 = \log_2 A - \sum\nolimits_{i=1}^{G} P_i \cdot \log_2 P_i \qquad (10)$$

Where A is the total number of atoms in the molecule and P_i is probability of randomly selecting an atom. For the molecules with complex structure and various types of atoms the value of CIC0 is small but for the simple molecules with relatively the same type of atoms the value of CIC0 is large [42].

Figure 5a demonstrates that for *CIC*0 values less than 4.624, the effect of this parameter on pIC_{50} values is considerable and constant but for larger values its contribution to the activity reduces.

The descriptor of nS represents the number of sulfur atoms of the molecules. As can be seen from Table 7, there is only one basis function (B_2) for describing the effect of this parameter in the model. The knot point for this function is selected to be zero, and the coefficient of this parameter is positive, so, by increasing the number of sulfur atoms in the molecules, the activity increases.

The parameter of R5m is R-autocorrelation of lag 5 weighted by atomic mass and shows that how the mass of the atoms in a given molecule is distributed in the space. The basis function for this variable is represented by B_3 in Table 7. The knot point for this function is 0.296, and the coefficient is negative. Therefore, for R5m values greater than 0.296, the contribution of this parameter to the activity reduces.

The descriptor of RDF075u is calculated by summing up the probability distribution of finding the atoms in spherical volume of radius 7.5 Å. RDF descriptors meet all the requirements for a 3D structure descriptor; it is independent of the number of atoms, i.e., the size of the molecule, it is unique regarding the three-dimensional arrangements of the atoms and is invariant against translation and rotation of the molecules [43]. The coefficient and the knot point for B_5 in Table 7 reveal that, for RDF075u values greater than 27.907 this parameter has a positive effect on pIC_{50} values.

VRA1 is Randic-type eigenvector-base index from adjacency matrix. This variable is among topological descriptors and represents the index of branching in the molecules [44]. Figure 5b represents how this parameter affects the inhibitory activity. It can be seen that just a specific branching index can lead to valuable effect on modulating behavior of the molecules.

5 Conclusions

The main aim of this work was developing an interpretable QSAR model to describe and predict the inhibition activity of a group of amides as CCR5 modulators. In this study we have applied both ANN and MARS techniques as modeling methods. Different feature selection techniques were applied for ANN, and BRGNNs was identified as the best. Although BRGNNs can model the data nicely but interpretation of its results is difficult. However, it is shown in this work that the MARS is able to model the data as robust as ANN and also can be used as a feature selection method. But the main advantage of MARS with respect to ANN is its ability in interpreting the parameters of the model. It is also able to clearly show how the dependent variable is related to the descriptors. Because the real three-dimensional structure of the human CCR5 is still unsolved [45], so the obtained results can be used as excessive information for the clarification mechanism of inhibition and studding the effects of various substitutions on the activity of 1-(3,3-diphenylpropyl)-piperidinyl amides and designing the new modulators of human CCR5. The importance of descriptors was calculated and the most important variable was found to be GGI9. It represents the total amount of charge transfer in a molecule and proposes that inhibition mechanism of these compounds is an electrical based procedure.

6 References

[1] World Health Organization AIDS Update 2008, http://www.who.int/hiv/en/ (accessed **2008**)

[2] S. Jiang, K. Lin, N. Strik, A. R. Neurath, *Biochem. Biophys. Res. Commun.* **1993**, *195*, 533 – 538.

[3] W. M. Kazmierski, J. P. Peckman, M. Duan, Curr. *Med. Chem. Anti-Infect. Agents* **2005**, 133 – 152.

[4] J. N. Burrows, J. G. Cumming, S. M. Fillery, G. A. Hamlin, J. A. Hudson, R. J. Jackson, S. Mc. Laughlin, J. S. Shaw, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 25 – 28.

[5] L. P. J. Veelenturf, *Analysis and Applications of Artificial Neural Networks*, Prentice Hall International, London **1995**, pp. 2 – 5.

[6] V. Neguyen-Cong, G. V. Dang, B. M. Rode, *Eur. J. Med. Chem.* **1996**, *31*, 797 – 803.

[7] Y. Fan, L. Shi, K. W. Kohn, Y. Pommier, J. N. Weinstein, *J. Med. Chem.* **2001**, *44*, 3254 – 3263.

[8] Q. S. Xu, D. L. Massart, Y. Z. Liang, K. T. Fang, *J. Chromatogr. A* **2003**, *98*, 155 – 167.

[9] M.Varcko, S. C. Basak, K. Geiss, F. Witzmann, *J. Chem. Inf. Model*, **2006**, *46*, 130 – 146.

[10] S. J. Cho, M. A. Hermsmeier, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 775 – 781.

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

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

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

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

[15] M. P. Gonzalez, J. Caballero, A. Tundidor-Camba, A. M. Helguera, M. Fernandez, *Bioorg. Med. Chem.* **2006**, *14*, 200 – 213.

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

[17] J. H. Friedman, *Ann. Stat.* **1991**, *19*, 1 – 141.

[18] L. Breiman, J. H. Friedman, R. A. Olshen, C. G. Stone, Wadsworth International Group, Belmont, CA **1984**.

[19] R. Put, Q. S. Xu, D. L. Massart, Y. V. Heyden, *J. Chromatogr. A* **2004**, *1055*, 11 – 19.

[20] P. Craven, G. Wahba, *Numer. Math.* **1979**, *31*, 317 – 403.

[21] M. Jalali-Heravi, M. Asadollahi-Baboli, P. Shahbazikhah, *Eur. J. Med. Chem.* **2008**, *43*, 548 – 556.

[22] D. J. C. Mackay, *Neural Comput.* **1992**, *4*, 415 – 447.

[23] D. J. C. Mackay, *Neural Comput*, **1992**, *4*, 448 – 472.

[24] M. Jalali-Heravi, A. Kyani, *Eur. J. Med. Chem.* **2007**, *42*, 649 – 659.

[25] M. Jalali-Heravi, A. Kyani, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1328 – 1335.

[26] J. Caballero, A. Candidor-Camba, M. Fernandez, *QSAR Comb. Sci.*, **2007**, *26*, 27 – 40.

[27] M. Fernandez, J. Caballero, *J. Mol. Graph. Model* **2006**, *25*, 410 – 422.

[28] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry*, Wiley-VCH, Weinheim **2000**.

[29] J. T. Leonard, K. Roy, *Bioorg. Med. Chem.* **2006**, *14*, 1039 – 1046.

[30] M. K. Gupta, R. Sagar, A. K. Shaw, Y. S. Prabhakar, *Bioorg. Med. Chem.* **2005**, *13*, 343 – 351.

[31] MATLAB, The Math Work Inc., http://www.mathworks.-com.

[32] J. T. Leonard, K. Roy, *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4467 – 4474.

[33] Y. Yuan, R. Zhang, R. Hu, X. Ruan, *Eur. J. Med. Chem.* **2008**, 1 – 10.

[34] S. Wold, *Quant. Struc. Act. Relat.* **1991**, *10*, 191 – 193.

[35] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model*, **2002**, *20*, 269 – 276.

[36] Y. Zhou, H. Leung, *J. Syst. Software*, **2007**, *80*, 1349–1361.
[37] P. B. Harrington, A. Urbas, C. Wan, *Anal. Chem.* **2000**, *72*, 5004–5013.
[38] R. Guha, P. C. Jurs, *J. Chem. Inf. Model*, **2005**, *45*, 800–806.
[39] C. J. Mantas, *Soft Comput.* **2008**, *12*, 493–512.
[40] J. S. Heh, J. C. Chen, M. Chang, *Neural Comput. Appl.* **2008**, *17*, 297–309.
[41] E. Deconinck, M. H. Zhang, F. Petitet, E. Dubus, I. Ijjaali, D. Coomans, *Anal. Chim. Acta* **2008**, *609*, 13–23.
[42] X. L. Huan, R. Sh. Zhang, X. J. Yoa, M. C. Liu, Z. D. Hu, C. T. Fan, *Anal. Chim. Acta* **2004**, *525*, 31.
[43] P. M. Gonzalez, C. Teran, M. Teijeria, M. A. Heluera, *Eur. J. Med. Chem*. **2006**, *41*, 52.
[44] T. A. Babalan, D. Ciubotariu, M. Medeleanu, *J. Chem. Comput. Sci*. **1991**, *31*, 517.
[45] L. Vangelista, L. Secchi, P. Lusso, *Vaccine* **2008**, *26*, 3008–3015.