Machine Learning Techniques and Drug Design

J.C. Gertrudes[a], V.G. Maltarollo[b], R.A. Silva[a], P.R. Oliveira[a], K.M. Honório[a,b] and A.B.F. da Silva*[,c]

[a]Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 03828-0000, São Paulo, SP, Brazil; [b]Centro de Ciências Naturais e Humanas, Universidade Federal do ABC, 09210-170, Santo André, SP, Brazil; [c]Departamento de Química e Física Molecular, Instituto de Química de São Carlos, Universidade de São Paulo, CP 780, 13560-970, São Carlos, SP, Brazil

Abstract: The interest in the application of machine learning techniques (MLT) as drug design tools is growing in the last decades. The reason for this is related to the fact that the drug design is very complex and requires the use of hybrid techniques. A brief review of some MLT such as self-organizing maps, multilayer perceptron, bayesian neural networks, counter-propagation neural network and support vector machines is described in this paper. A comparison between the performance of the described methods and some classical statistical methods (such as partial least squares and multiple linear regression) shows that MLT have significant advantages. Nowadays, the number of studies in medicinal chemistry that employ these techniques has considerably increased, in particular the use of support vector machines. The state of the art and the future trends of MLT applications encompass the use of these techniques to construct more reliable QSAR models. The models obtained from MLT can be used in virtual screening studies as well as filters to develop/discovery new chemicals. An important challenge in the drug design field is the prediction of pharmacokinetic and toxicity properties, which can avoid failures in the clinical phases. Therefore, this review provides a critical point of view on the main MLT and shows their potential ability as a valuable tool in drug design.

Keywords: Machine learning, drug design, QSAR, medicinal chemistry.

1. INTRODUCTION

The drug discovery process involves the use of hybrid methodologies for discovery and design of new bioactive substances, which can be candidate to new drugs. One valuable strategy is the accurate prediction of biological activity of chemical substances from a set of atomic and molecular descriptors, and this is known as quantitative structure-activity relationships or QSAR studies. It is well accepted that physicochemical and structural properties of chemical compounds are helpful to understand many aspects of chemical and biological interactions in drug design projects and many other chemical processes. QSAR methodologies are used to develop statistical models relating chemical structure and biological activity, as well as they are useful in elucidating the mechanisms of the chemical–biological interaction in several biomolecules. One of the most important characteristics of QSAR models is their predictive power [1-10].

As described above, QSAR studies employ extra-thermodynamically derived and computational-based descriptors in order to correlate them with biological activity. From the identification of a relatively small set of related or unrelated molecules with known biological activities, many descriptors for these molecules can be determined. These standard molecular descriptors routinely used in QSAR analysis can be grouped generally into four major classes: electronic, steric, hydrophobic and topological. The next step in QSAR studies is developing a suitable statistical model, employing the descriptors obtained previously for a compound set, and the main application of this model is the activity prediction of new compounds and, consequently, to use such model for understanding the possible mechanisms of action for a particular drug [1-10].

The quality and success of a QSAR model depend strictly on the accuracy of input data, selection of appropriate descriptors and statistical tools, and most importantly validation of the developed model. It is important to say that a QSAR model is valid only for analogue molecular structures used to build the model. The validation process can be performed using some strategies: (a) internal validation (or cross-validation); (b) external validation (test compound set, not used in the model training); (c) data randomization (or Y-scrambling), and others [6, 7, 9, 10].

In the past decades, several statistical methods have broadened the arsenal of tools that can be applied to QSAR studies. A certain number of computational techniques have been found useful for the establishment of these relationships such as multiple linear regression (MLR) [11-14] and partial least squares (PLS) [15-20]. Recently, there is also a growing interest in the application of artificial neural networks (ANNs) and support vector machines (SVM) in the field of QSAR, as well as other molecular modeling approaches have been recognized as important tools in drug discovery [21-29]. The most common machine learning techniques (MLT) can be classified as supervised (such as Multilayer Perceptrons, Bayesian Neural Network, and Support Vector Machines), unsupervised (for example, Self-Organizing Maps) and hybrid models, such as Counter Propagation Neural Network (CPN), which comprises the advantages of supervised and unsupervised learning techniques. MLT are suited for QSAR studies because there is a set of compounds with known biological activities available for the training step. There are many potential parameters for each compound and the contribution of each parameter is not known a priori. Below, details of some machine learning techniques will be discussed in more details.

2. ARTIFICIAL NEURAL NETWORKS (ANNS)

Structure-activity relationships are essentially a process of pattern recognition and, historically, QSAR models have been developed using linear methods such as linear discriminant analysis (LDA) [30, 31]. However, several nonlinear QSAR methods have been proposed in recent years [32-34]. In QSAR methods based on regression analysis, it is necessary to previously assume an input–output relation (e.g. linear or quadratic function), but ANN methodologies do not require any prior model of how input and output are connected and have the unique ability to adapt to highly complex non-linear relations. So, ANN techniques are being increasingly used in QSAR studies and are frequently used for data analysis with increasing interest in chemistry and related fields of research since 1988 [35]. ANNs are able to generate models for complex input–output relationships based on learning from examples and, therefore, are useful in prediction, i.e. ANNs can be used as detectors, regressors and classifiers. In sum, ANNs are nonparametric techniques and can effectively address a broader class of complex real problems.

In medicinal chemistry, ANNs have been applied in compound classification, modeling of structure–activity relationships, identification of potential drug targets and localization of structural and

*Address correspondence to this author at the Departamento de Química e Física Molecular, Instituto de Química de São Carlos, Universidade de São Paulo, CP 780, 13560-970, São Carlos, SP, Brazil; Tel: +551633739975; Fax: +551633739975; E-mail: alberico@iqsc.usp.br

1875-533X/12 $58.00+.00

functional features of biopolymers [36]. ANNs techniques have been also used in the fields of robotics, pattern identification, psychology, physics, computer science, biology and many others [37-40]. In addition, ANNs have been applied to the modeling of several systems in a wide range of applications such as animal science [41], cancer imaging extraction and classification [42, 43], pharmacodynamic and pharmacokinetic modeling [44, 45] and mapping of dose–effect relationships on pharmacological response [46], to predict secondary structures [47] and transmembrane segments [48], simulation of C13 nuclear magnetic resonance spectra [49], prediction of drug resistance of HIV-1 protease ligands [50], prediction of toxicity of chemicals to aquatic species [51], and as well as predicting physicochemical properties from the perspective of pharmaceutical research [52].

ANNs can be defined as a set of computational models based on the neuronal interconnections of natural organisms, i.e. they are inspired in the information-processing pattern of biological nervous systems. Some advantages of ANNs include flexibility to discover more complex relationships in data than traditional statistical models, capability for extracting essential process information from data and the ability of generalization, i.e. they can correctly process information that only broadly resembles the original training data. The essential features of ANNs involve independence of statistical and modeling assumptions, non-linearity, fault tolerance and universality, which make them particularly suitable for extremely complex data [53]. Another very important characteristic of neural networks is that they are adaptive, i.e. can learn based on the data given for training.

ANN techniques have stimulus–response transfer functions that accept some inputs (stimulus) and yield some outputs (response). They are used to typically learn an input–output mapping over a set of examples. ANNs can be viewed as a set of parallel structures (called neurons) consisting of nonlinear processing elements interconnected by fixed or variable weights. The ANN learning process involves a training stage in which the model parameters (weight connections) are determined from a known data (training set), a validation step in which the model performance is estimated and a phase of prediction in which the property of novel compounds are predicted by the optimized network. With only a few exceptions, ANNs are able to generate nonlinear models that are capable of learning several complex interactions among the input variables in a system even when they are difficult to find and describe.

A crucial step in the ANN design is related to data preprocessing, which mainly consists in encoding the input information into an object representation so that this could be suitably processed by the ANN. Each compound in the data set has many molecular descriptors (experimental and/or theoretical) and one biological parameter, which is given as input for ANN analyses. An ideal encoding scheme should extract maximal information from the input data and satisfy the basic coding assumption that similar items are represented by close vectors [54]. Two problems related to a variety of parameters are redundancy in information and chance correlations. Since it is not possible to know a priori which molecular properties are the most relevant to the problem at hand, ANNs are often used in conjunction with techniques for feature selection [55]. Therefore, the choice of a significant set of descriptors that can avoid redundancy and chance correlation is a very important step.

The main structure of ANN architecture usually includes input, hidden and output layers. Input layers receive information directly from input data. On the other hand, the output layer sends information directly to the outside world, to a secondary computer process or to other devices such as a mechanical control system, indicating a required response. A variable number of hidden layers can be found between the input and output layers; in this case, the hidden layers are interconnected to each other. Many models of ANNs have been proposed and applied in medicinal chemistry problems such as self-organizing map, multilayer perceptrons, probabilistic neural networks, counter propagation networks, and others [56].

One example of the applicability of ANN is described by Douali *et al.* [57]. In that study, the authors have studied a set of HEPT derivatives acting as nonnucleoside reverse transcriptase inhibitors (NNRTIs) by using ANN techniques. From that study it is possible to demonstrate the nonlinearity of the relationship between the molecular calculated descriptors and the anti-HIV-1 activity presented by the compounds. ANN analysis yielded predicted activities in good agreement with the experimental values and the effect of each molecular descriptor on the anti-HIV-1 activity variation was clearly elucidated [57].

2.1. Self-Organizing Maps

Self-organizing map (SOM) is a technique that can be used for clustering, visualization and extrapolation, keeping the topology of the data set. A SOM network can be used to study data of high-dimensional spaces by projection into a two-dimensional plane, i.e. a bidimensional array of neurons (each neuron in the grid is also an output neuron). In this technique, the neurons are connected only with their closest neighbors in the array according to a prescribed topological scheme. If a particular neuron represents a given pattern, then its neighbors represent similar patterns. The result corresponds to input data projections, so that points that are adjacent in the high-dimensional space will also be adjacent in the Kohonen map [58]. An important feature of SOM is the self-organization of the artificial neurons, implemented by a competitive unsupervised learning process. This characteristic generates topological feature maps, which can be utilized to analyze the shape and surface properties of the objects (chemical substances) responsible for the biological activity. A SOM process involves the following steps: (i) all neurons in the active layer obtain the same multidimensional input, and at the same time; (ii) a training pattern is presented to the network, and the winning neuron for representing such pattern is found; (iii) the winner neuron has its weights updated using the current learning rate, while the learning rate for the neighbors is scaled down proportional to its distance to the winner. Consequently, the knowledge of that pattern will be localized in the area of the winner neuron; (iv) each training step involves one pass through the data and the training is stopped when changes to the network's weights become insignificant; (v) a new pattern is classified according to the cluster associated with the corresponding winner neuron in the grid [58]. The main advantage of the SOM, in comparison to other projection methods, is that the algorithm is very simple, straightforward to implement and fast to compute. Fig. (**1**) exemplifies a Kohonen network.

The main applications of SOM are: the two-dimensional maps can be taken as a representation, an encoding, of the higher-dimensional information; samples mapped in the same or closely adjacent neurons can be considered as similar; points that form a group in such map can be taken as a class of samples having certain features in common [59]. This technique was successfully applied to chemical analysis problems using atomic absorption spectrometry [60] and to obtain the near-infrared spectral calibration of complex beverage samples [61]. Besides these applications, the SOM potentialities have been shown in the modeling toxicity [62] and design of novel trypanocidal quinone compounds [25]. Other use of SOM involves the separation of compounds in training and test. Several other studies have employed this technique with the same purpose [63-67].

2.2. Multilayer Perceptrons

Multilayer perceptrons (MLP) are considered feed-forward neural networks since all of the data information flows in only one direction, from the input to output units, and are probably the most common architecture used in supervised machine learning studies. MLP can be viewed as a universal approximator; this fact explains its success in QSAR analyses. MLP is very fast and easy to use, but its training is considered a very delicate task since it is slow and there is no guarantee that the achieved minimum is global [68]. A challenge in training the MLP is the choice of the appropriate network architecture, i.e. number of hidden layers and number of neurons in each layer. The speed of the training/learning is strongly affected by the learning rates and the momentum parameters involved in the training procedure for the adjustment of connections among the consecutive network layers [69].

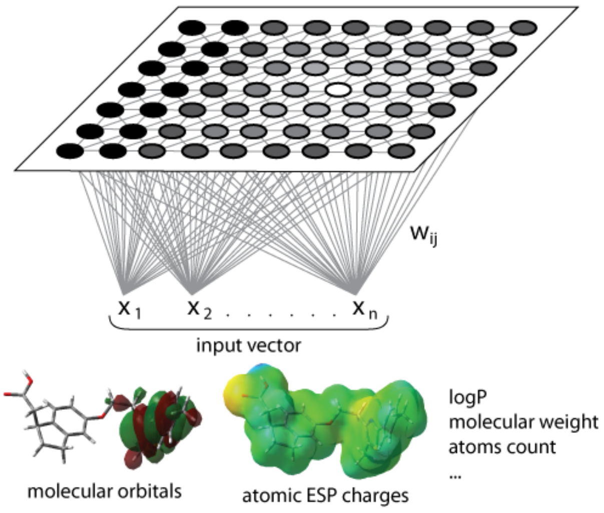

Fig. (1). Implementation of a generic SOM map.

This neural model is composed of multiple processing units (neurons) arranged in consecutive highly interconnected layers, in a way that the first layer consists of the input variables, and the last one consists of the output variables representing the estimated class. Intermediate layers, named as hidden layer, receive the entire input pattern that is modified by the passage through the weighted connections. A training step means a search process for the optimized set of weight values, which can minimize the squared error between the network's outputs and the desired outputs given in the training data set [68]. As an example, Fig. (**2**) shows how the input data, hidden neurons and output data can be modeled by a MLP network.

A MLP network can be trained using different algorithms such as backpropagation, conjugate gradient descent, quasi-Newton, Levenberg-Marquardt, quickpropagation, delta-bar-delta, etc. [70]. Backpropagation learning algorithm, which is the most widely used for MLP training, starts by calculating the error in the output layer and go backward through the network until the input layer. This training algorithm is used to adjust the network's weights and thresholds in order to minimize the error commited by the model. Because a target value is compared to the output value, the learning is called supervised [69, 70]. Although the use of backpropagation is common in neural network applications, it is quite limiting, as this procedure provides a guaranteed convergence, but only to a locally optimal solution. Even if the network's topology provides sufficient complexity to solve the given pattern recognition task completely, the backpropagation method may be incapable of discovering an appropriate set of weights to accomplish the task. When this occurs, an alternative solution is adding degrees of freedom to the network by increasing the number of nodes and/or connections. The problem of local convergence associated to the back propagation algorithm indicate the desirability of training with stochastic optimization methods, such as simulated evolution, which can provide convergence to globally optimal solutions [69, 70].

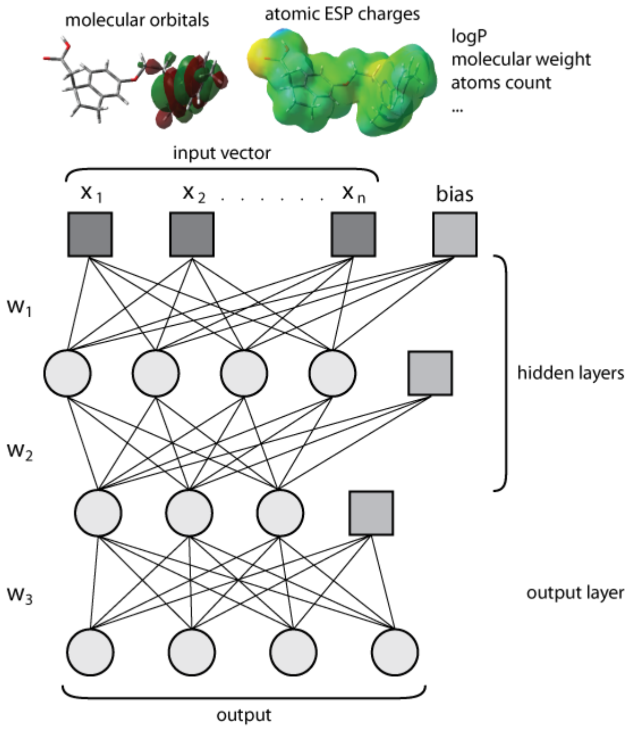

Fig. (2). An example of MLP architecture. The input vector presents the molecular orbitals and the atomic charges. The "bias" is an independent value which determines the bound of classification in a neural network.

An example that can be cited is the study of 50 cannabinoid compounds classified experimentally into psychoactives and psychoinactives [24]. In that study, the authors employed SOM and MLP techniques with the aim to predict the class of studied compounds (actives and inactives) and the obtained results indicated 96% of correct information. To plot the results and improve the physical interpretation, the Kohonen network was employed. Both MLP and SOM techniques showed reliable results and allowed good interpretation and visualization of the outcomes [24].

Many studies have used MLP for the modeling of biological/pharmacological systems, such as a study that applied MLP and Kohonen (SOM) techniques to model the relationship between quantum chemical and molecular descriptors and the trypanocidal activity of a compound set. Both obtained nonlinear models predicted the biological activity with good agreement to experimental data and the authors concluded that the employed descriptors were important to describe the main interactions between the substances and the biological target [25].

The first QSAR model for substances with affinity to vesicular monoamine transporter-2 (VMAT-2) was constructed by Zheng *et al.* [71] that used the MLP methodology. The interaction with this receptor can be a target to decrease the neurotoxicity of the methamphetamine (METH) usage; therefore, the studied ligands can be used for the METH dependence treatment. The authors compared MLP backpropagation neural network with PLS model to predict biological activity. The obtained properties (atomic distribution at the molecule, molecule size and steric descriptors, such as symmetry and shape indexes) were able to propose some protein-ligand interactions and guide the development of new VMAT-2 candidates. The statistical analysis from ANN analyses displayed until 8% more accurate than PLS model and shows a regression coefficient for the prediction set (r^2) of 0.93 [71].

2.3. Counter Propagation Neural Networks

The typical counter propagation neural network (CPNN) is a feedforward hybrid model which consists of three layers: the input layer, a hidden competitive layer represented by a SOM, and the output layer named Grossberg layer, fully connected to the competitive layer, providing a means of combining unsupervised and supervised learning features [72, 73]. According to Tso and Mather [73], the structure of a CPNN is similar to a multilayer perceptron. However, the hidden layer, and the training rules are quite different.

The training process of CPNN involves two main steps: (i) an unsupervised learning scheme is used by the hidden SOM layer for clustering the input data in distinct groups, and (ii) the output of the hidden layer feeds the Grossberg layer, which uses a supervised learning algorithm to adjust the weights between the Grossberg layer and the hidden one. This supervised stage aims at minimizing the classification error over the training data set.

This technique was successfully applied in chemistry to construct a model to predict a inhibitor into the active site of human thrombin [74], as well as to propose a novel 3-hydroxypyridine-4-one with improved antibacterial activity against *Staphylococcus aureus* [75] and to study some fluoroquinolones and their activity values from *in vitro* tests for inhibitory activity against tuberculosis [76].

2.4. Bayesian Neural Networks

A neural network model that incorporates Bayesian techniques in its learning process is known as Bayesian Neural Network (BNNs) and it has been successfully used in several chemical studies. According to Bishop [77], this approach has some suitable advantages over the traditional ANNs as better performance on error minimization, the ability for determining the best input set for a neural network and the possibility on using only training sets for comparing different models, which is an important issue, mainly when there is a small number of samples in a data set that can difficult the learning process of an ANN.

In the Bayesian framework, the weights of the network are set according to the priori distribution over its values, which can be determined by the Bayes' theorem [78]. These neural models have shown to perform efficient feature selection and generated robust QSAR models without committing overfitting, or risk of chance correlations [78]. This can be accomplished by including an error term that regularizes the weights by penalizing large magnitudes. Another important characteristic of this model is that the Bayesian neural networks have the potential to give results which are relatively independent of the network architecture, as well as the Bayesian method estimates the number of effective parameters. More details about this technique can be found in [79-81].

The main process used in the BNN studies involves the following steps: (i) the Bayesian training produces a posterior distribution over the network weights; (ii) when the inputs of the network are set to the values for some new pattern, the posterior distribution over network weights will give rise to a corresponding predictive distribution over the outputs; (iii) if a single-valued prediction is needed, then the mean of the predictive distribution may be considered as the network output, but the full predictive distribution also provides information on how uncertain this prediction is; (iv) choosing the appropriate network architecture (number of hidden layers and number of hidden units in each layer) and adapting to the relevant characteristics of the data depends on using the right prior distribution [82].

One of the most common approaches for such modeling is based on the Levenberg–Marquart algorithm with regularization, which has the aim to find the weights and hyperparameters that are the most probable, and then to approximate the distribution over weights [79, 80, 82]. Special automatic relevance detection methods were developed, allowing all input parameters to be used in the neural net, with the Bayesian inference eliminating those containing no information or redundant information [79, 80, 82]. The Bayesian training allows the use of all training data for building the model. Nevertheless, proof of the model generalization power may be confirmed only through external validation. This methodology requires at least twice as many training patterns as weights in the network, and the training is usually slow [82].

An interesting application of the BNN technique was performed by Caballero *et al.* [83] for a set of cyclin-dependent kinase (CDK) inhibitors, and the BNN model was the best one to perform external predictions. Ajay *et al.* [84] employed the BNN approach to classify molecules as drugs and non-drugs using the drug-like concept. To construct their models [84], the authors calculated various descriptors such as molecular weight, number of hydrogen bond donors and acceptors, number of rotatable bonds, aromatic density, log P and the kappa index. One of the constructed models showed the ability of generalizing about 80% of large libraries. Wang *et al.* [85] modeled the CYP3A4-mediated metabolism using the Km values (Michaelis-Menten constant) and electrotopological descriptors by means of the BNN technique. The authors performed outlier tests, comparing with the experimental data, and the obtained model had internal and external robustness [85]. Burden *et al.* [86] employed the BNN method and added the automatic relevance determination (ARD) method to improve the robustness of QSAR models. The ARD method allows the neural networks to classify the weight of each input and to estimate the "importance" of each input and turn off the less relevant inputs. The authors constructed more predictive models using BNN and BNN coupled to ARD than those using standard back-propagation neural networks and MLR models [86].

3. SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) is a supervised machine learning approach based on a statistical learning theory, called Vapnik-Chervonenkis (VC) theory [87]. Actually, this theory aims to propose computational learning techniques that maximize the generalization ability of classifiers [87].

To understand the fundamentals of SVM and its main benefits is important to explain the statistical concepts related to this technique. The first concept is related to the incorporation of the principle of structural risk minimization in its formulation, which means to consider, during the training process, the risk of structural models instead of the empirical risk. This feature yields to the SVM model a good generalization performance, allowing the correct classification of data that are not present in the training set [88].

Moreover, the SVM are also concerned about reducing the complexity of mathematical function used by the classifier. For this, it is necessary to control the VC dimension, which is a scalar index that measures the complexity of a function according to the maximum number of samples that are correctly classified.

The SVM approach consists in constructing a separating hyperplane that maximizes the distance between the classifier and the nearest sample of each class, which lies on the bounds of the margin of such hyperplane, known as margin of separation. The hyperplanes that define such margin are called support hyperplanes, and the data points that lie on these hyperplanes are called support vectors. More details on SVM concepts can be found in other studies in literature [89, 90]. Fig. (**3**) displays a SVM hyperplane.

4. COMPARISON OF METHODS

All machine learning techniques have advantages and disadvantages. As the research in drug design depends on a series of factors

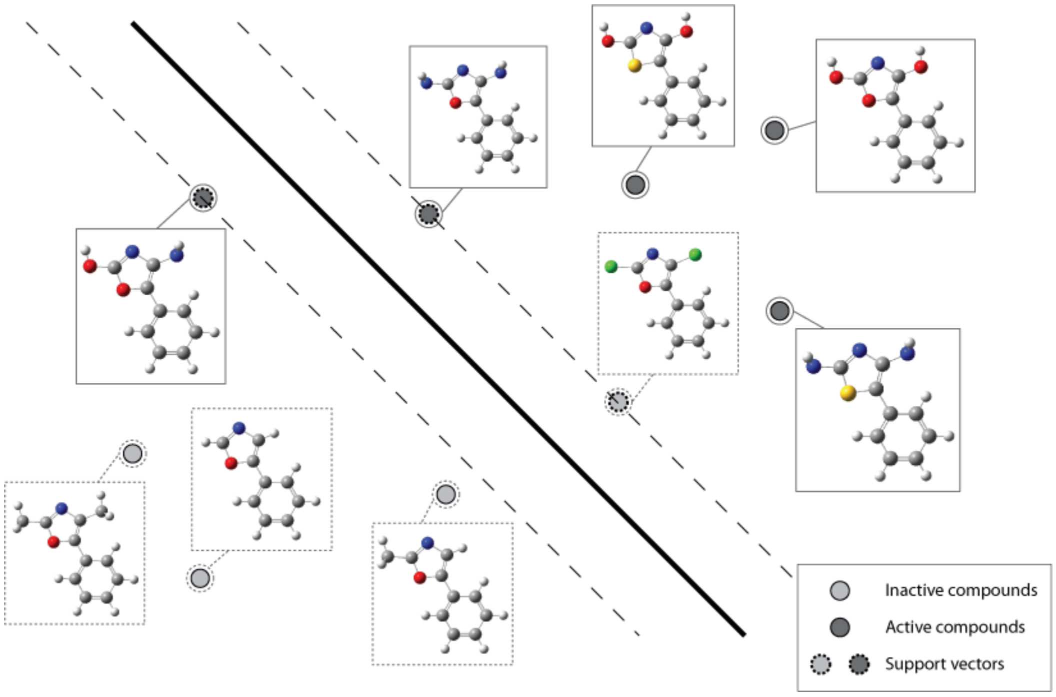

Fig. (3). Maximal hyperplane separation, with dashed lines as support hyperplanes, indicating the margin of separation between two classes.

such as biological data and a large number of molecular descriptors, each method could be efficient for specific problems. Therefore, it can be found in the literature several comparative studies of methodologies showing some trends in the use of techniques based on machine learning for QSAR applications.

Louis *et al.* [91] compared the modeling power of some techniques as MLR, ANN (three-layer MLP network using 10 fold cross-validation) and support vector machines (SVM) to evaluate the quantitative structure-property relationships (QSPR) regarding the intrinsic solubility of known generic drugs. The aqueous solubility is an important property of drug development due to its impact on pharmacokinetics, physical aspects and formulation of drugs [92-96]. The data set was composed of 74 generic drugs with intrinsic solubility (solubility of unionized species) measured at same experimental conditions. The best obtained ANN model contains 2 hidden neurons and was the best model to predict the test set (r^2 = 0.907 and r^2_{test} = 0.770); the SVM technique was the most accurate model to the training set (r^2 = 0.913 and r^2_{test} = 0.731).

Qin *et al.* [97] constructed and compared ANN and MLR models of chloroethylnitrosoureas (CENUs), which are potent alkylanting agents and employed against some kinds of tumors (as leukemias, melanomas, encephalomas and some solid tumors), since they can modify some nucleosides (possible source of citotoxicity). All selected descriptors described the interactions related to the DNA alkylation. From the comparison between the ANN and MLR models, the best neural network model displayed a higher accuracy than the multiple linear regression model (r^2 = 0.983 and r^2 = 0.506, respectively).

Another example of a QSAR study was described by Darnag *et al.* [98], which studied a set of tetrahydroimidazo[4,5,1-jk][1,4]benzodiazepinone (TIBO) derivatives, with anti-HIV activity, employing SVM, ANN and MLR techniques. From the results, the best QSAR model was obtained by using the SVM methodology and it was 6% more accurate than ANN and 16% more accurate than the MLR technique.

Sorich *et al.* [99] have compared PLS, BRANN and SVM models to predict the metabolism of some chemicals by human UDP-glucuronosyltransferase isoforms. The authors concluded that SVM has the highest accuracy among all the considered models, followed by BRANN and PLS, respectively.

Fatemi *et al.* [100] constructed some ANN, SVM and MLR models to predict the aqueous solubility of 145 drug-like compounds. The obtained statistical parameters revealed that the ANN model was superior to other methods, having r^2 equals to 0.816.

A QSAR study of melanocortin-4 receptor (MC4R) was constructed by Pourbasheer *et al.* [101] employing MLR and SVM techniques to construct the linear and nonlinear models. The SVM model was the most reliable, showing a regression coefficient equals to 0.908 for the MC4R binding affinity prediction.

MLT have advantages and disadvantages in their implementations. Kotsiantis [102] considers that, in most of cases, SVM and supervised neural networks tend to give better results when dealing with multi-dimensions and continuous features. These methods are also adequate when multicollinearity is present and there is a nonlinear relationship between the input and output features. Some disadvantages of supervised MLT include a large quantity of samples required to achieve the maximum prediction accuracy, the presence of irrelevant features that make the neural network training be inefficient. The use of ANN and SVM techniques in drug design studies has a significant limitation with respect to the interpretation/understanding of the extracted features in physicochemical terms that describe the interaction between bioactive substances and the biological target [103].

The most widely used MLP suffers with some limitations such as architecture and parameter setting, slow convergence of the back-propagation algorithm (with the risk to get stuck in a local minimum), poor ability of generalization, and lack of robustness due to random initialization of the weights [104-106]. The main disadvantages of SVM include the high complexity of the model and the long computing time if a significant prediction power is required [107].

The advantages of SOM and CPNN are related to their non-deterministic characteristics, as well as to the robustness of these techniques to outliers. Some limitations of these methods are related to the complexity of their training processes (due to parameters like number of neurons, network architecture, initial learning rate, neighborhood function) and long time for training, especially if large data sets are processed [108, 109]. Hussain [110] and Winkler [111] affirm that a few number of biological data (in particular, *in vivo* measurements) can also contribute to the poor performance (prediction accuracy) of MLT.

Several techniques of machine learning are very useful in drug design and Fig. (**4**) illustrates the current trends obtained by observing literature data from last decade. It is important to note that various machine learning techniques address the problem of high dimensionality of data in QSAR studies and Genetic Algorithm (GA) is one of the most applied techniques to solve this problem. This approach is based on natural selection, mutation, evolution and genetic crossover [112]. Another very used methodology in medicinal chemistry studies is Principal Component Analysis (PCA), which is a statistic method that can be used for reducing the dimensionality of multivariate data, preserving the relevant information in the original data set [113].

5. TRENDS OF MACHINE LEARNING TECHNIQUES IN DRUG DESIGN

There are several cases of discovering of a new lead compound from large chemical libraries employing distinct strategies. Each technique has its advantages, disadvantages and limitations, but each one can be employed to specific systems or can be combined with other techniques to improve its efficacy [114, 115].

QSAR methods have been used to predict the biological activity of unknown compounds and to optimize and/or improve the activity of known classes of compounds from molecular modifications [116-119]. Furthermore, QSAR models can be used as a powerful filter on virtual screening studies to discover new drug candidates [120, 121]. Virtual screening is considered an important and efficient tool in drug design research [122] and can be combined with machine learning methods.

A successful example of technique combination is reported by Hu *et al.* [123]. In that study, the authors identified 49 aldose reductase inhibitors (ARI) combining QSAR with ANN method, considering a structure-based virtual screening. The discovered hits showed binding energies similar to experimentally known ARI compounds.

Afantitis *et al.* [124], Noeske *et al.* [125] and Schneider *et al.* [126] have employed self- organizing maps to perform a ligand-based virtual screening on large chemical databases and the authors identified new hits with success. Therefore, artificial neural networks have been used successfully in drug design [127, 128]. Also, classification and regression methods based on neural networks and support vector machines can be used as valuable tools to develop new drugs [129, 130].

The main challenge in drug discovery is the prediction of pharmacokinetic properties [131-134]. There are several cases of developing drugs that cannot be used because of their side effects, toxicity, pharmacokinetics and formulation [135-139]. Pharmacokinetics involves four stages that are known as ADME (absorption, distribution, metabolism and excretion). Absorption, distribution and excretion stages depend on administration and formulation of the drug, target's location and solubility of the drug molecule. Therefore, several barriers must be considered on drug optimization stage. The metabolism process is more complex than the others because it can occur several side reactions (such as oxidation, reduction, hydrolysis, conjugation, addition of polar groups, dealkylation of tertiary and secondary amines and others) mainly in liver and blood plasma [138, 140]. In addition, the toxicity of drug candidates is a limiting factor to several compounds. So, there are some studies that employ neural networks in order to model the ADME and toxic properties of drugs and chemicals [141-148]. Nevertheless, it is interesting to notice the significant advantages for the use of ANN in a wide range of medicinal chemistry studies.

CONCLUSIONS

The use of machine learning techniques (MLT) has increased in the last decades and this indicates a tendency in the chemistry field. In fact, these techniques are useful tools in drug design studies nowadays. Support vector machines and artificial neural networks are methodologies used in various medicinal chemistry problems. In addition, several research groups have investigated the efficacy

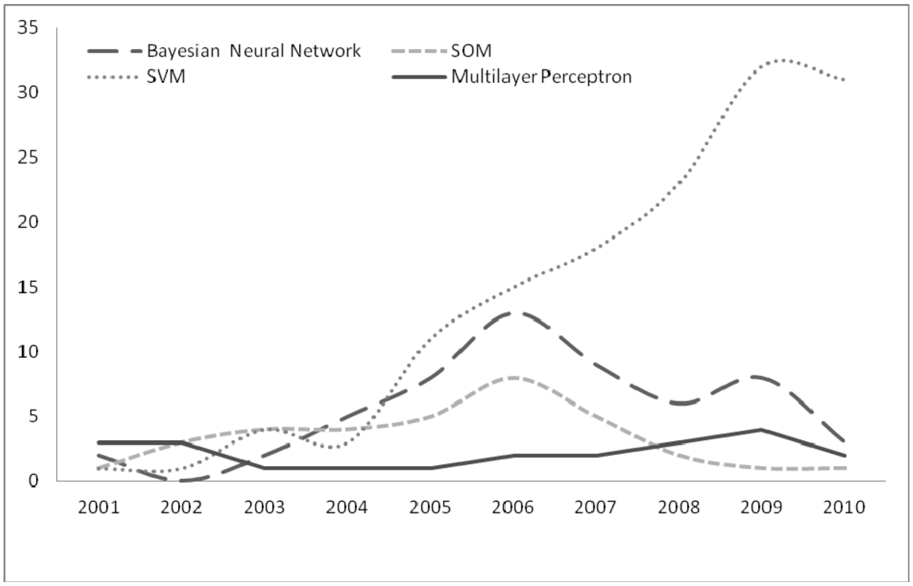

Fig. (4). Results (last decade) per year obtained from the search using QSAR keyword plus each one of word: "Bayesian Network", "SOM", "SVM" and "Multilayer perceptron".

of these techniques and their applicability. MLT have proved to be more effective than traditional clustering and multivariate analysis methods to solve actual problems such as prediction of biological activities, construction of QSAR/QSPR models, virtual screening and the prediction of pharmacokinetic properties. Therefore, there are many opportunities for the application of MLT in information management and analysis in the medicinal chemistry.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

The authors would like to thank the Brazilian agencies CNPq, FAPESP and CAPES for providing the financial support.

REFERENCES

[1] De Benedetti, P.G.; Fanelli, F. Computational quantum chemistry and adaptive ligand modeling in mechanistic QSAR. *DDT,* **2010,** *15*(19–20), 859–866.

[2] Castillo-Garit, J.A.; Abad, C.; Rodríguez-Borges, J.E.; Marrero-Ponce, Y.; Torrens, F. A review of QSAR studies to discover new drug-like compounds actives against leishmaniasis and trypanosomiasis. *Curr. Top. Med. Chem.,* **2012,** *12*(8), 852-65.

[3] Sharma, O.P.; Saini, N.K.; Gupta, V.; Sachdeva, K.; Arya, H. Evolutionary History of QSAR: A Review. *J. Natur. Cons.,* **2011,** *1*(4), 266-272.

[4] Oprea, T.I. On the information content of 2D and 3D descriptors for QSAR. *J. Braz. Chem. Soc.,* **2002,** *13*(6), 811-815.

[5] Tavares, L.C. QSAR: a abordagem de Hansch. *Quim. Nova,* **2004,** *27*(4), 631-639.

[6] Montanari, M.L.C.; Montanari, C.A.; Gaudio, A.C. Validação lateral em relações quantitativas entre estrutura e atividade farmacológica. *Quim. Nova,* **2002,** *25*(2), 231-240.

[7] Gaudio, A.C.; Zandonade, E. Proposição, validação e análise dos modelos que correlacionam estrutura química e atividade biológica. *Quim. Nova,* **2001,** *24*(5), 658-671.

[8] Arroio, A.; Honório, K.M.; da Silva, A.B.F. da. Propriedades químico-quânticas empregadas em estudos das relações estrutura-atividade. *Quim. Nova,* **2010,** *33*(3), 694-699.

[9] Kiralj, R.; Ferreira, M.M.C. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *J. Braz. Chem. Soc.,* **2009,** *20*(4), 770-787.

[10] Ferreira, M.M.C. Multivariate QSAR. *J. Braz. Chem. Soc.,* **2002,** **13**(6), 742-753.

[11] Goudarzi, N.; Goodarzi, M.; Chen, T. QSAR prediction of HIV inhibition activity of styrylquinoline derivatives by genetic algorithm coupled with multiple linear regressions. *Med. Chem. Res.,* **2012,** *21*(4), 437-443.

[12] Xu, J.; Wang, L.; Wang, L.; Shen, X.; Xu, W. QSPR Study of Setschenow Constants of Organic Compounds Using MLR, ANN, and SVM Analyses. *J. Comp. Chem.,* **2011,** *32*(15), 3241-3252.

[13] Alves, C.N.; Pinheiro, J.C.; Camargo, A.J.; Ferreira, M.M.C.; Ramero, R.A.; da Silva, A.B.F. A multiple linear regression and partial least squares study of flavonoid compounds with anti-HIV activity. *J. Mol. Struct. THEOCHEM,* **2001,** *541*(1-3), 81-88.

[14] Yu, Y.; Su, R.; Wang, L.; Qi, W.; He, Z. Comparative QSAR modeling of antitumor activity of ARC-111 analogues using stepwise MLR, PLS, and ANN techniques. *Med. Chem. Res.,* **2010,** *19*(9), 1233-1244.

[15] Nandi, S.; Bagchi, M.C. Activity Prediction of Some Nontested Anticancer Compounds Using GA-Based PLS Regression Models. *Chem. Biol. Drug Des.,* **2011,** *78*(4), 587-595.

[16] Noolvi, M.N.; Patel, H.M.; Bhardwaj V.A. Comparative QSAR Analysis of Quinazoline Analogues as Tyrosine Kinase (erbB-2) Inhibitors. *Med. Chem.,* **2011**, *7*(3), 200-212.

[17] Weber, K.C.; Honório, K.M.; Bruni, A.T.; Andricopulo, A.D.; da Silva, A.B.F. A partial least squares regression study with antioxidant flavonoid compounds. *Struct. Chem.,* **2006,** *17*(3), 307-313.

[18] Garcia, T.S.; Honorio, K.M. Two-Dimensional Quantitative Structure-Activity Relationship Studies on Bioactive Ligands of Peroxisome Proliferator-Activated Receptor delta. *J. Braz. Chem. Soc.,* **2011**, *22*(1), 65-72.

[19] Weber, K.C.; Honório, K.M.; Andricopulo, A.D.; da Silva, A.B.F. Pharmacophore-based 3D QSAR studies on a series of high affinity 5-HT1A receptor ligands. *Eur. J. Med. Chem.,* **2010,** *45*(4), 1508-1514.

[20] Maltarollo, V.G.; Homem-de-Mello P.; Honório K.M. Role of physicochemical properties in the activation of peroxisome proliferator-activated receptor delta. *J. Mol. Model.,* **2011,** *17*(10), 2549-2558.

[21] Witten, I.H.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann. **2005.**

[22] Moss, G.P.; Shah, A.J.; Adams, R.G.; Davey, N.; Wilkinson, S.C.; Pugh, W.J.; Sun, Y. The application of discriminant analysis and Machine Learning methods as tools to identify and classify compounds with potential as transdermal enhancers. *Eur. J. Pharm. Sci.,* **2012,** *45*(1-2), 116-127.

[23] Yang, X.G.; Lv, W.; Chen, Y.Z.; Xue, Y. In Silico Prediction and Screening of gamma-Secretase Inhibitors by Molecular Descriptors and Machine Learning Methods. *J. Comp. Chem.,* **2010,** *31*(6), 1249-1258.

[24] Honório, K.M.; de Lima, E.F.; Quiles, M.G.; Romero, R.A.F.; Molfetta, F.A.; da Silva, A.B.F. Artificial Neural Networks and the Study of the Psychoactivity of Cannabinoid Compounds. *Chem. Biol. Drug Des.,* **2010,** *75*(6), 632–640.

[25] Molfetta, F.A.; Angelotti, W.F.D.; Romero, R.A.F.; Montanari, C.A.; da Silva, A.B.F. A neural networks study of quinone compounds with trypanocidal activity. *J. Mol. Model.,* **2008,** *14*(10), 975-985.

[26] Hameed, A.J.; Ibrahim, M.; ElHaes, H. Computational notes on structural, electronic and QSAR properties of Fulleropyrrolidine-1-carbodithioic acid 2; 3 and 4-substituted-benzyl esters. *J. Mol. Struct. THEOCHEM*, **2007,** *809*(1-3), 131-136.

[27] Essa, A.H.; Ibrahem, M.; Hameed, A.J.; Al-Masoudi, N.A. Theoretical investigation of 3'-subtituted-2'-3'-dideoxythymidines related to AZT. QSAR infrared and substituent electronic effect studies. *ARKIVOC*, **2008,** *xiii*, 255-265.

[28] Ibrahim, M.; Saleh, N.A.; Hameed, A.J.; Elshemey, W.M.; Elsayed, A.A. Structural and Electronic Properties of new Fullerene Derivatives and their Possible Application as HIV-1 Protease Inhibitors. *Spectrochim. Acta A,* **2010,** *75*(2), 702-709.

[29] Ibrahim, M.; Saleh, N.A.; Elshemey, W.M.; Elsayed, A.A. Fullerene Derivative as anti-HIV Protease Inhibitor: Molecular Modeling and QSAR Approaches. *Mini Rev. Med. Chem.,* **2012,** *12*(6), 447-451.

[30] Garcia, I.; Fall, Y.; Gomez, G.; Gonzalez-Diaz, H. First computational chemistry multi-target model for anti-Alzheimer, anti-parasitic, anti-fungi, and anti-bacterial activity of GSK-3 inhibitors *in vitro*, *in vivo*, and in different cellular lines. *Mol. Divers.,* **2011,** *15*(2), 561-567.

[31] Casanola-Martin, G.M.; Marrero-Ponce, Y.; Khan, M.T.H.; Khan, S.B.; Torrens, F.; Perez-Jimenez, F.; Rescigno, A.; Abad, C. Bond-Based 2D Quadratic Fingerprints in QSAR Studies: Virtual and *In vitro* Tyrosinase Inhibitory Activity Elucidation. *Chem. Biol. Drug Des.,* **2010,** *76*(6), 538-545.

[32] Goodarzi, M.; Freitas, M.P.; Vander Heyden, Y. Linear and nonlinear quantitative structure-activity relationship modeling of the HIV-1 reverse transcriptase inhibiting activities of thiocarbamates. *Anal. Chim. Acta,* **2011,** *705*(1-2), 166-173.

[33] Qin, Y.; Deng, H.; Yan, H.; Zhong, R. An accurate nonlinear QSAR model for the antitumor activities of chloroethylnitrosoureas using neural networks. *J. Mol. Graph. Model.,* **2011,** *29*(6), 826-833.

[34] Fu, G.H.; Cao, D.S.; Xu, Q.S.; Li, H.D.; Liang, Y.Z. Combination of kernel PCA and linear support vector machine for modeling a nonlinear relationship between bioactivity and molecular descriptors. *J. Chemometr.,* **2011,** *25*(2), 92-99.

[35] Hoskins, J.C.; Himmelbau, D.M. Artificial Neural Network Models of Knowledge Representation in Chemical Engineering. *Comput. Chem. Eng.,* **1988,** *12*(9-10), 881-890.

[36] Zhou, P.; Tian, F.; Chen, X.; Shang, Z. Modeling and Prediction of Binding Affinities Between the Human Amphiphysin SH3 Domain and Its Peptide Ligands Using Genetic Algorithm-Gaussian Processes. *Biopolymers*, **2008,** *90*(6), 792-802.

[37] Soyguder, S. Intelligent control based on wavelet decomposition and neural network for predicting of human trajectories with a novel vision-based robotic. *Expert Syst. Appl.,* **2011,** *38*(11), 13994-14000.

[38] Aitkenhead, M.J.; McDonald, A.J.S. The state of play in machine/environment interactions. *Artif. Intell. Rev.,* **2006,** *25*(3), 247-276.

[39] Fogel G.B. Computational intelligence approaches for pattern discovery in biological systems. *Brief Bioinform.,* **2008,** *9*(4), 307-316.

[40] Perlovsky, L.I. Toward physics of the mind: Concepts, emotions, consciousness, and symbols. *Phys. Life. Rev.,* **2006,** *3*(1), 23-55.

[41] Fernandez, C.; Soria, E.; Martin, J.D.; Serrano, A.J. Neural networks for animal science applications: Two case studies. *Expert Syst. Appl.,* **2006,** *31*(2), 444-450.

[42] Cammann, H.; Jung, K.; Meyer, H.A.; Stephan, C. Avoiding Pitfalls in Applying Prediction Models, As Illustrated by the Example of Prostate Cancer Diagnosis. *Clin. Chem.,* **2011,** *57*(11), 1490-1498.

[43] Seker, H.; Odetayo, M.O.; Petrovic, D.; Naguib, R.N.G.; Bartoli, C.; Alasio, L.; Lakshmi, M S.; Sherbet, G.V. Assessment of nodal involvement and survival analysis in breast cancer patients using image cytometric data: Statistical, neural network and fuzzy approaches. *Anticancer Res.,* **2002,** *22*(1A), 433–438.

[44] Li H.; Yap, C. W.; Xue, Y.; Li, Z.R.; Ung, C.Y.; Han, L.Y.; Chen Y.Z. Statistical learning approach for predicting specific pharmacodynamic, pharmacokinetic, or toxicological properties of pharmaceutical agents. *Drug Develop. Res.,* **2005,** *66*(4), 245–259.

[45] Mager, D.E.; Shirey, J.D.; Cox, D.; Fitzgerald, D.J.; Abernethy, D.R. Mapping the dose-effect relationship of orbofiban from sparse data with an artificial neural network. *J. Pharm. Sci.,* **2005,** *94*(11), 2475–2486.

[46] Schneider, G.; Coassolo, P.; Lavé, T. Combining *in vitro* and *in vivo* pharmacokinetic data for prediction of hepatic drug clearance in humans by artificial neural networks and multivariate statistical techniques. *J. Med. Chem.,*

1999, *42*(25), 5072–5076.

[47] Guo, J.; Rao, N. Predicting Protein Folding Rate From Amino Acid Sequence. *J. Bioinform. Comp. Biol.,* **2011,** *9*(1), 1-13.

[48] He, J.Y.; Hu, H.J.; Harrison, R.; Tai, P.C.; Pan, Y. Transmembrane segments prediction and understanding using support vector machine and decision tree. *Expert Syst. Appl.,* **2006,** *30*(1), 64-72.

[49] Jalali-Heravi, M.; Shahbazikhah, P.; Zekavat, B.; Ardejani, M.S. Principal component analysis-ranking as method for the simulation of 13 C nuclear a variable selection ma netic resonance spectra of xanthones using artificial neural networks. *QSAR Comb. Sci.,* **2007,** *26*(6), 764-772.

[50] Reddy, A.S.; Kumar, S.; Garg, R. Hybrid-genetic algorithm based descriptor optimization and QSAR models for predicting the biological activity of Tipranavir analogs for HIV protease inhibition. *J. Mol. Graph. Model.,* **2010,** *28*(8), 852-862.

[51] Carafa, R.; Faggiano, L.; Real, M.; Munne, A.; Ginebreda, A.; Guasch, H.; Flo, M.; Tirapu, L.; von der Ohe, P.C. Water toxicity assessment and spatial pollution patterns identification in a Mediterranean River Basin District. Tools for water management and risk analysis. *Sci. Tot. Environ.,* **2011,** *409*(20), 4269-4279.

[52] Sprous, D.G.; Palmer, R.K.; Swanson, J.T.; Lawless, M. QSAR in the Pharmaceutical Research Setting: QSAR Models for Broad, Large Problems. *Curr. Top. Med. Chem.,* **2010,** *10*(6), 619-637.

[53] Milac, A.L.; Avram, S.; Petrescu, A.J. Evaluation of a neural networks QSAR method based on ligand representation using substituent descriptors Application to HIV-1 protease inhibitors. *J. Graph. Mol. Model.,* **2006,** *25*(1), 37–45.

[54] Wu, C.H. ; McLarty, J.W. *Neural Networks and Genome Informatics*, Elsevier, **2000**.

[55] Ghosh, P.; Bagchi, M.C. QSAR Modeling for Quinoxaline Derivatives using Genetic Algorithm and Simulated Annealing Based Feature Selection. *Curr. Med. Chem.,* **2009,** *16*(30), 4032-4048.

[56] Caballero, J.; Fernandez, M. Artificial Neural Networks from MATLAB (R) in Medicinal Chemistry. Bayesian-Regularized Genetic Neural Networks (BRGNN): Application to the Prediction of the Antagonistic Activity Against Human Platelet Thrombin Receptor (PAR-1). *Curr. Top. Med. Chem.,* **2008,** *8*(18), 1580-1605.

[57] Douali, L.; Villemin, D.; Cherqaoui, D. Neural networks: Accurate nonlinear QSAR model for HEPT derivatives. *J. Chem. Inf. Comp. Sci.,* **2003,** *43*(4), 1200-1207.

[58] Schneider, P.; Tanrikulu, Y.; Schneider, G. Self-Organizing Maps in Drug Discovery: Compound Library Design, Scaffold-Hopping, Repurposing. *Curr. Med. Chem.,* **2009,** *16*(3), 258-266.

[59] Kohonen, T. *Self-Organizing Maps.* 3rd Ed. Springer. **2000.**

[60] Balbinot, L.; Smichowski, P.; Farias, S.; Arruda, M.A.Z.; Vodopivez, C.; Poppi, R.J. Classification of Antarctic algae by applying Kohonen neural network with 14 elements determined by inductively coupled plasma optical emission spectrometry. *Spectrochim. Acta B,* **2005,** *60*(5), 725-730.

[61] Tan, C.; Qin, X.; Li, M. An ensemble method based on a self-organizing map for near-infrared spectral calibration of complex beverage samples. *Anal. Bioanal. Chem.,* **2008,** *392*(3), 515-521.

[62] Vracko, M. Kohonen Artificial Neural Network and Counter Propagation Neural Network in Molecular Structure-Toxicity Studies. *Curr. Comput-Aid. Drug.,* **2005,** *1*(1), 73-78.

[63] Guha, R.; Serra, J.R.; Jurs, P.C. Generation of QSAR sets with a self-organizing map. *J. Mol. Graph. Model.,* **2004,** *23*(1), 1–14.

[64] Hoshi, K.; Kawakami, J.; Kumagai, M.; Kasahara, S.; Nishimura, N.; Nakamura, H.; Sato, K. An analysis of thyroid function diagnosis using Bayesian-type and SOM-type neural networks. *Chem. Pharm. Bull.,* **2005,** *53*(12), 1570–1574.

[65] Nandi, S.; Vracko, M.; Bagchi, M.C. Anticancer activity of selected phenolic compounds: QSAR studies using ridge regression and neural networks. *Chem. Biol. Drug Des.,* **2007,** *70*(5), 424–436.

[66] Xiao, Y.D.; Clauset, A.; Harris, R.; Bayram, E.; Santago, P.; Schmitt, J.D. Supervised self-organizing maps in drug discovery. 1. Robust behavior with overdetermined data sets. *J. Chem. Inf. Model.,* **2005,** *45*(6), 1749–1758.

[67] Yan, A.; Wang, Z.; Cai, Z. Prediction of human intestinal absorption by GA feature selection and support vector machine regression. *Int. J. Mol. Sci.,* **2008,** *9*(10), 1961–1976.

[68] Hornik, K.; Stinchcombe M.; White H. Multilayer feedforward networks are universal approximators. *Neural Networks,* **1989,** *2*(5), 359-366.

[69] Kovacs, Z.L. *Redes Neurais Artificiais.* 3rd Ed. Livraria da Física. **2002.**

[70] Haykin, S.S. *Neural networks: a comprehensive foundation*, 2nd Ed. Prentice Hall, **1999**.

[71] Zheng, F.; Zheng, G.; Deaciuc, A.G.; Zhan, C.G.; Dwoskin, L.P.; Crooks, P.A. Computational neural network analysis of the affinity of lobeline and tetrabenazine analogs for the vesicular monoamine transporter-2. *Bioorg. Med. Chem.,* **2007,** *15*(8), 2975–2992.

[72] Kuzmanovski, I.; Novič, M. Counter-propagation neural networks in Matlab. *Chemometr. Intell. Lab.,* **2008,** *90*(1), 84-91.

[73] Tso, B.; Mather, P.M. Classification Methods for Remotely Sensed Data. CRC Press. **2009.**

[74] Mlinsek, G.; Novič, M.; Hodoscek, M.; Solmajer, T. Prediction of Enzyme Binding: Human Thrombin Inhibition Study by Quantum Chemical and Artificial Intelligence Methods Based on X-ray Structures. *J. Chem. Inf. Comp. Sci.,* **2001,** *41*(3-6), 1286-1294.

[75] Sabet, R.; Fassihi, A.; Hemmateenejad, B.; Saghaei, L.; Miri, R.; Gholami, M. Computer-aided design of novel antibacterial 3-hydroxypyridine-4-ones: application of QSAR methods based on the MOLMAP approach. *J. Comput-Aid. Mol. Des.,* **2012,** *26*(3), 349-361.

[76] Minovski, N.; Vračko, M.; Šolmajer, T. Quantitative structure-activity relationship study of antitubercular fluoroquinolones. *Mol. Divers.,* **2011,** *15*(2), 417-426.

[77] Bishop, C.M. *Neural Networks for Pattern Recognition.* Oxford University Press, **2005.**

[78] Winkler, D.A.; Burden, F.R. Application of Neural Networks to Large Dataset QSAR, Virtual Screening, and Library Design. In: *Combinatorial Library, Methods in Molecular Biology*. English, L. B. (Eds.), Springer Protocols, **2002,** 325-367.

[79] de Freitas, J.F.G. *Bayesian Methods for Neural Networks.* PhD thesis, Cambridge University Engineering Department, **2003.**

[80] Neal, R.M. *Bayesian Learning for Neural Networks.* PhD thesis, University of Toronto, **1995.**

[81] Fernandez, M.; Caballero, J. Modeling of activity of cyclic urea HIV-1 protease inhibitors using regularized-artificial neural networks. *Bioorg. Med. Chem.,* **2006,** *14*(1), 280–294.

[82] Neal, R.M. *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics, Springer: New York, **1996.**

[83] Caballero, J.; Fernandez, M.; Gonzalez-Nilo, F.D. Structural requirements of pyrido[2,3-d]pyrimidin-7-one as CDK4/D inhibitors: 2D autocorrelation, CoMFA and CoMSIA analyses. *Bioorg. Med. Chem.,* **2008,** *16*(11), 6103-6115.

[84] Ajay; Walters, W.P.; Murcko, M.A. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.,* **1998,** *41*(18), 3314-3324.

[85] Wang, Y.; Li, Y.; Li, Y.; Yang, S.; Yang, L. Modeling Km values using electrotopological state: Substrates for cytochrome P450 3A4-mediated metabolism. *Bioorg. Med. Chem. Letters.,* **2005,** *15*(18), 4076-4084.

[86] Burden, F.R.; Ford, M.G.; Whitley, D.C.; Winkler, D.A. Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Neural Networks. *J. Chem. Inf. Comp. Sci.,* **2000,** *40*(6), 1423-1430.

[87] Schölkopf, B.; Burges, C.J.C.; Smola, A.J. *Advances in kernel methods: support vector learning.* MIT Press: Cambridge, **1999**.

[88] Taylor, J.S.; Cristianini, N. An introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, **2005.**

[89] Boser, E.; Vapnik, N.; Guyon, I. M. Training Algorithm Margin for Optimal Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory.* **1992,** *6*, 144–152.

[90] Burges, C.J.C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.,* **1998,** *2*(2), 121–167.

[91] Louis, B.; Agrawal, V.K.; Khadikar, P.V. Prediction of intrinsic solubility of generic drugs using mlr, ann and svm analyses. *Eur. J. Med. Chem.,* **2010,** *45*(9), 4018–4025.

[92] Palmer, A.M. New horizons in drug metabolism, pharmacokinetics and drug discovery. *Drug News Perspect.*, **2003,** *16*(1), 57–62.

[93] Ansel, H.C.; Popovich, N.G.; Loyd, V. Farmacotécnica: Formas Farmacêuticas e Sistemas de Liberação de Fármacos. 6th Ed. Premier, **2003**.

[94] Nigsch, F.; Klaffke, W.; Miret, S. *In vitro* models for processes involved in intestinal absorption. *Expert. Opin. Drug Met.,* **2007,** *3*(4), 545–556.

[95] Patrick, G.L. *An introduction to Medicinal Chemistry.* Oxford University Press, **2009.**

[96] Kennedy, T. Managing the drug discovery/development interface. *DDT.,* **1997,** *2*(10), 436–444.

[97] Qin, Y.; Deng, H.; Yan, H.; Zhong, R. An accurate nonlinear qsar model for the antitumor activities of chloroethylnitrosoureas using neural networks. *J. Mol. Graph. Model.,* **2011,** *29*(6), 826–833.

[98] Darnag, R.; Mostapha Mazouz, E.L.; Schmitzer, A.; Villemin, D.; Jarid, A.; Cherqaoui, D. Support vector machines: Development of qsar models for predicting anti-hiv-1 activity of tibo derivatives. *Eur. J. Med. Chem.,* **2010,** *45*(4), 1590–1597.

[99] Sorich, M.J.; Miners, J.O.; McKinnon, R.A.; Winkler, D.A.; Burden, F.R.; Smith, P.A. Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human udp-glucuronosyltransferase isoforms. *J. Chem. Inf. Comp. Sci.,* **2003,** *43*(6), 2019–2024.

[100] Fatemi, M.H.; Heidari, A.; Ghorbanzade, M. Prediction of aqueous solubility of drug-like compounds by using an artificial neural network and least-squares support vector machine. *Bull. Chem. Soc. Jpn.*, **2010**, *83*(11), 1338–1345.

[101] Pourbasheer, E.; Riahi, S.; Ganjali, M.R.; Norouzi, P. Qsar study on melanocortin-4 receptors by support vector machine. *Eur. J. Med. Chem.,* **2010,** *45*(3), 1087–1093.

[102] Kotsiantis, S.B. Supervised machine learning: A review of classification techniques. *Informatica,* **2007,** *31*(3), 249-268.

[103] Thai, K.; Nguyen, T.; Ngo, T.; Tran, T.; Huynh, T. A Support Vector Machine Classification Model for Benzo[c]phenathridine Analogues with Topoisomerase-I Inhibitory Activity. *Molecules,* **2012,** *17*(4), 4560-4582.

[104] Doucet, J.P.; Barbault, F.; Xia, H.; Panaye, A.; Fan, B.T. Nonlinear SVM Approaches to QSPR/QSAR Studies and Drug Design. *Curr. Comput-Aid. Drug.,* **2007,** *3*(4), 263-289.

[105] Yao, X.J.; Panaye, A.; Doucet, J.P.; Zhang, R.S.; Chen, H.F.; Liu, M.C.; Hu,

Z.D.; Fan, B. T. Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. *J. Chem. Inf. Comput. Sci.,* **2004,** *44*(4), 1257-1266.

[106] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.,* **2001,** *26*(1), 5–14.

[107] Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In: Lipkowitz, K. B.; Cundari, T. R. Reviews in Computational Chemistry. Wiley-VCH, John Wiley & Sons, Inc. **2007,** *23,* 291-400.

[108] Yin, L.; Huang, C.H.; Ni, J. Clustering of gene expression data: Performance and similarity analysis. *BMC Bioinformatics,* **2006,** *7*(4), S19.

[109] Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.,* **1998,** *70*(3), 175-222.

[110] Hussain, A.S. Artificial neural network based *in vitro-in vivo* correlations. *Adv. Exp. Med. Biol.,* **1997,** *423*, 149-58.

[111] Winkler, D.A. Neural networks in ADME and toxicity prediction. *Drug Future,* **2004,** *29*(10), 1043-1057.

[112] Fogel, G.; Corne, D. *Evolutionary computation in bioinformatics*. Imprint: Morgan Kaufmann, **2002.**

[113] Jolliffe, I.T. *Principal Component Analysis.* 2nd Ed. Springer, **2002.**

[114] Cronin, M.T.D.; Schultz, T.W. Pitfalls in qsar. *J. Mol. Struct. THEOCHEM,* **2003,** *622*(1-2), 39–51.

[115] Muegge, I.; Oloff, S. Advances in virtual screening. *DDT.,* **2006,** *3*(4), 405–411.

[116] Seifert, M.H.J. Targeted scoring functions for virtual screening. *DDT.,* **2009,** *14*(11-12), 562–569.

[117] Oloff, S.; Mailman, R.B.; Tropsha, A. Application of validated qsar models of d1 dopaminergic antagonists for database mining. *J. Med. Chem.,* **2005,** *48*(23), 7322–7332.

[118] Mishra, P.; Tripathi, V.; Yadav, B.S. Insilco qsar modeling and drug development process. *GERF Bull. Biosci.,* **2010,** *1*(1), 37–40.

[119] Shen, M.; Béguin, C.; Golbraikh, A.; Stables, J.P.; Kohn, H.; Tropsha, A. Application of predictive qsar models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.,* **2004,** *47*(9), 2356-2364.

[120] Cherkasov, A.; Shi, Z.; Fallahi, M.; Hammond, G.L. Successful in silico discovery of novel nonsteroidal ligands for human sex hormone binding globulin. *J. Med. Chem.,* **2005,** *48*(9), 3203–3213.

[121] Tsai, K.C.; Chen, S.Y.; Liang, P.H.; Lu, I.L.; Mahindroo, N.; Hsieh, H.P.; Chao, Y.S.; Liu, L.; Liu, D.; Lien, W.; Lin, T.H.; Wu, S.Y. Discovery of a novel family of sars-cov protease inhibitors by virtual screening and 3d-qsar studies. *J. Med. Chem.,* **2006,** *49*(12), 3485–3495.

[122] Reddy, A.S.; Pati, S.P.; Kumar, P.P.; Pradeep, H.N.; Sastry, G.N. Virtual screening in drug discovery – a computational perspective. *Curr. Prot. Pept. Sc.,* **2007,** *8*(4), 329–351.

[123] Hu, L.; Chen, G.; Chau, R.M.W. A neural networks-based drug discovery approach and its application for designing aldose reductase inhibitors. *J. Mol. Graph. Model.,* **2006,** *24*(4), 244–253.

[124] Afantitis, A.; Melagraki, G.; Koutentis, P.A.; Sarimveis, H.; Kollias, G. Ligand - based virtual screening procedure for the prediction and the identification of novel β-amyloid aggregation inhibitors using Kohonen maps and Counterpropagation Artificial Neural Networks. *Eur. J. Med. Chem.,* **2011,** *46*(2), 497–508.

[125] Noeske, T.; Trifanova, D.; Kauss, V.; Renner, S.; Parsons, C.G.; Schneider, G.; Weil, T. Synergism of virtual screening and medicinal chemistry: Identification and optimization of allosteric antagonists of metabotropic glutamate receptor 1. *Bioorg. Med. Chem.,* **2009,** *17*(15), 5708–5715.

[126] Schneider, G.; Nettekoven, M. Ligand-based combinatorial design of selective purinergic receptor (a2a) antagonists using self-organizing maps. *J. Comb. Chem.,* **2003,** *5*(3), 233–237.

[127] Karpov, P.V.; Osolodkin, D.I.; Baskin, I.I.; Palyulin, V.A.; Zefirov, N.S. One-class classification as a novel method of ligand-based virtual screening: The case of glycogen synthase kinase 3β inhibitors. *Bioorg. Med. Chem. Lett.,* **2011,** *21*(22), 6728–6731.

[128] Molnar, L.; Keseru, G.M. A neural network based virtual screening of cytochrome p450 3a4 inhibitors. *Bioorg. Med. Chem. Lett.,* **2002,** *12*(3), 419–421.

[129] Jorissen, R.N.; Gilson, M.K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.,* **2005,** *45*(3), 549–561.

[130] Plewczynski, D.; Spieser, S.A.; Koch, U. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.,* **2006,** *46*(3), 1098–1106.

[131] Lobanov, V. Using artificial neural networks to drive virtual screening of combinatorial libraries. *DDT.,* **2004,** *2*(4), 149–156.

[132] van de Waterbeemd, H.; Gifford, E. Admet in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.,* **2003,** *2*(3), 192–204.

[133] Walker, D.K. The use of pharmacokinetic and pharmacodynamic data in the assessment of drug safety in early drug development. *Brit. J. Clin. Pharmaco.,* **2004,** *58*(6), 601–608.

[134] Xu, J.; Hagler, A. Chemoinformatics and drug discovery. *Molecules*, **2002,** *7*(8), 566–600.

[135] Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *J. Chem. Inf. Comput. Sci.,* **2003,** *43*(6), 1882-1889.

[136] Ma, L.; Cheng, C.; Liu, X.; Zhao, Y.; Wang, A.; Herdewijn, P. A neural network for predicting the stability of RNA/DNA hybrid duplexes. *Chem. Int. Lab. Sys.,* **2004,** *70*(2), 123-128.

[137] Niwa, T. Prediction of Biological Targets Using Probabilistic Neural Networks and Atom-Type Descriptors. *J. Med. Chem*., **2004,** *47*(10), 2645-2650.

[138] Taskinen, J.; Yliruusi, J. Prediction of physicochemical properties based on neural network modelling. *Adv. Drug Deliv. Rev.,* **2003,** *55*(9), 1163-1183.

[139] Niculescu, S.P. Artificial neural networks and genetic algorithms in QSAR. *J. Mol. Struct. THEOCHEM*, **2003,** *622*(1-2), 71-83.

[140] Moda, T.L.; Honório, K.M.; Andricopulo, A.D. Propriedades farmacocinéticas no planejamento de fármacos. In: *Química Medicinal: Métodos em Planejamento Racional de Fármacos.* Montanari, C.A. (Org.). EDUSP: São Paulo, **2011,** 515-556.

[141] Belic, A.; Grabnar, I.; Belic, I.; Karba, R.; Mrhar, A. Predicting the antihypertensive effect of nitrendipine from plasma concentration profiles using artificial neural networks. *Comput. Biol. Med.,* **2005**, *35*(10), 892–904.

[142] Funar-Timofei, S.; Ionescu, D.; Suzuki, T. A tentative quantitative structure-toxicity relationship study of benzodiazepine drugs. *Toxicol. In Vitro,* **2010,** *24*(1), 184–200.

[143] Jalali-Heravi, M.; Kyani, A. Comparative structure-toxicity relationship study of substituted benzenes to Tetrahymena pyriformis using shuffling-adaptive neuro fuzzy inference system and artificial neural networks. *Chemosphere,* **2008,** *72*(5), 733–740.

[144] Huuskonen, J. Qsar modeling with the electrotopological state indices: predicting the toxicity of organic chemicals. *Chemosphere,* **2003,** *50*(7), 949-953.

[145] Mazzatorta, P.; Vracko, M.; Jezierska, A.; Benfenati, E. Modeling toxicity by using supervised kohonen neural networks. *J. Chem. Inf. Comp. Sci.,* **2003,** *43*(2), 485–492.

[146] Singh, S.K.; Saini, S.; Verma B.; Mishra, D.N. Quantitative structure pharmacokinetic relationship using artificial neural network: A review. *IJPSDR.,* **2009,** *1*(3), 144–153.

[147] Stojic, N.; Eric, S.; Kuzmanovski, I. Prediction of toxicity and data exploratory analysis of estrogen-active endocrine disruptors using counter-propagation artificial neural networks. *J. Mol. Graph. Model.,* **2010,** *29*(3), 450-460.

[148] Winkler, D. Neural networks as robust tools in drug lead discovery and development. *Mol. Biotechnol.,* **2004,** *27*(2), 139–167.

Received: April 26, 2012 **Revised: July 03, 2012** **Accepted: July 05, 2012**