New Hybrid Genetic Based Support Vector Regression as QSAR Approach for Analyzing Flavonoids-GABA(A) Complexes

Mohammad Goodarzi,[†] Pablo R. Duchowicz,*[,†] Chih H. Wu,[‡] Francisco M. Fernández,[†] and Eduardo A. Castro[†]

Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, C.C. 16, Suc.4, 1900 La Plata, Argentina,

and Department of Digital Content and Technology, National Taichung University No. 140, Ming-Shen Road, 40306 Taichung, Taiwan

Received February 26, 2009

Several studies were conducted in past years which used the evolutionary process of Genetic Algorithms for optimizing the Support Vector Regression parameter values although, however, few of them were devoted to the simultaneously optimization of the type of kernel function involved in the established model. The present work introduces a new hybrid genetic-based Support Vector Regression approach, whose statistical quality and predictive capability is afterward analyzed and compared to other standard chemometric techniques, such as Partial Least Squares, Back-Propagation Artificial Neural Networks, and Support Vector Machines based on Cross-Validation. For this purpose, we employ a data set of experimentally determined binding affinity constants toward the benzodiazepine binding site of the GABA (A) receptor complex on 78 flavonoid ligands.

INTRODUCTION

The basis of the Quantitative Structure Activity (QSAR) or Property (QSPR) Relationships Theory is the assumption that the variation of the behavior of chemical compounds, as expressed by any experimentally measured biological or physicochemical property on such compounds, can be correlated with numerical entities related to some aspect of the chemical structure termed molecular descriptors.[1−3] Descriptors are generally used to describe different characteristics/attributes of a certain structure in order to yield information about the activity/property being studied. QSAR-QSPR techniques are usually based on statistically determined linear or nonlinear models that relate the chemical behavior of compounds with descriptors.[4]

Various different factors find crucial importance in the QSAR-QSPR field: (a) the chosen set of descriptors employed, carrying suitable information of the molecular structure; (b) the modeling method employed; (c) the number of descriptors to be included in the model; (d) the composition of the training and test sets; and (e) the choice of the validation techniques to be applied. As will be seen later, in this work we use the Replacement Method (RM)[5−9] as a variable subset selection approach, as this technique has been successful for selecting relevant descriptors influencing several molecular properties/activities. We also make use of some linear and nonlinear methods such as Support Vector Machine (SVM), which was originally proposed by Vapnik et al. in 1995 and was then further developed to handle regression problems as Support Vector Regression (SVR).[10] It has been confirmed by a number of studies that SVM outperforms those widely used standard statistical tools, namely Multiple Linear Regression (MLR),[11] Hybrid Genetic Algorithm (HGA),[12] and Artificial Neural Networks (ANN)[13] in most of the cases or, at least, performs no worse than the others because of its characteristics of dimensional independence, restricted number of degrees of freedom, excellent generalization capability, global optimum, and ease of implementation.[14]

Several studies have proposed optimization methods which used a Genetic Algorithm (GA) for simultaneously optimizing the SVM parameter values, and the resulting hybrid GA-SVM methods were employed, i.e. by Fang et al.[15] in forecasting atmospheric corrosion of zinc and steel. In another work, Min et al.[16] pointed out that these hybrid models have improved on the previous single prediction models. In addition, GA has been increasingly applied in combination with other Artificial Intelligence techniques such as ANNs and Case-Based Reasoning (CBR). Their approach focuses on the improvement of the SVM-based model by integration of GA and SVM. However, few studies have focused on concurrently optimizing the type of SVM kernel function and the parameters of SVM kernel function although it has been shown that there is a great potential for useful applications in this area.[16] Present research work introduces a novel and specialized hybrid GA-SVR method and therefore constitutes one of the pioneering studies in using this optimization technique in the field of Chemical Informatics and Molecular Modeling. The results found are properly compared with other techniques such as Back Propagation Artificial Neural Network (BPN), Support

* Corresponding author phone: (+54)(221)425-7430/(+54)(221)425-7291; fax: (+54) 221 425 4642; e-mail: pabloducho@gmail.com/duchow@inifta.unlp.edu.ar.

[†] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA.

[‡] National Taichung University No. 140.

10.1021/ci900075f CCC: $40.75 © 2009 American Chemical Society
Published on Web 06/03/2009

Vector Machines based on Cross-Validation, and Partial Least Squares.

More than 4000 chemically unique flavonoids (phenylbenzopyranes) that have been isolated from various types of vascular plants during past years are becoming very popular due to their health promoting effects.[17] Some of the biological activities attributed to this family of chemical compounds involve the following: antiallergic, anticancer, antioxidant, anti-inflammatory, tranquilizers, and antiviral. Such kinds of compounds are important constituents of the human diet, being derived largely from fruits, vegetables, nuts, seeds, stems, and flowers, and thus constitute one of the most important classes of metabolites.

It is known that the overall balance between neuronal excitation and inhibition in the central nervous system (CNS) is due to the affinity of such type of ligands to the benzodiazepine binding site (BZD-bs) of the γ-aminobutyric acid type A (GABA(A)) receptor complex.[18] Chrysin (5,7-dihydroxyflavone) and Apigenin (5,7,40-trihydroxyflavone) are among the first compounds from the natural flavone family that demonstrated to possess a potent in vivo anxiolitic activity[19] and do not involve unwanted size-effects. This anxio-selective property of some flavones arises as a consequence of the absence of a simultaneous myorelaxant, amnestic, or sedative effect.[20] Despite their selectivity, flavonoid compounds often exhibit only moderate affinities in vitro. Several research groups have been able to generate synthetic flavone derivatives with higher affinities for the GABA(A) receptor, by means of the synthesis of small organic molecules libraries prepared by combinatorial chemistry, performed on solid or solution phases, and assisted with the molecular modeling of the flavonoid binding to the BZD-bs pharmacophore.[21−23]

In continuation with our efforts for analyzing the behavior of flavonoid compounds as effective psychoactive drugs,[24] we employ the proposed methodology for modeling the interaction of such type of ligands to the BZD-bs of the GABA(A) receptor complex. Up to now, only a few Quantitative Structure−Activity Relationships (QSAR) were developed for analyzing this interaction.[25−30] It has to be mentioned, however, that most of these QSAR involved only a few flavone derivatives during the training stage of the model, thus resulting in an incomplete description of the chemical universe. Another main interest of present research is to apply the so derived QSAR models for estimating the binding affinities of some new 2-,7-substituted benzopyranes,[31] for which there still are no experimental potencies. Up to now, few attempts have been carried out to synthesize flavonoids with substitutions of such types.

METHODS

The following subsections briefly describe the theory of Support Vector Regression (SVR), the Least Squares Support Vector Machine (LSSVM) Nonlinear SVR model, and Optimization of SVM Using Genetic Algorithms and Hybrid Genetic Algorithms.

Theory of Support Vector Regression. This subsection briefly introduces SVR, which can be used for time-series forecasting. Given training data $(\mathbf{x_1},\mathbf{y_1}), ...,(\mathbf{x}_l,\mathbf{y}_l)$, where $\mathbf{x_i}$ are the input vectors $(1 \times l)$ and $\mathbf{y_i}$ are the associated output values $(1 \times l)$ of $\mathbf{x_i}$, the support vector regression is an optimization problem

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\mathbf{w^T w} + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \quad (1)$$

restricted to

$$\mathbf{y_i} - (\mathbf{w^T}\phi(\mathbf{x_i}) + b) \le \varepsilon + \xi_i \quad (2)$$

$$(\mathbf{w^T}\phi(\mathbf{x_i}) + b) - \mathbf{y_i} \le \varepsilon + \xi_i^* \quad (3)$$

$$\xi_i, \xi_i^* \ge 0, i = 1, ..., l \quad (4)$$

where l denotes the number of samples, b is the bias term, vector of i-sample is data set mapped to a higher dimensional space by the kernel function ϕ vector, ξ_i represents the upper training error, and ξ_i*is the lower training error subject to ε-insensitive tube $|y - (\mathbf{w}^T\phi(\boldsymbol{x}) + b)| \le \varepsilon$. Three parameters determine the SVR quality: error cost C, width of tube, and mapping function (also called kernel function). The basic idea in SVR is to map the data set $\mathbf{x_i}$ into a high-dimensional feature space via nonlinear mapping. Kernel functions perform nonlinear mapping between the input space and a feature space. The approximating feature map for the Mercer kernel performs nonlinear mapping. In machine learning theories, the popular kernel functions are

In eq 5, $\mathbf{x_i}$ and $\mathbf{x_j}$ are input space vectors; in eq 6, t is the

$$\text{Linear kernel: } k(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i^T x_j} \quad (5)$$

$$\text{Polynomial kernel: } k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i^T x_j} + t)^d \quad (6)$$

$$\text{Gaussian (RBF) kernel: } k(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\frac{||\mathbf{x_i} - \mathbf{x_j}||^2}{2\sigma^2}\right) \quad (7)$$

intercept constant term and d the degree of the polynomial; in eq 7, σ^2 denotes the width of the Gaussian kernel.

The Least Squares Support Vector Machine Nonlinear SVR Model. This study uses LSSVM as a tool for our nonlinear SVR model. The motivation for choosing LSSVM as the approximation tool is its higher generalization capability as well as the achievement of an almost global solution within a reasonably short training time.[32] In feature space, the primary form of a LSSVM regression model in the optimization problem is formulated as

$$\min_{w,b,\xi} \frac{1}{2}\mathbf{w^T w} + \frac{1}{2}\gamma\sum_{i=1}^{l}\zeta_i^2 \quad (8)$$

restricted to

$$\mathbf{y_i} = \mathbf{w^T}\phi(\mathbf{x_i}) + b + \xi_i i = 1, 2, ..., l \quad (9)$$

where γ is a parameter and each error variable $\zeta_i \in R$ for $i = 1,2, ...,l$. In this nonlinear optimization problem, the Lagrangian is

$$L(\mathbf{w}, b, \xi; \alpha) = \frac{1}{2}\mathbf{w^T w} + \gamma\sum_{i=1}^{l}\xi_i^2 - \sum_{i=1}^{l}\alpha_i(\mathbf{w^T}\phi(\mathbf{x_i}) + b + \xi_i - y_i) \quad (10)$$

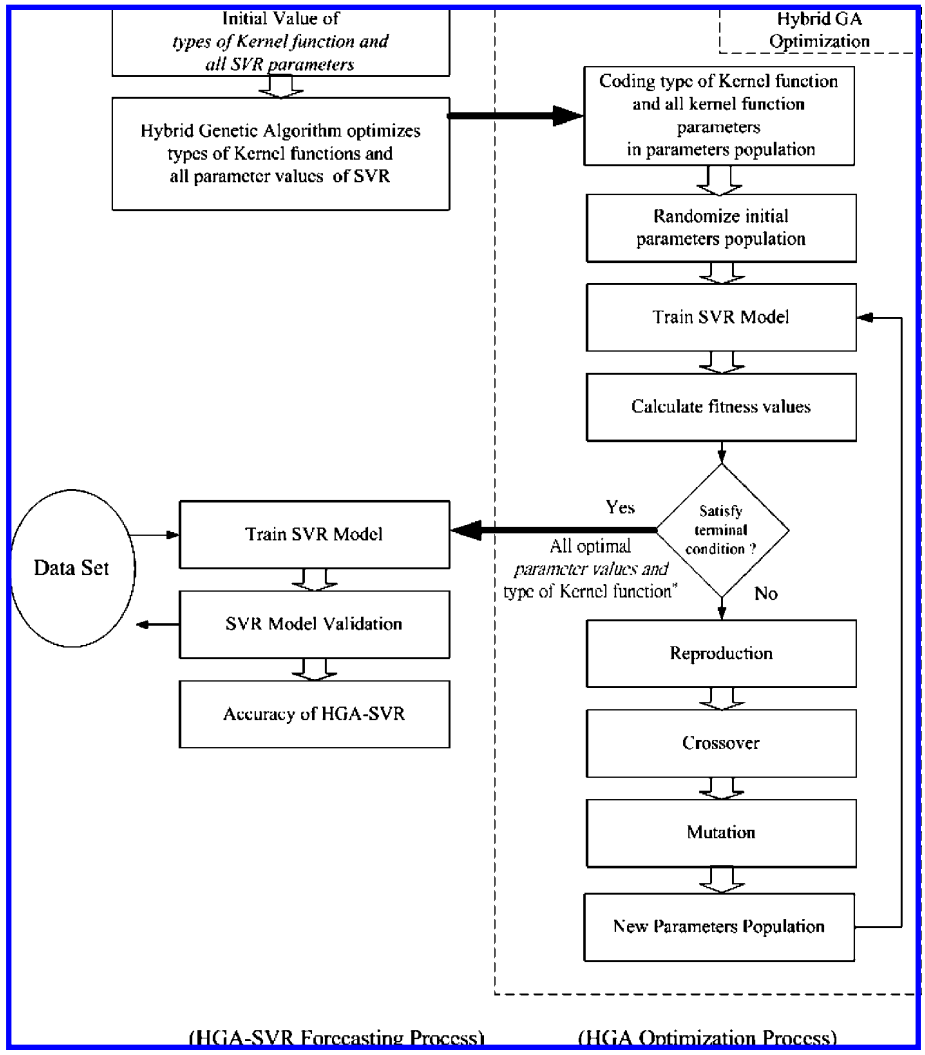

Figure 1. Schematic diagram for the optimization process involved in the HGA-SVR approach.

where the α_i are Lagrange multipliers. The first order conditions are

$$\frac{\partial L(\mathbf{w}, b, \xi; \alpha)}{\partial \mathbf{w}} = 0 \rightarrow \mathbf{w} = \sum_{i=1}^{l} \alpha_i \phi(\mathbf{x_i}) \tag{11}$$

$$\frac{\partial L(\mathbf{w}, b, \xi; \alpha)}{\partial b} = 0 \rightarrow \sum_{i=1}^{l} \alpha_i = 0 \tag{12}$$

$$\frac{\partial L(\mathbf{w}, b, \xi; \alpha)}{\partial \xi_i} = 0 \rightarrow \alpha_i = \gamma \xi_i, \quad i = 1, 2, ..., l \tag{13}$$

$$\frac{\partial L(\mathbf{w}, b, \xi; \alpha)}{\partial \alpha_i} = 0 \rightarrow \mathbf{w}^{\mathbf{T}} \phi(\mathbf{x_i}) + b + \xi_i - y_i = 0, \quad i = 1, 2, ..., l \tag{14}$$

We can eliminate $\mathbf{w}$ and ξ, and then the Karush-Kuhn-Tucker (KKT) system is obtained as

$$\begin{bmatrix} \mathbf{0}_{1\times l} & \mathbf{1}_{1\times l} \\ \mathbf{1}_{l\times 1} & \mathbf{K} + \gamma^{-1}\mathbf{E}_l \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{15}$$

Here $\mathbf{1}_{m\times n}$ denotes the $m \times n$ matrix of ones, $0_{m\times n}$ is the $m \times n$ matrix of zeros, $\mathbf{K}$ is the $l \times l$-dimensional kernel matrix of which the $i^{th}j^{th}$ entry is $[\mathbf{K}]_{ij} = \langle \phi(\mathbf{x_i}), \phi(\mathbf{x_j}) \rangle = K(\mathbf{x_i}, \mathbf{x_j})$,

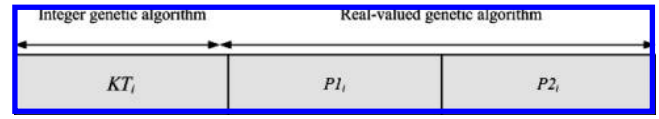

Figure 2. Gene structure for the HGA proposed in this work (population *i*).

and $\mathbf{E}_{nh}$ is the $nh \times nh$ identity matrix. Finally, the output of the LSSVM model in response to the input x is obtained from[32,33]

$$f(\mathbf{x}) = \sum_{k=1}^{l} \alpha_k^* K(\mathbf{x}_k, \mathbf{x}) + b^* \tag{16}$$

where α_k^* and b^* are optimal solutions of eq 15.

Optimization of SVM Using Genetic Algorithms and Hybrid Genetic Algorithms. As mentioned earlier, when designing an effective model, values of the two essential parameters in SVR have to be chosen carefully in advance.[34] These parameters include (1) regularization parameter C, which determines the trade-off cost between minimizing the training error and minimizing model complexity and (2) parameter σ (or d) of the kernel function, which defines the nonlinear mapping from the input space to some high-dimensional feature space. Generally speaking, model selection by SVM is still performed in the standard way: by

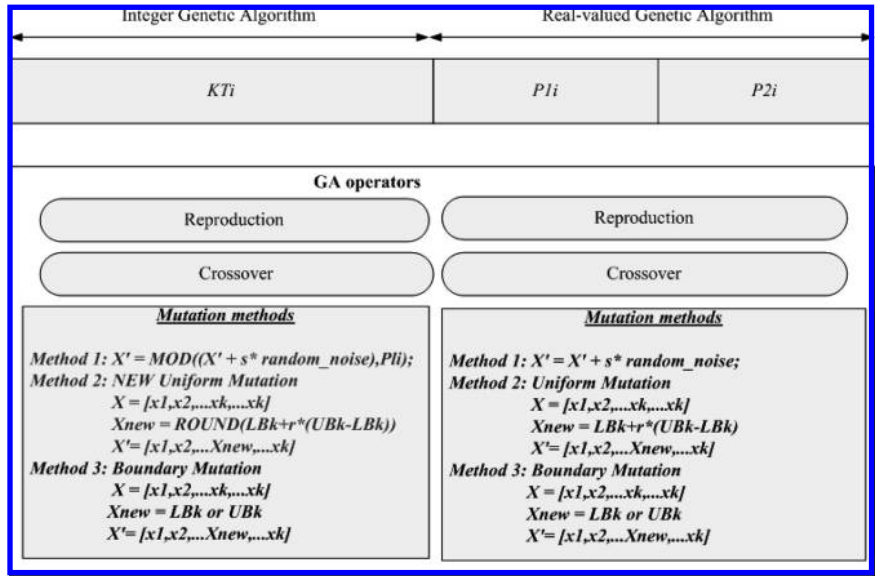

Figure 3. The new GA operators for the proposed HGA.

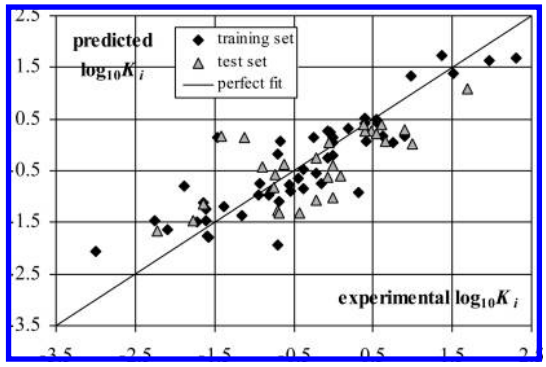

Figure 4. Predicted versus experimental log K_i for the training and test sets using the linear QSAR model.

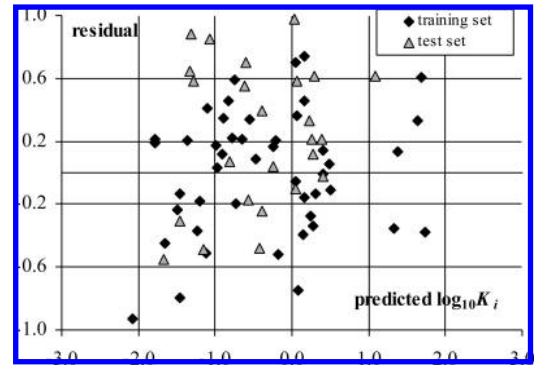

Figure 5. Dispersion plot of the residuals for the training and test sets achieved with the linear QSAR model.

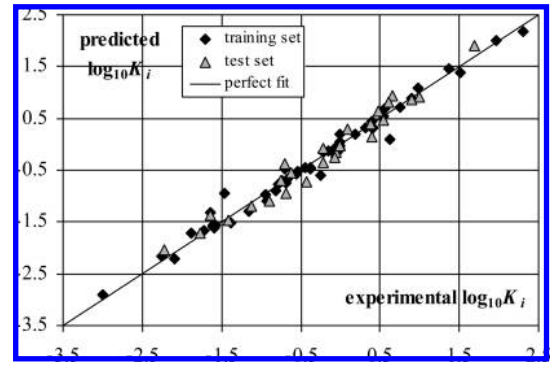

Figure 6. Predicted versus experimental log K_i for the training and test sets using the HGA-SVR model.

learning different SVMs and testing them on a validation set to determine the optimal value of the kernel parameters. Therefore, the Kernel-Adatron algorithm was proposed, which can automatically perform model selection without being tested on a validation.[35] Unfortunately, this algorithm is ineffective if the data have a flat ellipsoid distribution,[36] and a possible solution is to consider the data distribution.

Three various aspects of optimizing SVM have been proposed so far.[37] First of all appears "the kernel function and kernel parameters". Recently, GA can be adopted to automatically determine the optimal hyper-parameters for SVR to improve predictive accuracy and generalization ability. Pai and Hong[38] used GA to optimize the necessary parameters C, σ^2, and ε used in the Gaussian RBF kernel function of SVM. Howley and Madden[39] proposed a globally genetic kernel SVM to extend the aspects of SVM optimization for the parameters (e.g., C, σ^2, d, ε) via genetic GA. Wu et al.[40] proposed the simultaneous optimization approach, hybrid genetic-based SVR which integrates integer GA and real-valued GA in SVR, to optimize types of kernel function and its parameters simultaneously.

The second aspect of SVM optimization via GA is 'feature subset selection'. In most classification systems, the proper feature subset selection is able to enhance the classification performance and reduce computational requirements by selecting the most relevant small subset of features to the target data. Therefore, Lee and Byun[41] contributed a new face authentication system for memory-constrained devices for image identification. Sun et al.[42] use GA for searching a subset of eigenvectors encoding important information about the target concept of PCA for feature extraction and SVMs for classification. Li et al.[43] applied SVM-based with statistical testing and GA-based methods for feature selection. On the other hand, GA has been used to optimize both the feature subset and parameters of SVM simultaneously for bankruptcy prediction. Min et al.[16] presented a hybridizing method for the simultaneous optimization of SVM using GA for improving the predictive accuracy of SVM in two aspects: feature subset selection and parameter optimization.

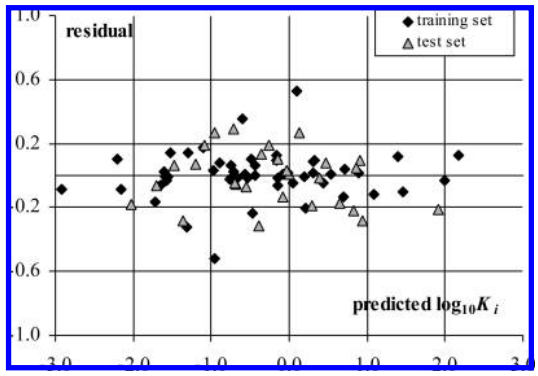

Figure 7. Dispersion plot of the residuals for the training and test sets with the HGA-SVR model.

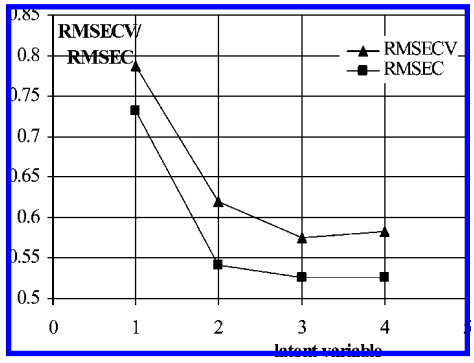

Figure 8. Determination of the optimum number of factors in the PLS method.

The last aspect of optimization is the "simultaneous optimization of kernel parameters and feature subset selection". This optimization approach may be more effective by simultaneously optimizing kernel parameters and feature selection, and these factors may affect the classification accuracy of SVM. Hence, various successful applications of this technique such as machinery fault detection,[44] SVM based Intrusion Detection System (IDS) with GA,[45] and protein sequence classification[46] have been proposed for improving performance of SVM. Besides these methods, Anh et al.[37] proposed a global optimization aspect that extends the simultaneous optimization to include one more factor-optimal instance selection by using GA in their study. The major idea of instance selection is that it selects appropriate reduced subsets from the data set and then is used as the training sample for SVM to prevent the distorted or noised training data. The advantage can be demonstrated in the applications of gear fault detection using ANN and SVR with GA[47] and GA-based prototype selection.[48]

A Novel Hybrid Genetic-Based Support Vector Regression (HGA-SVR). In our proposed novel HGA-SVR model, the types of kernel and parameter values of SVR are dynamically and simultaneously optimized by implementing the evolutionary process, and the resulting SVR model with these optimal values is then used for prediction. In the first stage, our approach tries to determine the appropriate type of kernel function for optimizing the SVR model, and then the RGA searches the optimal values of SVR parameters to enable SVR to fit various data sets. The overall process of our proposed approach is illustrated in Figure 1.

The types of kernel function and optimal values of SVR's parameters are searched by HGAs from a randomly generated initial population of chromosomes. The types of kernel function (Gaussian (RBF) kernel, Polynomial kernel, and Linear kernel) and all the values of parameters are directly coded in the chromosomes with integer and real-valued number, respectively. The proposed model implements the tournament method for selecting chromosomes. Adewuya's crossover method and boundary mutation method were used to modify the chromosome. The single best chromosome in each generation survives in the succeeding one. The proposed model was developed and implemented in the MATLAB 7.1.[49] The major tool, namely LIBSVM, for training and validating the SVM is the one developed by Pelckmans et al.[50] The SVR tool was also integrated with fuzzy weight preprocessing for medical decision making system and obtained the highest classification accuracy in their data set.[51] Thus, we believe that the proposed model is able to handle huge data sets and easily and efficiently be combined with the real-valued genetic algorithm in the MATLAB environment.

Table 1. Types of Various Kernel Function and Necessary Kernel Function Parameters[a]

KTi		*P1i* (parameter 1)	*P2i* (parameter 2)
0	Linear kernel	gamma	--
1	Polynomial kernel	*d*	*T*
2	RBF kernel	*C*	σ

[a] -- denotes no parameter needed; gamma, *d, t, C,* and σ denote various types of kernel function parameters.

Unlike applying traditional GAs, when using HGA for optimization problems all of the corresponding parameters and types of kernel function can be coded directly to form a chromosome. Hence, the representation of the chromosome is straightforward in HGA. All the parameters of SVR are directly coded to form the chromosome in the present approach. Figure 2 shows the gene structure of our proposed HGA. Consequently, chromosome X is represented as $X = \{KT, P_1, P_2\}$, where P_1 and P_2 denote the parameters of the kernel function (the first and second parameter values, respectively), while *KTi* denotes the types of kernel function which includes three types of them as indicated previously by eqs 5−7. The values zero, one, or two denote that the system would choose a 'Linear kernel', 'Polynomial kernel', or a 'Gaussian (RBF) kernel', respectively. The first part of the HGA will be implemented in the integer value type GA.

The various types of SVM kernel functions and the necessary kernel function parameters that need to be optimized are summarized in Table 1. The definition and type of essential parameters in SVR is based on the definition of the LSSVM tool. C (or γ) is the penalty (cost) parameter of the training error in the Linear/Polynomial/RBF kernel function. The parameter d denotes the degree of polynomial kernel function, t denotes the constant term of the polynomial kernel function, and σ denotes the squared root of the variance of the Gaussian kernel.

Another mathematical problem to solve involves various kernel function parameters with a different range of parameter values. Therefore, this research proposes the new GA operators in the novel HGA to deal with the range of the SVM parameter value. These new GA operators are shown in Figure 3, and, as can be appreciated, the HGA is divided into two parts—integer GA and real-valued GA. Our method selects the same GA reproduction operator and crossover operators. However, this study designs a different GA

mutation operator (i.e., Method 1 and Method 2 of Figure 3 differ in the real and integer parts) for limiting the range of parameter value. The revised mutation operator in *KTi* (new Method 1) is designed by MOD function calculation (remainder) and ROUND function calculation (convert the real-value into integer value) to limit the range of value. The revised mutation operator in *KTi* (new Method 2) is first used for the calculation via uniform mutation operators and then for converting the real value into integer value (the *KTi* value must be an integer value for mapping the coding design). Finally, we believe that it is unnecessary to redesign the boundary mutation which adopts upper and lower bounds.

MATERIALS

Experimental Data. The experimental binding affinity constants (K_i [μM]) of flavonoid ligands for the benzodiazepine site of the GABA(A) receptor complex are obtained from our previous QSAR study.[24] These data are converted into logarithm scale ($\log_{10}K_i$) for modeling purposes and are presented in Table 2. We partition the complete data set into a training set of 50 flavone derivatives and employ the rest of them as a test set (molecules **51**−**78** from Table 2) for assessing whether these data are correctly predicted by the best QSAR finally derived. Both sets are chosen in such a way they expand a similar range of variation of the observed K_i values.

Calculation of Molecular Descriptors. Parametric Method 3 (PM3)[52] optimized structures are taken from our previous publication, and the total pool of explored Dragon descriptors[53] consists of $D = 1176$ variables.[24] The initial conformations of the flavonoids are obtained by means of the "model build" modulus of the HyperChem program.[52]

RESULTS AND DISCUSSION

In recent years researchers have focused an increasing attention on finding the most efficient tool for variable selection in QSAR/QSPR studies. In fact, many techniques were employed that were not successful for all the cases under consideration, especially when the data set results large. One of the best methods is the Replacement Method,[5−9] which we employ in present research. Table 2 shows the experimental and predicted $\log_{10}K_i$ binding affinity constants of flavonoid ligands toward the GABA(A) receptor, calculated through several linear and nonlinear methods: RM, PLS, SVM, and ANN, for comparing with our developed HGA-SVM method.

We begin the analysis with the application of the RM variable subset selection technique on the training set composed of 50 flavone derivatives, in order to search for the optimal linear regression model that minimizes its standard deviation (*S*) and includes the best "representative" molecular descriptors, extracted from the pool containing $D = 1176$ variables. The following four-descriptors linear QSAR is obtained, which accomplish with the semiempirical "Rule of Thumb",[54] stating that at least five or six data points should be present for each fitting parameter

$$\log_{10}K_i = 103.240(\pm 15) - 8.541(\pm 2)\cdot J - 50.328(\pm 7)\cdot BELe2 - 5.650(\pm 0.7)\cdot MATS8v + 4.213(\pm 0.7)\cdot HATS7u \quad (17)$$

$$N = 50, R = 0.876, S = 0.554, p < 10^{-4}$$

$$R_{loo} = 0.845, S_{loo} = 0.615, R_{l-10\%-o} = 0.784, S_{l-10\%-o} = 0.728, S_{Rand} = 0.908$$

$$N = 28, R_{val} = 0.712, S_{val} = 0.712$$

In this equation, *N* is the number of compounds used, *R* is the correlation coefficient, *S* stands for the model's standard deviation of calibration, *p* is the significance of the model, and subindices *loo* and *l-10%-o* stand for the Leave-One-Out and Leave-10%-Out Cross Validation techniques, respectively.[55]

Both parameters of *loo* and *l-10%-o*, measure the internal validation of the developed QSPR upon inclusion/exclusion of compounds. The S_{Rand} parameter represents the standard deviation according to the Y-Randomization technique,[56] which is obtained by scrambling the experimental $\log_{10}K_i$ values in such a way that they do not correspond to the respective flavonoids. After analyzing 5000000 cases of Y-Randomization for eq 17, the smallest S_{rand} achieved is compared to the one found when considering the true calibration (*S*). Therefore, as $S_{rand} > S$, it is expected that the correlation given by eq 17 is not fortuitous and results in real structure−activity relationship. Subindex *val* stands for the test set statistical parameters. As can be appreciated, the QSAR model is predictive on 18 "fresh" test set compounds and thus is able to capture the essential structural features of the flavonoids that relate to the GABA response. Table 3 includes the correlation matrix for the four descriptors appearing in eq 17, revealing that the selected variables are not seriously correlated (maximum intercorrelation coefficient of 0.6). On the other hand, the Variance Inflation Factor (*VIF*)[57,58] is a method of detecting the severity of multicollinearity. *VIF* can be easily calculated as

$$VIF = \frac{1}{1 - R^2} \quad (18)$$

$$Tolerance = \frac{1}{VIF} \quad (19)$$

In practice, when $VIF > 5$ or higher, or if the tolerance remains under the value 0.20, then this would indicate that there exists multicollinearity among the descriptors. Obviously, as can be seen from Table 3, there is no multicollinearity problem on the selected subset of molecular descriptors.

Figure 4 plots the RM predictions of binding affinity constants for the 78 flavonoids included in the training and test sets, as a function of their experimental values. Figure 5 shows the plot of the propagation of residuals at both sides of the zero line for this model. The proposed linear QSAR model involves several two- and three-dimensional aspects of the molecular structure, which can be classified as follows: (i) topological: *J*, the Balaban *J* index;[59] (ii) a BCUT descriptor: *BELe2*, the lowest eigenvalue no. 2 of Burden matrix/weighted by atomic Sanderson electronegativities;[60] (iii) a 2D Autocorrelation: *MATS8v*, Moran autocorrelation-lag 8/weighted by atomic van der Waals volumes;[61] and finally, a GETAWAY descriptor: *HATS7u*, the leverage-weighted autocorrelation of lag 7/unweighted.[62]

Table 2. Experimental and Predicted Binding Affinities of Flavonoids Towards GABA (A) Receptor

no.	chemical compound	exp.	RM	PLS	SVM	BPN	HGA-SVM
1	6-fluoro-3′-methoxyflavone	0.398	0.507	0.505	0.386	0.264	0.442
2	6-bromo-3′-methoxyflavone	−0.215	−0.555	−0.571	−0.722	−0.466	−0.150
3	6-bromo-4′-methoxyflavone	0.322	−0.913	−0.928	0.352	0.337	0.308
4	6-chloro-2′-fluoroflavone	−0.380	−0.835	−0.856	−0.402	−0.401[a]	−0.439
5	6,3′-difluoroflavone	−0.036	0.243	0.233	−0.150	0.238	−0.162
6	6-chloro-3′-fluoroflavone	−0.932	−0.737	−0.760	−0.922	−0.664	−1.105
7	6-bromo-3′-fluoroflavone	−1.377	−1.193	−1.222	−1.222	−1.224	−1.519
8	4′-fluoroflavone	0.556	0.411	0.407	−0.040	0.392[a]	0.690
9	6,4′-difluoroflavone	0.398	0.410	0.399	−0.124	0.152	0.310
10	6,3′-dichloroflavone	−1.638	−1.124	−1.151	−1.159	−1.120	−1.317
11	3′-bromoflavone	−0.384	−0.471	−0.480	−0.595	−0.237	−0.483
12	6,3′-dibromoflavone	−1.721	−1.484	−1.514	−1.376	−1.562[a]	−1.658
13	6-bromo-4′-nitroflavone	−0.699	−1.947	−1.949	−0.752	−0.608	−0.719
14	6-bromoflavone	−1.155	−1.360	−1.382	−1.255	−1.397	−1.298
15	6-chloroflavone	−0.785	−0.904	−0.921	−0.952	−0.809	−0.762
16	6-nitroflavone	−0.561	−0.785	−0.771	−0.547	−0.214[a]	−0.567
17	6-methoxyflavone	−0.066	0.270	0.263	−0.043	0.005	−0.076
18	6-bromo-3′-nitroflavone	−3.000	−2.073	−2.071	−2.426	−3.059	−2.913
19	6-methyl-3′-nitroflavone	−2.252	−1.456	−1.418	−1.571	−1.813	−2.161
20	6-chloro-3′-nitroflavone	−2.097	−1.651	−1.644	−1.139	−2.075[a]	−2.202
21	6−3′-dinitroflavone	−1.585	−1.776	−1.738	−0.870	−1.534	−1.609
22	3′-nitroflavone	−0.545	−0.894	−0.869	−0.430	−0.499	−0.531
23	6-methyl-3′-bromoflavone	−1.886	−0.790	−0.787	−1.599	−1.467	−1.723
24	6-nitro-3′-bromoflavone	−1.602	−1.234	−1.231	−0.950	−1.182[a]	−1.594
25	6-methoxy-3′-bromoflavone	0.000	−0.205	−0.224	−0.510	−0.231	0.047
26	6−3′-dimethylflavone	−0.682	−1.095	−1.054	−0.221	−0.660	−0.743
27	5,2′-dihydroxy-6,7,8,6′-tetramethoxyflavone	−0.444	−0.655	−0.664	−0.344	−0.475	−0.442
28	2′-hydroxy-β-napthoflavone	−1.569	−1.780	−1.725	−1.053	−0.966[a]	−1.558
29	6,2′-dihydroxyflavone	−1.469	0.139	0.146	−1.121	−0.764	−0.952
30	2′-hydroxyflavone	−0.678	0.074	0.082	−0.542	−0.979	−0.664
31	7,2′.dihydroxyflavone	−0.252	0.141	0.149	−0.184	−0.803	−0.605
32	5,7-dihydroxy-8-methoxyflavone	0.182	0.316	0.318	0.500	−0.300[a]	0.188
33	6-hydroxyflavone	0.422	0.062	0.065	−0.094	0.519	0.325
34	7-hydroxyflavone	0.623	0.170	0.176	0.058	0.275	0.099
35	5,6,7-trihydroxyflavone	0.747	0.045	0.059	0.282	0.614	0.708
36	6-hydroxy-2′-methoxyflavone	0.976	1.330	1.339	1.199	1.587[a]	1.091
37	2′-methoxyflavone	1.508	1.374	1.384	1.248	1.600	1.394
38	2′-amino-6-methoxyflavone	0.544	0.487	0.494	0.443	0.489	0.535
39	Flavone	0.000	0.053	0.056	−0.100	0.400	0.203
40	5,3′,4′-trihydroxy-6,7-dimethoxyflavone	2.301	1.693	1.695	1.460	1.465[a]	2.174
41	5,4′-dihydroxy-6,7-dimethoxyflavone	1.362	1.742	1.742	1.462	0.833	1.461
42	5,7,4′-trihydroxy-6-methoxyflavone	0.000	0.155	0.157	0.129	0.235	−0.019
43	5,7-dihydroxy-2′-chloroflavone	0.903	0.166	0.176	0.337	0.460	0.889
44	5,7-dihydroxy-6,8-dibromoflavone	−0.155	−0.746	−0.743	−0.255	0.099[a]	−0.140
45	3,5,7,4′-tetrahydroxyflavone	1.969	1.634	1.638	1.483	2.192	2.002
46	6-nitro-4′-bromoflavone	−1.600	−1.466	−1.469	−1.192	−1.440	−1.570
47	6-chloro-3′-methoxyflavone	−0.072	−0.240	−0.251	−0.467	−0.206	−0.169
48	6-bromo-4′-fluoroflavone	−0.939	−0.970	−1.000	−1.146	−1.360[a]	−0.968
49	6-fluoro-3′-chloroflavone	−0.701	−0.179	−0.193	−0.485	−0.109	−0.468
50	6-bromo-3′-methylflavone	−0.812	−0.982	−0.975	−0.646	−0.513	−0.890
51	6-chloro-4′-methoxyflavone	0.097	−0.602	−0.612	0.084	0.175	0.284
52	6-bromo-2′-fluoroflavone	−0.424	−1.308	−1.334	−0.430	−0.512	−0.718
53	6-chloro-4′-fluoroflavone	−0.742	−0.570	−0.594	−0.876	−0.844	−0.688
54	3′-chloroflavone	−0.212	−0.251	−0.258	−0.435	−0.036	−0.076
55	6-fluoro-3′-bromoflavone	−0.627	−0.382	−0.398	−0.627	−0.271	−0.552
56	6-chloro-3′-bromoflavone	−1.638	−1.153	−1.179	−1.178	−1.357	−1.359
57	6-fluoroflavone	0.653	0.070	0.067	0.477	0.289	0.939
58	6-fluoro-3′-nitroflavone	−0.745	−0.817	−0.798	−0.474	−0.778	−0.694
59	6-hydroxy-3′-bromoflavone	0.000	−0.390	−0.399	−0.556	−0.191	−0.006
60	3′-methylflavone	1.000	0.026	0.053	0.375	0.637	0.902
61	5,7,2′-trihydroxy-6,8-dimethoxyflavone	−2.215	−1.663	−1.680	−1.173	−2.154	−2.034
62	5,7,2′-trihydroxy-6-methoxyflavone	−1.420	0.182	0.184	−0.914	−0.924	−1.479
63	5,7,2′-trihydroxyflavone	−1.125	0.153	0.167	−0.921	−0.899	−1.195
64	5,7-dihydroxy-6,8-dimethoxyflavone	−0.699	−1.280	−1.300	−1.009	−0.721	−0.386
65	5,7-dihydroxy-6-methoxyflavone	−0.051	0.049	0.051	0.125	−0.634	−0.150
66	5,7-dihydroxyflavone	0.477	0.267	0.281	0.356	0.407	0.650
67	5,7-dihydroxy-2′-fluoroflavone	0.903	0.289	0.298	0.394	0.448	0.858
68	5,7,4′-trihydroxyflavone	0.602	0.391	0.403	0.312	0.849	0.820
69	5-hydroxy-7-methoxyflavone	1.699	1.087	1.103	1.062	1.444	1.909
70	5,7-dihydroxy-6,8-diiodoflavone	0.000	−1.027	−1.025	−0.464	−0.401	−0.034
71	6-fluoro,3′-hydroxyflavone	0.400	0.280	0.280	0.004	0.669	0.135
72	6-chloro,3′-hydroxyflavone	−0.070	−0.620	−0.633	−0.074	−0.452	−0.261

Table 2. Continued

no.	chemical compound	exp.	RM	PLS	SVM	BPN	HGA-SVM
73	6-bromo,3′-hydroxyflavone	−0.220	−1.072	−1.090	−0.206	−0.118	−0.355
74	6-bromo-2′-nitroflavone	−0.680	−1.326	−1.325	−0.836	−0.867	−0.951
75	3′-methoxyflavone	0.380	0.403	0.406	0.164	0.374	0.395
76	3′-fluoroflavone	0.550	0.223	0.219	−0.072	0.405	0.471
77	6-bromo-3′-chloroflavone	−1.770	−1.462	−1.494	−1.377	−1.549	−1.703
78	6-methylflavone	−0.903	−0.421	−0.407	−0.897	−0.535	−1.092
79	7-methoxyflavone	−	0.561	0.575	0.417	0.597	0.531
80	7-chloroflavone	−	0.272	0.276	0.091	0.416	0.368
81	7-bromoflavone	−	0.280	0.286	0.129	0.391	0.239
82	2-(α-naphthyl)-4H-1-benzopyran-4-one	−	−0.649	−0.574	−0.468	−0.600	−0.638
83	2-(β-naphthyl)-4H-1-benzopyran-4-one	−	−2.611	−2.538	−1.806	−2.013	−1.872
84	7-bromo-2-(α-naphthyl)-4H-1-benzopyran-4-one	−	−1.089	−1.027	−0.719	−0.801	−0.934
85	7-bromo-2-(β-naphthyl)-4H-1-benzopyran-4-one	−	−2.241	−2.164	−1.782	−1.888	−1.652
86	7-chloro-2-(α-naphthyl)-4H-1-benzopyran-4-one	−	−0.833	−0.769	−0.831	−0.732	−0.826
87	7-chloro-2-(β-naphthyl)-4H-1-benzopyran-4-one	−	−2.288	−2.214	−1.831	−1.927	−1.867
88	7-methoxy-2-(α-naphthyl)-4H-1-benzopyran-4-one	−	−1.178	−1.097	−0.926	−1.162	−1.215
89	7-methoxy-2-(β-naphthyl)-4H-1-benzopyran-4-one	−	−2.093	−2.011	−1.607	−1.846	−1.617
90	7-fluoro-2-(α-naphthyl)-4H-1-benzopyran-4-one	−	−0.404	−0.336	−0.553	−0.617	−0.427
91	7-methyl-2-(α-naphthyl)-4H-1-benzopyran-4-one	−	−2.172	−2.079	−1.831	−1.688	−1.773
92	7-methyl-2-(β-naphthyl)-4H-1-benzopyran-4-one	−	−2.599	−2.502	−1.961	−1.857	−2.127

[a] Compound used as a member of the monitoring set in ANN.

Table 3. Correlation Matrix for Descriptors Appearing in the Linear QSAR Found by the RM Technique on 50 Training Flavonoids

-	*J*	*BELe2*	*MATS8v*	*HATS7u*	*Tolerance*	*VIF*
J	1	0.397	0.436	0.594	0.457	2.186
BELe2		1	0.054	0.096	0.759	1.318
MATS8v			1	0.367	0.740	1.351
HATS7u				1	0.617	1.621

Table 4. Architecture of the Back-Propagation ANN Used and Its General Specifications

parameters	detail
no. of nodes in the input layer	4 + 1[a]
no. of nodes in the hidden layer	6
no. of nodes in the output layer	1
learning rate	0.527
momentum	0.644
no. of iterations	37
transfer function	Sigmoid

[a] Bias.

The next step is to construct the Back Propagation ANN model. Before training the network, the input values are normalized between −1 and 1. The initial weights are randomly selected between −0.3 and 0.3, and after that the parameters of the nodes in the hidden layer, weights and biases learning rates, and momentum values are optimized. The proper number of nodes in the hidden layer is determined by training the network with different number of nodes in the hidden layer. The root-mean-square error (RMSE) value measures how good the outputs are in comparison with the target values. Table 4 shows the architecture and specifications of the optimized network. After optimization of the ANN parameters, the network is trained for the adjustment of weights and biases values. It should be noted that for evaluating the overfitting in all the nonlinear modeling methods, we separate some molecules from the training set, employing 38 of them for calibrating the model (76% out of 50 molecules, with range of activities from −3.0585 to 2.1919) and 12 molecules as a monitoring set (24% out of 50 molecules, with range of activities from −2.0751 to 1.4651). Therefore, the training of the ANN model of K_i must stop when the RMSE of the monitoring set begins to increase, while the RMSE of the training set continues to decrease.[63] In order to select the best weight update function we consider the prediction standard error of estimation (SEE).

The SVM prediction ability for regression is affected by several parameters: kernel type K, which determines the sample distribution in the mapping space and its corresponding parameters γ; capacity parameter C; and ε-insensitive loss function. All these three parameters are optimized using K-fold cross-validation. In this work, we use 5-fold cross-validation for optimizing the parameters and the final optimal model is determined to be $C = 46$, $\gamma = 0.819$, and $\varepsilon = 0.07$. The optimal value for ε depends upon the type of noise present in the data. Even though enough knowledge of the noise is available to select an optimal value for ε, there would always exist some practical consideration of the number of resulting support vectors. The ε-insensitivity prevents the entire training set to meet boundary conditions and allows the possibility of scarcity in the dual formulation solutions. This is the reason why choosing the appropriate value for ε is a critical step. Another involved parameter is the regularization C that controls the trade-off between maximizing the margin and minimizing the training error. If C is too small, then an insufficient stress would be placed on fitting the training data. On the other hand, if C is too large, then the SVM model would overfit the training data set. Finally, with the purpose of inspecting any possible interaction between C and ε, we vary the value of ε after performing the optimization of C. The results indicate that the value of the optimized ε does not change during this stage, thus concluding that both parameters are independent from each other.

Figure 6 shows the HGA-SVR predictions of binding affinity constants for the 70 training and test set flavonoids, as a function of their experimentally observed values. Figure 7 shows the plot of the propagation of residuals at

Table 5. Statistical Results Achieved for Different Linear and Nonlinear Techniques and Our Proposed HGA-SVM Model[a]

parameters	sets	RM	PLS	SVM	BPN	HGA-SVM
RMSEP	training set	0.5257	0.5258	0.3943	0.2983	0.1513
	test set	0.6460	0.6466	0.3938	0.2785	0.1731
RSEP(%)	training set	45.5501	45.5651	34.1691	27.0392	13.1111
	test set	69.9414	70.0096	42.6391	30.1535	18.7424
MAE(%)	training set	8.9371	8.9084	7.9102	7.9074	4.4308
	test set	13.6642	13.6314	10.4308	9.1250	7.2139
R^2	training set	0.7671	0.7669	0.9042	0.9208	0.9809
	test set	0.5065	0.5081	0.8846	0.9097	0.9637
F	training set	158.0549	157.9194	452.9992	418.5660	2468.016
	test set	26.6809	26.8542	199.211	261.9565	690.2593
Ttest	training set	12.5720	12.5666	21.2838	20.4589	49.6791
	test set	5.1654	5.1821	14.1142	16.1851	26.2728

[a] ANN monitoring set: $RMSEP = 0.4340$; $RSEP(\%) = 33.3306$; $MAE(\%) = 17.3606$; $R^2 = 0.880$ $F = 73.27789$, $Ttest = 8.560251$.

both sides of the zero line. Both figures illustrate that our novel HGA-SVR approach is a powerful technique for the QSAR prediction of the flavonoids - GABA(A) interaction.

We also apply the PLS method as this technique is one of the most popular linear methods.[64,65] The optimum number of factors (latent variables) to be included in the calibration model is determined by computing the RMSE on the training set (RMSEC) and cross-validation (RMSECV). A reasonable choice for the optimum number of factors would be the one that yields the minimum for the RMSEC/RMSECV ratio. However, the best latent variable is selected based on minimization of RMSECV and/or stabilization of RMSEC, which is shown in Figure 8. From this plot, it can be see that the RMSECV begins to increase after latent variable 3. Therefore, the optimum number of components for PLS is 3.

The statistical parameters for all the calculated models are included in Table 5, and, as can be appreciated, present results clearly indicate that the proposed novel HGA-SVR method achieves a much better statistical performance than the other ones. This conclusion is demonstrated by comparison of the *F* statistical, *t* test, correlation coefficient (R^2), root-mean-square error of prediction (*RMSEP*), relative standard error of prediction (*RSEP*), and mean absolute error (*MAE*) values.

In order to generate predictive models of general applicability that are not limited to operate as a mere correlation, the design of a properly validated model constitutes the most important step in the QSAR analysis. We apply a realistic validation process over the compounds which constitute the 'test set', with the main purpose of assessing whether the derived QSAR relationships are able to estimate the binding affinity values on independent test set compounds, that are not involved during the model fitting using the 'training set' compounds. As Table 2 demonstrates, the best predicted $\log_{10}K_i$ values in the test set correspond to our novel HGA-SVR procedure, thus demonstrating its acceptable predictive capability.

Finally, we apply the so derived HGA-SVR model to estimate the affinity $\log_{10}K_i$ on some derivatives synthesized by our colleagues.[31] Their list is denoted as 'estimation set' in Table 2, consisting on benzopyrane compounds substituted in positions 2 and 7. The predictions of binding affinity constants toward the benzodiazepine binding site of the GABA (A) receptor complex achieved for this estimation set of ligands are in line with the results found for our previously reported RM based linear QSAR.[24]

CONCLUSIONS

In this research we introduce for the first time the novel HGA-SVR hybrid method in the QSAR-QSPR field, and its statistical performance is then compared to other powerful linear and nonlinear chemometric tools such as Partial Least Squares, Back-Propagation Artificial Neural Networks, and Support Vector Machines. It is worth mentioning that when we compare the predictions achieved through HGA-SVR with those found in previous studies on the same class of flavone derivatives, we completely improve the results, both for the training set of compounds by which the HGA-SVR model is calibrated and also for the test set which is employed to validate the model and check its predictive power. On the other, based on the results using different linear and nonlinear methods, these indicated that this class of compounds has a nonlinear behavior, because nonlinear modeling methods behave better than linear ones. Among all the mathematical relationships analyzed in this work, the HGA-SVR is the most reliable and predictive one.

ACKNOWLEDGMENT

We thank the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) of Argentina. C. H. Wu thanks to the National Science Council (NSC97-2410-H-142-005) of Taiwan, Republic of China, R.O.C.

REFERENCES AND NOTES

(1) Hansch, C.; Leo, A. In *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(2) Verma, R. P.; Hansch, C. An Approach towards the Quantitative Structure-Activity Relationships of Caffeic Acid and its Derivatives. *ChemBioChem* **2004**, *5*, 1188–1195.
(3) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279–287.
(4) Goodarzi, M.; Freitas, M. P. Augmented three-mode MIA-QSAR modelling for a series of anti-HIV-1 compounds. *QSAR Comb. Sci.* **2008**, *27*, 1092–1098.
(5) Duchowicz, P. R.; Castro, E. A.; Fernández, F. M.; González, M. P. A new search algorithm for QSPR/QSAR theories: Normal boiling points of some organic molecules. *Chem. Phys. Lett.* **2005**, *412*, 376–380.

(6) Duchowicz, P. R.; Fernández, M.; Caballero, J.; Castro, E. A.; Fernández, F. M. QSAR of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. *Bioorg. Med. Chem.* **2006**, *14*, 5876–5889.
(7) Duchowicz, P. R.; González, M. P.; Helguera, A. M.; Cordeiro, M. N. D. S.; Castro, E. A. Application of the Replacement Method as Novel Variable Selection in QSPR. 2. Soil Sorption Coefficients. *Chemom. Intell. Lab. Syst.* **2007**, *88*, 197–203.
(8) Duchowicz, P. R.; Mercader, A. G.; Fernández, F. M.; Castro, E. A. Prediction of Aqueous Toxicity for Heterogeneous Phenol Derivatives by QSAR. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 97–107.
(9) Mercader, A. G.; Duchowicz, P. R.; Fernández, F. M.; Castro, E. A.; Bennardi, D. O.; Autino, J. C.; Romanelli, G. P. QSAR Prediction of Inhibition of Aldose Reductase for Flavonoids. *Bioorg. Med. Chem.* **2008**, *16*, 7470–7476.
(10) Cortes, C.; Vapnik, V. Support vector network, Machine Learning. *Machine Learning* **1995**, *20*, 273–293.
(11) Egan, W. J., Jr.; Merz, K. M.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43*, 3867–3877.
(12) Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Model.* **2003**, *43*, 1077–1084.
(13) Huuskonen, J.; Salo, M.; Taskinen, J. Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *J. Pharm. Sci.* **1997**, *86*, 450–454.
(14) Schölkopf, B.; Smola, A. In *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, 1st ed., MIT Press: Cambridge, MA, 2002.
(15) Fang, S. F.; Wang, M. P.; Qi, W. H.; Zheng, F. Hybrid genetic algorithms and support vector regression in forecasting atmospheric corrosion of metallic materials. *Comput. Mater. Sci.* **2008**, *44*, 647–655.
(16) Min, S.-H.; Lee, J.; Han, I. Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert Syst. Appl.* **2006**, *31*, 652–660.
(17) Andersen, O. M.; Markham, K. R. In *Flavonoids: Chemistry, Biochemistry and Applications*; CRC Press: Boca Raton, FL, 2006.
(18) Medina, J. H.; Viola, H.; Wolfman, C.; Marder, M.; Wasowski, C.; Calvo, D.; Paladini, A. C. Overview-Flavonoids: A New Family of Benzodiazepine Receptor Ligands. *Neurochem. Res.* **1997**, *22*, 419–425.
(19) Medina, J. H.; Paladini, A. C.; Wolfman, M.; de Stein, L.; Calvo, D.; Diaz, L. E.; Pena, C. Chrysin (5,7-dihydroxyflavone), a naturally-occuring ligand for benzodiazepine receptors, with anticonvulsant properties. *Biochem. Pharmacol.* **1990**, *40*, 2227–2231.
(20) Argyropoulos, S. V.; Nutt, D. The use of benzodiazepines in anxiety and other disorders. *Eur. Neuropsychopharmacol.* **1999**, *6*, s407–es412.
(21) Marder, M.; Viola, H.; Bacigaluppo, J. A.; Colombo, M. I.; Wasowski, C.; Wolfman, C.; Medina, J. H.; Ruveda, E.; Paladini, A. C. Detection of benzodiazepine receptor ligands in small libraries of flavone derivatives synthesized by solution phase combinatorial chemistry. *Biochem. Biophys. Res. Commun.* **1998**, *249*, 481–485.
(22) Dekermendjian, K.; Kahnberg, P.; Witt, M. R.; Sterner, O.; Nielsen, M.; Liljefors, T. Structure-activity relationships and molecular modeling analysis of flavonoids binding to the benzodiazepine site of the rat brain GABA(A) receptor complex. *J. Med. Chem.* **1999**, *42*, 4343–4350.
(23) Hong, X.; Hopfinger, A. J. 3D-Pharmacophores of Flavonoid Binding at the Benzodiazepine GABAA Receptor Site Using 4D-QSAR Analysis. *J. Chem. Inf. Model.* **2003**, *43*, 324–326.
(24) Duchowicz, P. R.; Vitale, M. G.; Castro, E. A.; Autino, J. C.; Romanelli, G. P.; Bennardi, D. O. QSAR Modeling of the Interaction of Flavonoids with GABA(A) Receptor. *Eur. J. Med. Chem.* **2008**, *43*, 1593–1602.
(25) Huang, X.; Liu, T.; Gu, J.; Luo, X.; Ji, R.; Cao, Y.; Xue, H.; Wong, J. T.; Wong, B. L.; Pei, G.; Jiang, H.; Chen, K. 3D-QSAR Model of Flavonoids Binding at Benzodiazepine Site in $GABA_A$ Receptors. *J. Med. Chem.* **2001**, *44*, 1883–1891.
(26) Blair, T.; Webb, G. A. Electronic factors in the structure-activity relations of some 1,4-benzodiazepin-2-ones. *J. Med. Chem.* **1977**, *20*, 1206–1210.
(27) Greco, G.; Novellino, E.; Silipo, C.; Vittoria, A. Study of Benzodiazepines Receptor Sites Using a Combined QSAR-CoMFA Approach. *Quant. Struct.-Act. Relat.* **1992**, *11*, 461–477.
(28) Gupta, S. P.; Paleti, A. Quantitative Structure - Activity Relationship Studies on Benzodiazepine Receptor Binding of Some Nonbenzodiazepine Series of Ligands. *Quant. Struct.-Act. Relat.* **1996**, *15*, 12–16.
(29) Marder, M.; Estiú, G.; Bruno-Blanch, L. E.; Viola, H.; Wasowski, C.; Medina, J. H.; Paladini, A. C. Molecular modeling and QSAR analysis of the interaction of flavone derivatives with the benzodiazepine binding site of the $GABA_A$ receptor complex. *Bioorg. Med. Chem.* **2001**, *9*, 323–335.
(30) Hadjipavlou-Litina, D.; Garg, R.; Hansch, C. Comparative Quantitative Structure-Activity Relationship Studies (QSAR) on Non-Benzodiazepine Compounds Binding to Benzodiazepine Receptor (BzR). *Chem. Rev.* **2004**, *104*, 3751–3793.
(31) Bennardi, D. O.; Romanelli, G. P.; Jios, J. L.; Vázquez, P. G.; Cáceres, C. V.; Autino, J. C. Synthesis of substituted flavones and arylchromones using P and Si Keggin heteropolyacids as catalysts. *Heterocycl. Commun.* **2007**, *13*, 77–81.
(32) Iplikci, S. Dynamic Reconstruction of Chaotic Systems from Interspike Intervals Using Least Squares Support Vector Machines. *Physica D* **2006**, *216*, 282–293.
(33) Ding, M.; Yang, W. Deterministic Point Processes Generated by Threshold Crossings: Dynamics Reconstruction and Chaos Control. *Phys. Rev. E* **1997**, *55*, 2397–2402.
(34) Duan, K.; Keerthi, S. S.; Poo, A. N. Evaluation of Simple Performance Measures for Tuning SVM Hyperparameters. *Neurocomputing* **2003**, *51*, 41–59.
(35) Cristianini, N.; Campell, C.; Taylor, J. S. Dynamically adapting kernels in Support Vector Machines. In *Advances in Neural Information Processing Systems (NIPS)*; MIT press: 1999; Vol. 11(2), pp 204−210.
(36) Campbell, C. Kernel Methods: A Survey of Current Techniques. *Neurocomputing* **2002**, *48*, 63–84.
(37) Ahn, H.; Lee, K.; Kim, K. J. Global Optimization of Support Vector Machines Using Genetic Algorithms for Bankruptcy Prediction . *International Conference on Neural Information Processing*; N°13, Hong Kong, China (2006), 2006; Vol. 4234, pp 420−429.
(38) Pai, P.-F.; Hong, W.-C. Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. *Electr. Pow. Syst. Res.* **2005**, *74*, 417–425.
(39) Howley, T.; Madden, M. G. The Genetic Kernel Support Vector Machine: Description and Evaluation. *Artif. Intell. Rev.* **2005**, *24*, 379–395.
(40) Wu, C. H.; Tseng, G. H.; Lin, R. H. A Novel Hybrid Genetic Algorithm for Kernel Function and Parameter Optimization in Support Vector Regression. *Expert Syst. Appl* **2008**, *42*, In Press. doi:10.1016/j.eswa.2008.06.046.
(41) Lee, K.; Byun, H. A New Face Authentication System for Memory-Constrained Devices. *IEEE T. Consum Electr.* **2003**, *49*, 1214–1222.
(42) Sun, Z.; Bebis, G.; Miller, R. Object detection using feature subset selection. *Pattern Recognit.* **2004**, *37*, 2165–2176.
(43) Li, L.; Tang, H.; Wu, Z.; Gong, J.; Gruidl, M.; Zou, J.; Tockman, M.; Clark, R. A. Data mining techniques for cancer detection using serum proteomic profiling. *Artif. Intell. Med.* **2004**, *32*, 71–83.
(44) Jack, L. B.; Nandi, A. K. Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms. *Mech. Syst. Signal Pr.* **2002**, *16*, 373–390.
(45) Kim, D. S.; Nguyen, H.-N.; Park, J. S. Genetic algorithm to improve SVM based network intrusion detection system. *Proceedings of the 19th International Conference on Advanced Information Networking and Applications* (AINA 2005), 28−30 March 2005, Taipei, Taiwan, IEEE Computer Society 2005, pp 155−158.
(46) Zhao, X.-M.; Cheung, Y. -M.; Huang, D. -S. A novel approach to extracting features from motif content and protein composition for protein sequence classification. *Neural Networks* **2005**, *18*, 1019–1028.
(47) Samanta, B. Gear fault detection using artificial neural networks and support vector machines with genetic algorithms. *Mech. Syst. Signal Pr.* **2004**, *18*, 625–644.
(48) Babu, T. R.; Murty, M. N. Comparison of genetic algorithm based prototype selection schemes. *Pattern Recognit.* **2001**, *34*, 523–525.
(49) *Matlab, Version 7.5*; MathWorks Inc.: Natick, 2007.
(50) Pelckmans, K.; Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; Lukas, L.; Hamers, B.; De Moor, B.; Vandewalle, J. *LS-SVMlab Toolbox User's Guide version 1.4;* November, 2002. Software available at http://www.esat.kuleuven.ac.be/sista/lssvmlab (accessed Apr 16, 2009).
(51) Comaka, E.; Polatb, K.; Güneşb, S.; Arslana, A. A new medical decision making system: Least square support vector machine (LSSVM) with Fuzzy Weighting Pre-processing. *Expert Syst. Appl.* **2007**, *32*, 409–414.
(52) *HyperChem, Version 7.0*; Hypercube, Inc.: Gainesville, 2007.
(53) Milano Chemometrics and QSAR Research Group. http://michem.disat.unimib.it/chm (accessed Apr 16, 2009).
(54) Hansch, C. In *Comprehensive drug design*; Pergamon Press: New York, 1990.
(55) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross Validation. *J. Chem. Inf. Model.* **2003**, *43*, 579–586.
(56) Wold, S.; Eriksson, L. In *Chemometrics methods in molecular design;* VCH: Weinheim, Germany, 1995.
(57) Curto, J. D.; Pinto, J. C. New multicollinearity indicators in linear regression models. *Int. Statist. Rev.* **2007**, *75*, 114–121.

(58) Johnson, R. A.; Wichern, D. W. *Applied multivariate statistical analysis*; Prentice-Hall: New York, 1992.
(59) Balaban, A. T. Highly discriminating distance based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.
(60) Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.
(61) Moreau, G.; Broto, P. Autocorrelation of a molecular structure: a new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359–360.
(62) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. *J. Chem. Inf. Model.* **2002**, *42*, 693–705.
(63) Goodarzi, M.; Freitas, M. P.; Jensen, R. Feature Selection and Linear/Non-Linear Regression Methods for the Accurate Prediction of Glycogen Synthase Kinase-3β Inhibitory Activities. *J. Chem. Inf. Model.* 2009In press. doi:10.1021/ci9000103.
(64) Goodarzi, M.; Freitas, M. P. Predicting Boiling Points of Aliphatic Alcohols through Multivariate Image Analysis Applied to Quantitative Structure-Property Relationships. *J. Phys. Chem. A* **2008**, *112*, 11263–11265.
(65) Goodarzi, M.; Freitas, M. P. On the use of PLS and N-PLS in MIA-QSAR: Azole antifungals. *Chemom. Intell. Lab. Syst.* **2008**, *96*, 59–62.

CI900075F