

Genetic neural network modeling of the selective inhibition of the intermediate-conductance Ca^{2+} -activated K^{+} channel by some triarylmethanes using topological charge indexes descriptors

Julio Caballero^a, Miguel Garriga^b & Michael Fernández^{a,*}

^aMolecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740, Matanzas, Cuba; ^bPlant Biotechnology Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, C.P. 44740, Matanzas, Cuba

Received 25 June 2005; accepted 19 October 2005
© Springer 2005

Key words: Bayesian regularization, clotrimazole, genetic algorithm, ion channel, neural networks, QSAR

Summary

Selective inhibition of the intermediate-conductance Ca^{2+} -activated K^{+} channel (IK_{Ca}) by some clotrimazole analogs has been successfully modeled using topological charge indexes (TCI) and genetic neural networks (GNNs). A neural network monitoring scheme evidenced a highly non-linear dependence between the IK_{Ca} blocking activity and TCI descriptors. Suitable subsets of descriptors were selected by means of genetic algorithm. Bayesian regularization was implemented in the network training function with the aim of assuring good generalization qualities to the predictors. GNNs were able to yield a reliable predictor that explained about 97% data variance with good predictive ability. On the contrary, the best multivariate linear equation with descriptors selected by linear genetic search, only explained about 60%. In spite of when using the descriptors from the linear equations to train neural networks yielded higher fitted models, such networks were very unstable and had relative low predictive ability. However, the best GNN BRANN 2 had a Q^2 of LOO of cross-validation equal to 0.901 and at the same time exhibited outstanding stability when calculating 80 randomly constructed training/test sets partitions. Our model suggested that structural fragments of size three and seven have relevant influence on the inhibitory potency of the studied IK_{Ca} channel blockers. Furthermore, inhibitors were well distributed regarding its activity levels in a Kohonen self-organizing map (KSOM) built using the inputs of the best neural network predictor.

Introduction

K^{+} channels constitute a remarkably diverse family of membrane-spanning proteins that have a wide range of functions in electrically excitable and unexcitable cells. One important class opens in response to a calcium concentration increase within the cytosol. This effect was first demonstrated in red blood cells with which it was found that metabolic

inhibition which caused cytosolic Ca^{2+} to rise also resulted in a large increase in K^{+} permeability (the Gárdos effect) [1]. The change in permeability was subsequently shown to be mediated by the opening of Ca^{2+} -activated K^{+} channels (K_{Ca}) in the membrane of red cells. Afterwards, pharmacological and electrophysiological evidence [2–4] and, more recently, structural evidence from cloning studies [5], have established that exists several kinds of Ca^{2+} -activated K^{+} channels. Those in the red cell belong to the intermediate conductance (IK_{Ca}) subtype, so named because its single channel conductance (10–30 pS) lies between that of the

*To whom correspondence should be addressed. Phone: +53-45-26-1251; Fax: +53-45-25-3101; E-mail: michael.fernandez@umcc.cu

small conductance and high conductance subtypes [2, 3].

Resting human T lymphocytes possess about 400 *Kv1.3* channels and roughly 2–20 functional *IKCa1* channels, products of the *IKCa1* gene. The membrane potential of resting T cells is maintained by *Kv1.3* channels rather than by *IKCa1*, and selective inhibitors of *Kv1.3* suppress the activation response. In contrast, mitogen-activated human T lymphocytes exhibit 300–800 functional *IKCa1* channels along with 400–500 *Kv1.3* channels [6]. Because expression of *IKCa1* channels is dramatically enhanced in activated T cells, in parallel with enhanced $[Ca^{2+}]_i$ signaling, a strategy targeting *IKCa1* channels could be especially effective in suppressing chronically activated T cells and could perhaps lead to therapy for autoimmune disorders.

Several compounds have been shown to block the *IKCa*-mediated Ca^{2+} -activated K^+ permeability in red blood cells. These include quinine and quinidine, some carbocyanine dyes, charybdotoxin, dequalinium and some of its derivatives, nifedipine, nitrendipine, cetiedil and clotrimazole and some related compounds [7].

Clotrimazole directly block the *IKCa* channel in human erythrocytes, colonic epithelium, and human T lymphocytes at nanomolar concentrations. Because of its potent channel-blocking activity, clotrimazole is being clinically valued for the treatment of erythrocyte dehydration in sickle cell disease and secretory diarrheas [6]. However, it had been also reported that this compound inhibits mammalian P450-dependent enzymes, as well. This lack of specificity limits the use of clotrimazole as a pharmacological tool, creating a need for a truly selective *IKCa1* inhibitor [6].

Computational models that are able to predict the biological activity of compounds by its structural properties, are powerful tools to design highly active molecules. In this sense, quantitative structure–activity relationships (QSAR) studies have been successfully applied for modeling biological activities of natural and synthetic chemicals [8]. Based on the molecular graph representation of a chemical structure, Graph-Theoretical and Topological methods are included in the most QSAR studies [9–17]. In such studies, a set of topological descriptors encoding structural features of the compounds under investigation are computed over the molecular graphs, afterwards

multivariate linear or/and non-linear relationships are established between a reduced subset of variables and the biological property. Among these approaches, topological charge indexes (TCI) have been successfully introduced for modeling physicochemical [9] and biological properties [15] and for pharmaceutical research [10–12].

Since interactions between a chemical and its biological target are often non-linear, artificial neural network (ANN) methodology rather than multilinear regression analysis (MRA) had been successfully applied in QSAR studies of biological properties in the last decade [15–25]. Besides the non-linearity among biological activities and the computed molecular descriptors, another major problem arises when the number of calculated variables exceeds the number of compounds in the data set, so that one is dealing with an undetermined problem where undesirable overfitting can result [26]. This problem can be handled by implementing a feature selection routine that determines which of the descriptors have a significant influence on the activity of a set of compounds. Genetic algorithm (GA) rather than forward or backward elimination procedure has been successfully applied for feature selection in QSAR studies when the dimensionality of the data set is high and/or the interrelations between variables are convoluted [18–23, 25].

In this work, we employed genetic neural networks (GNNs) for predicting the selective inhibition of the intermediate-conductance Ca^{2+} -activated K^+ channel, *IKCa1*, by thirty triaryl methane (clotrimazole analogs) reported by Wulff et al. [6]. TCI descriptors were used for encoding structural information from the studied compounds and a GA was applied for relevant feature extraction from the descriptor matrix. In order to gain in performance in both robustness of predictions and speed of computation, Bayesian regularization was implemented in a Levenberg–Marquardt algorithm for mean square error (MSE) minimization during supervised training of full-connected feed-forward ANN. Using this strategy, three, four and five-descriptors models were achieved. Reliability of the models was settled examining the square correlation coefficient (R^2) and the root mean square error (RMSE) of the data fitting and the square correlation coefficient of the leave-one-out (LOO) cross-validation (Q^2) and the RMSE of the LOO cross-validation

(RMSE_{cv}). The stability and predictive power of the best predictor was examined by performing calculation of several randomly constructed training/test set partitions in the data set. In addition, the versatility of ANNs was used also for mapping the *IK*_{Ca} inhibitory activities on a topological map using competitive neural networks in order to address structural features related with the blocking activity of the studied compounds.

Charge topological indexes approach

The binding of a substrate to its receptor is dependent on the shape of the substrate and on a variety of effects such as the molecular electrostatic potential, polarizability, hydrophobicity and lipophilicity. Therefore, in a QSAR study the strategy for encoding molecular information must in some way, either explicitly or implicitly, account for these physicochemical effects. Furthermore, usually data sets include molecules of different size with different numbers of atoms, so the structural encoding structures must allow comparing such molecules [8].

The molecular charge distribution plays an important role in many biological and pharmacological activities. It can be evaluated through physicochemical parameters such as dipole moment and electronic polarizability. However, gathering these interactions in certain molecular code is more useful for QSAR purpose. In this sense, efforts had been focused to develop charge-related descriptors. Kier and Hall [27, 28] tried to encode how each atom in a given molecule ‘reflects’ the steric and electronic effects of the surrounding atoms and as such, could be best described as information rich atomic descriptors. In this sense, they developed the concept of E-states, an electrotopological-state index for atoms in a molecule. More recently, Carbó-Dorca et al. [29] proposed the use of electron–electron repulsion energy, in connection with molecular quantum similarity measures derived from quantum-chemical calculations, as a molecular descriptor in QSAR studies of biological activities [29].

In this context, TCI descriptors, a kind of Graph-Theoretical and Topological descriptors, were defined by Galvez et al. and their ability to evaluate the charge transfers between pairs of atoms and the global charge transfer was demon-

strated [9, 10]. For calculating TCI descriptors, H-depleted molecular structure is represented as a graph *G* (Table 1). Differently to the common topological 2D descriptors, TCI are calculated using the “inverse square topological distance matrix” instead the topological distance matrix. Since the charge influence decreases with the square of the distance, Galvez et al. introduced the “inverse square topological distance matrix” denoted by *D*^{*} in which matrix elements are the inverse square of the corresponding element in the topological distance matrix *D* (Table 1). In order to avoid division by zero, the diagonal entries of the topological distance matrix remain the same, so diagonal entries of *D*^{*} are 0 (Table 1). In Table 1, *A* is the adjacency (*N*×*N*) matrix of the molecular graph *G*, where *N* is the number of vertices (atoms different to hydrogen). Finally, *M* matrix is defined as *M* = *A*×*D*^{*}. Regarding this, TCI descriptors *GGIk*, *JGIk* and *JGT* are defined as:

$$GGIk = \sum_{i=1, j=i+1}^{i=N-1, j=N} |CT_{ij}| \delta(k, D_{ij}) \quad (1)$$

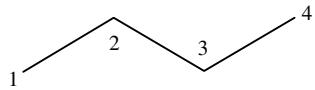
$$JGIk = \frac{GGIk}{(N-1)} \quad (2)$$

$$JGT = \sum_{k=1}^{k_{\max}} JGIk \quad (3)$$

where δ is Kronecker’s delta and $CT_{ij} = m_{ij} - m_{ji}$. “*m*” stands for the elements of the *M* matrix. Thus, *GGIk* represents the sum of all the CT_{ij} terms, with $D_{ij}=k$, being D_{ij} the entries of the topological distance matrix (*D*) and *k* ranging from 1 up to 10. These indexes represent a strictly topological quantity plausibly correlating with the charge distribution inside the molecule [9, 10].

In a previous work of our group [15], TCI descriptors yielded remarkable better results than others, such as physico-chemical quantities, 2D topological and 3D descriptors, for modeling Aldose Reductase inhibition by flavonoid compounds. Consequently, we try TCI descriptors now for modeling the selective inhibition of the intermediate-conductance Ca^{2+} -activated K^{+} channel by some triarylmethanes. Similarly to the above mentioned flavonoids compounds [15],

Table 1. Representation of a molecular graph G and their D , D^* , A and M matrixes.

Molecular Graph G	
D	$\begin{vmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{vmatrix}$
D^*	$\begin{vmatrix} 0 & 1 & \frac{1}{4} & \frac{1}{9} \\ 1 & 0 & 1 & \frac{1}{4} \\ \frac{1}{4} & 1 & 0 & 1 \\ \frac{1}{9} & \frac{1}{4} & 1 & 0 \end{vmatrix}$
A	$\begin{vmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{vmatrix}$
$M = A \times D^*$	$\begin{vmatrix} 1 & 0 & 1 & \frac{1}{4} \\ \frac{1}{4} & 2 & \frac{1}{4} & \frac{10}{9} \\ \frac{10}{9} & \frac{1}{4} & 2 & \frac{1}{4} \\ \frac{1}{4} & 1 & 0 & 1 \end{vmatrix}$

the triarylmethanes here studied do not vary appreciably in size and/or shape in such a way that mainly relative positions of substituents change in the triarylmethane-like structures. Then, electrotopological conformation is the main feature that it is varying from one compound to another. Those evidences, in conjunction with the fact that TCI describe molecular topology and charge transfer through the molecule, support our interest in such descriptors for building highly reliable models for the inhibitory activity of IK_{Ca} blockers. Furthermore, we prefer 2D topological or 3D descriptors over other simpler and more physical descriptors such as octanol/water partition coefficient or molar refractivity because of, in spite of their difficult interpretation, they usually yield outstanding QSAR models when are used in combination with ANNs [15–20].

Using Dragon computer software [30], a set of 21 TCI descriptors was computed and thus a 21×30 data matrix was obtained for modeling the selective inhibition of the intermediate-conductance Ca^{2+} -activated K^+ channel, $IKCa1$, by thirty clotrimazole analogs.

Genetic neural networks

Genetic algorithm

The GA is used to select the features that are most significant for the molecular data set. GAs are stochastic optimization methods inspired by evolutionary principles [31]. The distinctive aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores different regions in parameter space [32]. The GA implemented in this paper is a version of the So and Karplus report [21] programmed within the Matlab environment using GA and Neural Networks Tool Boxes [33] that was previously reported by our group [20].

An individual in the population is represented by a string of integers which means the numbering of the columns in the data matrix. In the original study, the fitness of the individual was determined by a variety of fitness functions which are proportional to the residual error of the training set, the test set, or even the cross-validation set from the neural network simulations. In our approach, we tried the MSE of data fitting as the individual fitness function. The basic design of the implemented GA is summarized in the flow diagram shown in Figure 1. The first step is to create a gene pool (population) of N individuals. Each individual encodes the same number of descriptors; the descriptors are randomly chosen from a common data matrix, and in a way such that (1) no two individuals can have exactly the same set of descriptors and (2) all descriptors in a given individual must be different. The fitness of each individual in this generation is determined by the MSE of a trained ANN and scaled using and scaling function. A top scaling fitness function scaled a top fraction of the individuals in a population equally; these individuals have the same probability to be reproduced while the rest are assigned the value 0 [34].

The next step, a fraction of children of the next generation is produced by crossover (crossover children) and the rest by mutation (mutation children) from the parents [34]. Sexual and asexual reproductions take place so that the new offspring contains characteristics from two or one of its parents (Figure 2). In a sexual reproduction two individuals are selected probabilistically on the basis of their scaled fitness scores and serve as

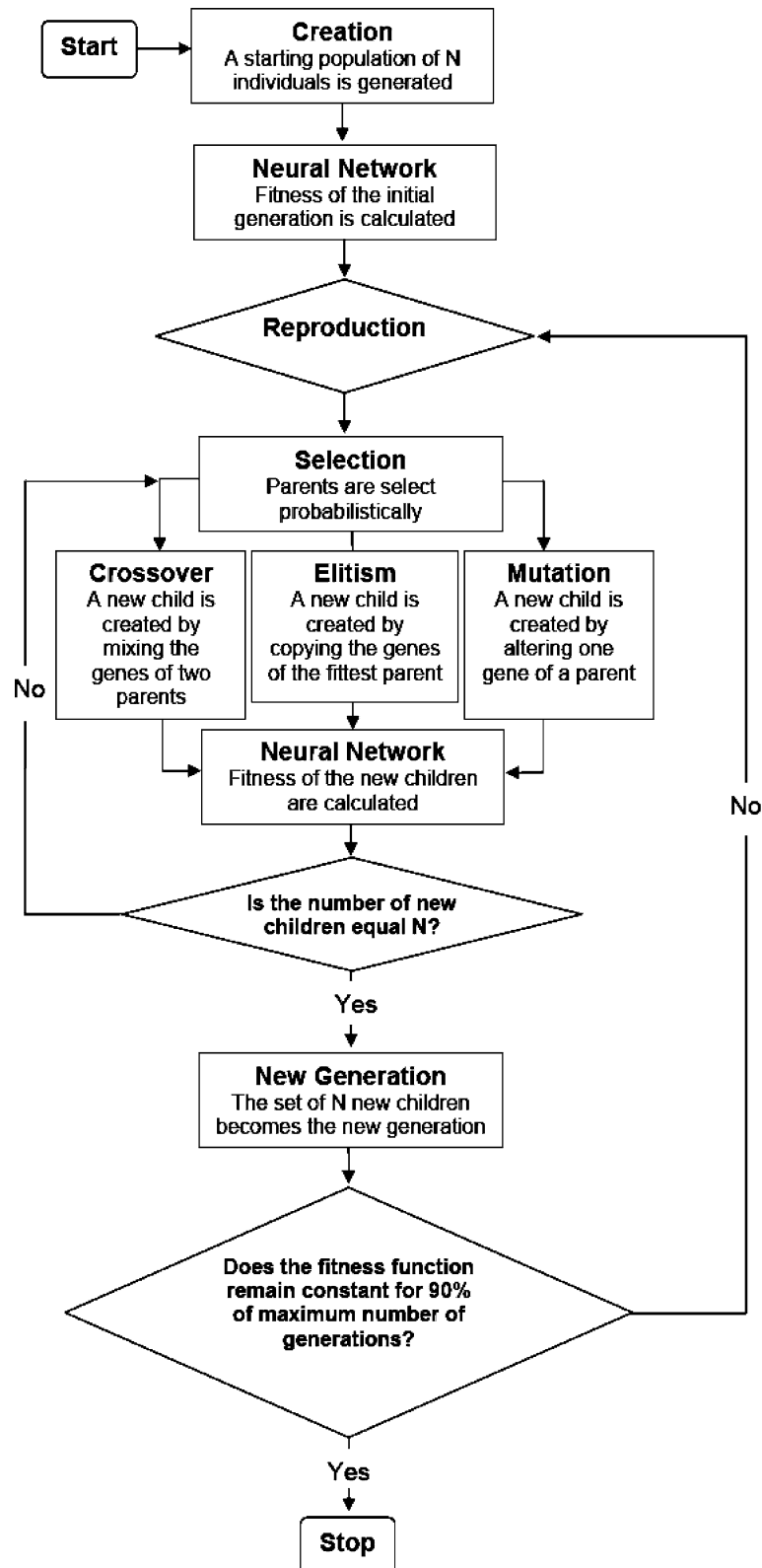


Figure 1. Flow diagram describing the strategy for the GA. See Figure 2 for the detailed descriptions of the reproduction strategy.

parents. Next, in a crossover each parent contributes a random selection of half of its descriptor set and a child is constructed by combining these two halves of “genetic code”. Finally, the rest of the individuals in the new generation are obtained by asexual reproduction when parents selected randomly are subjected to a random mutation in one of its genes; i.e., one descriptor is replaced by another.

Similarly to So and Karplus [21], we also included elitism which protects the fittest individ-

ual in any given generation from crossover or mutation during reproduction. The genetic content of this individual simply moves on to the next generation intact. This selection, crossover and mutation process is repeated until all of the N parents in the population are replaced by their children [34]. The fitness score of each member of this new generation is again evaluated, and the reproductive cycle is continued until a 90% of the generations showed the same target fitness score [25].

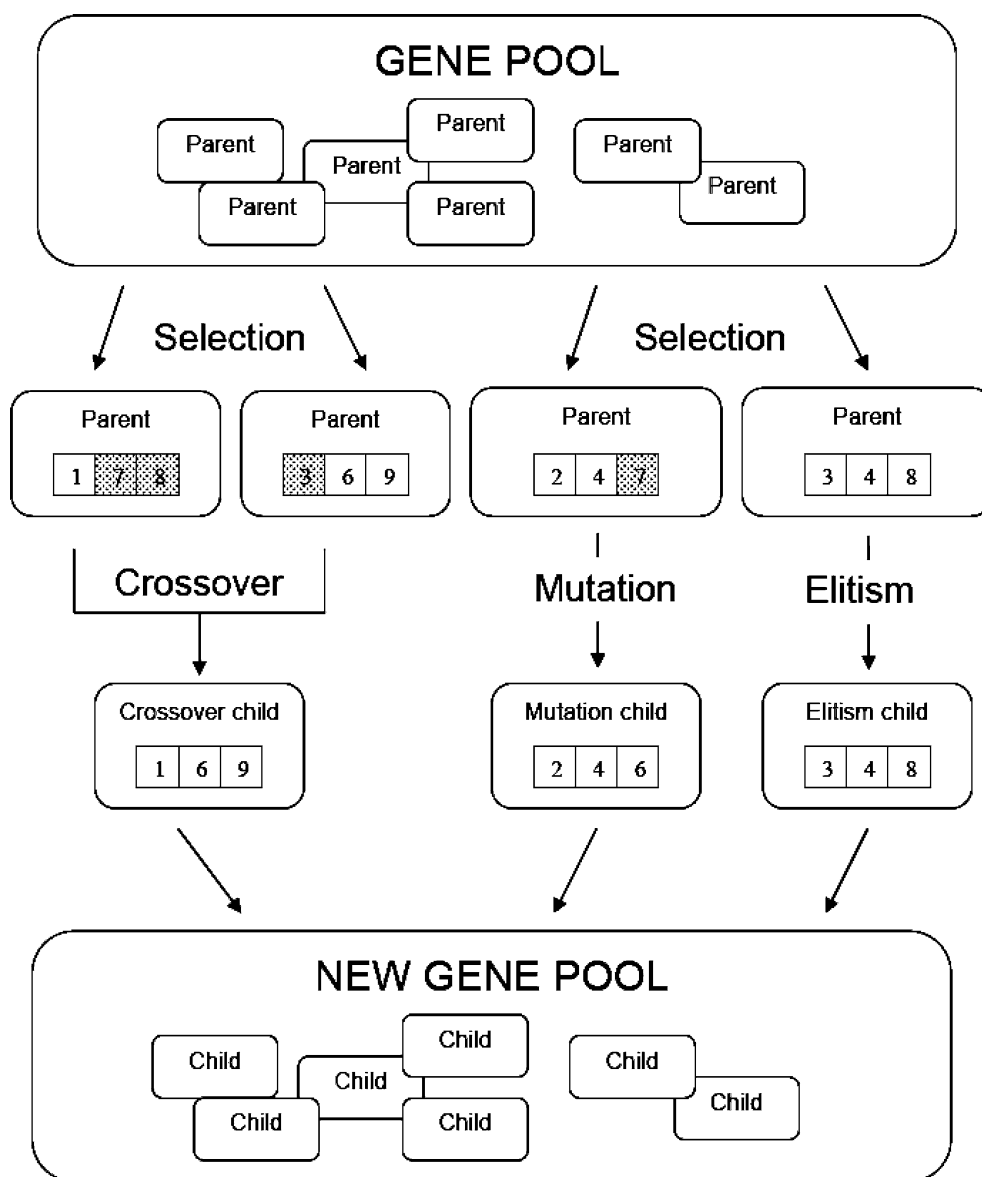


Figure 2. Schematic diagram describing the reproduction strategy in GA algorithm.

Artificial neural networks

Feed-forward neural networks

ANNs are computer-based models in which a number of processing elements (also called neurons, units, or nodes) are interconnected by links in a netlike structure forming “layers” [35]. A variable value is assigned to every neuron. The neurons can be one of three different kinds. The input neurons receive their values from independent variables forming the input layer. The hidden neurons collect values from other neurons, giving a result that is passed to a successor neuron. The output neurons take values from other units and correspond to different dependent variables forming the output layer. In this sense, network architecture is commonly represented as I–H–O, where I, H and O are the number of neurons in the input, hidden and output layers respectively (Figure 3).

The links between units have associated values, named weights, that condition the values assigned to the neurons. There exist additional weights assigned to bias values that act as neuron value offsets. The weights are adjusted through a training process in order to minimize network MSE. Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output.

The characteristics of the ANNs have been found to be suitable for data processing, in which the functional relationship between the input and the output is not previously defined. This is due to the fact that structure–activity relationships are often non-linear and very complex and neural networks are able to approximate any kind of

analytical continuous function, according to Kolmogorov’s theorem [36].

Matlab computer software [33] was used for implementing fully connected, three-layer, back-propagated feed-forward Bayesian Regularized Artificial Neural Networks (BRANN). In these networks, the transfer function of input and output layers was linear, and the hidden layer had neurons with a hyperbolic tangent transfer function. Inputs and target were normalized prior to network training. BRANN training was carried out according to the Levenberg–Marquardt optimization. The initial value for μ was 0.005 with decrease and increase factors of 0.1 and 10 respectively. The training was stopped when μ became larger than 10^{10} [37].

When parameters (weights and biases) increase, network loses its ability to generalize. Error on the training set is driven to a very small value, but when new data is presented to the network the error is large. The predictor has memorized the training examples, but it has not learned to generalize to new situations, it means network overfits the data.

Typically, training aims to reduce the sum of squared errors:

$$F = \text{MSE} = \frac{1}{N} \sum_{i=1}^N (Y_i - A_i)^2 \quad (4)$$

Bayesian regularization involves modifying the performance function (F), it is possible to improve generalization by adding an additional term [38].

$$F = \beta \times \text{MSE} + \alpha \times \text{MSW} \quad (5)$$

$$\text{MSW} = \frac{1}{N} \sum_{j=1}^n w_j^2 \quad (6)$$

In these equations F is the network performance function, MSE is the mean of the sum of squares of the network errors, N is the number of compounds, Y_i is the predicted biological activity of the compound i , A_i is the experimental biological activity of the compound i , MSW is the mean of the sum of the squares of the network weights, w_j are the weights of the neuron j , n is the number of network weights and α and β are objective function parameters [38].

The relative size of the objective function parameters dictates the emphasis for training

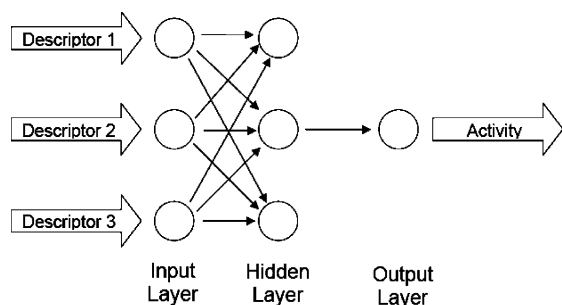


Figure 3. A prototype back-propagation neural network used in this study with a 3-3-1 architecture. Descriptors chosen by GA constitute the inputs and network is trained against the target IK_{Ca} channel-blocking activity.

getting a smoother network response. MacKay's Bayesian regularization automatically sets the correct values for the objective function parameters [38], in this sense the regularization is optimized. In the Bayesian framework the weights of the network are considered random variables. After the data is taken, the density function for the weights can be updated according to Bayes' rule:

$$P(w|D, \alpha, \beta, M) = \frac{P(D|w, \beta, M) \times P(w|\alpha, M)}{P(D|\alpha, \beta, M)} \quad (7)$$

where D represents the data set, M is the particular neural network model used, and w is the vector of network weights. $P(w|D, \alpha, \beta, M)$ is the posterior probability, that is the plausibility of a weight distribution considering the information of the data set in the model used. $P(w|\alpha, M)$ is the prior density, which represents our knowledge of the weights before any data is collected. $P(D|w, \beta, M)$ is the likelihood function, which is the probability of the data occurring, given the weights. $P(D|\alpha, \beta, M)$ is a normalization factor, which guarantees that the total probability is 1.

Considering that the noise in the training set data is Gaussian and that the prior distribution for the weights is Gaussian, the posterior probability fulfills the relation:

$$P(w|D, \alpha, \beta, M) = \frac{1}{Z_F} \exp(-F) \quad (8)$$

where Z_F depends of objective function parameters. So under this framework, minimization of F is identical to find the (locally) most probable parameters [38].

Bayesian regularization overcomes the remaining deficiencies of neural networks [39]. By using Bayesian regularization it can be obtained robust models, well matched to the data and able to make accurate predictions. Since the algorithm automatically regularizes the training process usually non-validation set is recurred so all data set can be devoted to train the network [39]. The Bayesian neural net has the potential to give models which are relatively independent of neural network architecture, above a minimum architecture, and the Bayesian regularization method estimates the number of effective parameters. BRANNs have been successfully used on QSAR studies of biological activities and specifically in drug discovery

[18–20, 39–42]. So, taking into account the kindness of the BRANNs, we used these networks for implementing our GNN algorithm.

Self-organizing maps

In order to settle structural similarities among the clotrimazole analogs, a Kohonen self-organizing map (KSOM) was built. Kohonen [43] introduced a neural network model that generates KSOM. Neurons are arranged in a 2-dimensional network. Molecules characterized by m descriptors are projected into this network. With $m > n$ a Kohonen network can be used to project a higher-dimensional space into a lower-dimensional space. Such maps of surface properties have been used for comparing wide variety of biologically active compounds [44].

$$\text{out}_{cs} \leftarrow \min \left[\sum_{i=1}^m (X_{si} - w_{ij})^2 \right] \quad (9)$$

Kohonen network is training using an unsupervised and competitive learning process. In our case, a molecule s characterized by m descriptors, x_{si} , will be projected into that (central) neuron, c_s , that has weights, w_{ji} , most similar to the input variables (Equation 9). During the learning process, weights of the neurons in the network are changed to make them even more similar to the input variables. The weights of all neurons are adjusted but to an extent that decreases with increasing distance from the central, winning neuron, cs . Finally, a molecule is projected into that neuron of the network with weights that come closest to the description of the molecule by the descriptors.

It should be noticed that the criterion embedded in Equation 9 for determining the winning neuron for a molecule basically constitutes the measure determining the similarity of molecular structures. Molecules with similar autocorrelation vectors, Xs , are projected into the same or closely adjacent neurons. KSOM were implemented in Matlab environment, neurons were initially located at a grid topology. The ordering phase was developed in 1000 steps with 0.9 learning rate until tuning neighborhood distance (1.0) was achieved. The tuning phase learning rate was 0.02. Training was performed for a period of 2000 epochs in an unsupervised manner [37].

Models evaluation and fitness function

The quality of the fit of the data set by a specific model was measured by its R^2 and the RMSE value.

$$R^2 = 1 - \frac{\sum_{i=1}^N (Y_i - A_i)^2}{\sum_{i=1}^N (Y_i - \bar{A})^2} \quad (10)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - A_i)^2} \quad (11)$$

where N is the number of compounds, Y_i and A_i are the predicted and experimental biological activities of i compound respectively, \bar{A}_i is the average experimental activity.

However, a most important measure is the prediction quality of models [26]. An internal LOO cross-validation process was carried out in such a way that the predictive power of the models was estimated by calculating Q^2 and RMSE_{cv} of LOO cross-validation. A data point is removed (left-out) from the set, and the model refitted; the predicted value for that point is then compared to its actual value. This is repeated until each datum has been omitted once; the sum of squares of these deletion residuals can then be used to calculate Q^2 , an equivalent statistic to R^2 , as well as to calculate RMSE_{cv} .

$$Q^2 = 1 - \frac{\sum_{i=1}^N (Y_i - A_i)^2}{\sum_{i=1}^N (Y_i - \bar{A})^2} \quad (12)$$

$$\text{RMSE}_{\text{cv}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - A_i)^2} \quad (13)$$

where N is the number of compounds, Y_i and A_i are the predicted and experimental biological activities of i left-out compound respectively, \bar{A}_i is the average experimental activity of left-in compounds different to i .

In addition to the internal validation, we settled the stability and predictive ability of our best predictor by performing calculation of several randomly constructed training/test set partitions in the data set. In this regard, we consider the

criterion proposed by Golbraikh and Tropsha [45] which states that a good QSAR model should have high values of both Q^2 of internal LOO cross-validation and R^2 of an external test set fitting.

As we previously pointed out, in the GNN framework implemented in this paper the MSE of trained BRANN was used as fitness function. The features of the BRANN predictors addressed in Section 'Artificial neural networks', robustness and good generalization abilities, should assure the statistical quality of the selected models.

IKCa1 channel-blocking activity data set

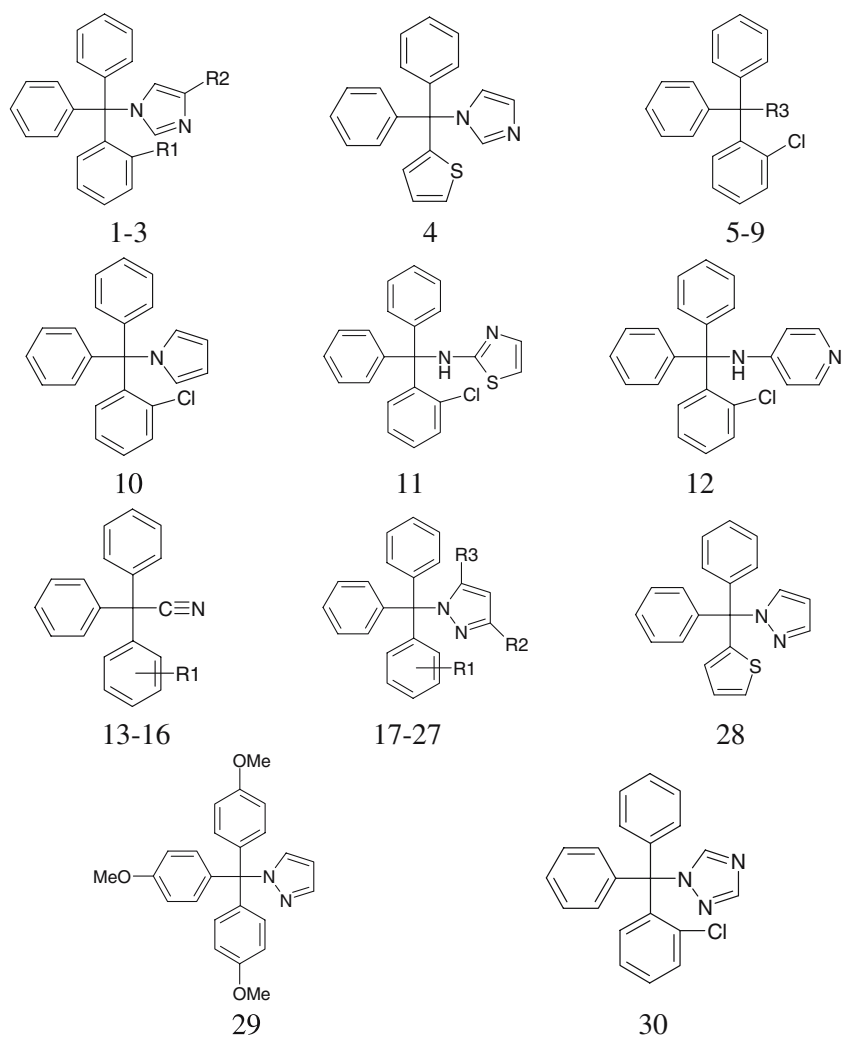
Selective blocking activities of the intermediate-conductance Ca^{2+} -activated K^+ channel, *IKCa1*, of thirty clotrimazole analogs was taken from the literature [6]. In such report, cells were studied in the whole-cell configuration of the patch-clamp technique. The holding potential in all experiments was 280 mV. Wulff et al. [6] used for measurement of *IKCa1* an internal pipette solution containing (in mM): 145 K^+ aspartate, 2 MgCl_2 , 10 Hepes, 10 K_2 EGTA, and 8.5 CaCl_2 (1 mM free Ca^{2+}), $\text{pH}=7.2$, 290–310 mOsm. To reduce currents from the native chloride channels in COS-7 cells expressing *IKCa1* channels, Na^+ aspartate Ringer was used as an external solution (in mM): 160 Na^+ aspartate/4.5 KCl /2 CaCl_2 /1 MgCl_2 /5 Hepes, $\text{pH}=7.4$ /290–310 mOsm. IK_{Ca} currents in COS-7 cells were elicited by 200 ms voltage ramps from -120 to 40 mV applied every 10 s and the reduction of slope conductance at -80 mV by drug taken as a measure of channel block. K_d at nanomolar concentrations and Hill coefficient (Hill coefficient=1.2) were determined by fitting the Hill equation to the reduction of slope conductance at 280 mV. Molecular structures, numbering of the substituents, experimental and predicted biological activities ($-\log(Kd)$) of the studied triarylmethanes are summarized in Table 2.

Results and discussion

GA based multilinear regressions analysis

As a first approach, we sought for linear relationships between the TCI descriptors and the *IKCa1*

Table 2. Chemical structures, experimental, calculated and residual IK_{Ca} blocking activities of the clotrimazole analogs using the best predictor BRANN 2



	R1	R2	R3	$-\log(Kd)$		
				Experimental	Calculated	Residual
1	Cl	H		-1.845	-1.611	-0.234
2	H	H		-3.176	-3.250	0.074
3	Cl	CH ₃		-3.079	-2.978	-0.101
4				-3.000	-3.046	0.046
5			OH	-2.716	-2.786	0.070
6			NH ₂	-3.000	-2.786	-0.214
7			NHCOCH ₃	-3.079	-3.387	0.307
8			NHCONH ₃	-3.699	-3.387	-0.313
9			CH(COOC ₂ H ₅)	-2.602	-2.610	0.008
10 ^a				-3.845	-2.121	1.724
11				-4.176	-4.154	-0.022

Table 2. Continued.

	R1	R2	R3	-log(Kd)		
				Experimental	Calculated	Residual
12				-4.477	-4.493	0.016
13	<i>o</i> -Cl			-1.778	-1.904	0.126
14	<i>o</i> -F			-1.845	-1.904	0.059
15	<i>p</i> -Cl			-2.875	-2.856	-0.019
16	<i>p</i> -F			-2.903	-2.856	-0.047
17	H	H	H	-3.398	-3.250	-0.148
18	<i>o</i> -Cl	H	H	-1.301	-1.611	0.310
19	<i>o</i> -F	H	H	-1.602	-1.611	0.009
20	<i>p</i> -Cl	H	H	-1.954	-2.163	0.209
21	<i>p</i> -F	H	H	-2.301	-2.163	-0.138
22	<i>o</i> -Cl	CH ₃	H	-3.580	-3.630	0.050
23	<i>o</i> -CF ₃	H	H	-3.176	-3.105	-0.071
24	<i>p</i> -CF ₃	CF ₃	H	-3.041	-2.978	-0.063
25	<i>o</i> -Cl	CH ₃	CH ₃	-4.079	-4.120	0.040
26	<i>o</i> -Cl	CF ₃	H	-3.301	-3.393	0.092
27	<i>o</i> -CF ₃	CF ₃	H	-4.398	-4.379	-0.019
28				-3.041	-3.046	0.005
29				-4.301	-4.290	-0.011
30				-1.653	-1.611	-0.042

^aThe blocker activity for the outlier (compound 10) was calculated by predictor BRANN 2 before its removal.

Table 3. Statistics of the best linear models of the IK_{Ca} channel-blocking activity of the clotrimazole analogs obtained by linear genetic search.

Descriptors in linear models	Full data set	Data set without compound 10			
	R^2	R^2	RMSE	Q^2	RMSE _{cv}
<i>GGI5, JGI1, JGT</i>	0.514	0.625	0.535	0.264	0.604
<i>GGI1, GGI2, GGI6, JGI6</i>	0.566	0.670	0.502	0.397	0.633
<i>GGI1, GGI6, JGI1, JGI2, JGI6</i>	0.587	0.707	0.472	0.456	0.361

channel-blocking activities of the clotrimazole analogs. In this connection, we implemented a GA routine that searches for the best fitted multivariate linear equations having three, four and five variables. Table 3 showed the statistical parameters of data fitting and LOO cross-validation of the multilinear regression models obtained. Compound 10 was considered as general outliers due to its large residual when was predicted for all the models, so we removed this triarylmethane with the aim of improving model quality. As can be observed, removing the outlier remarkably increases R^2 values for all models. However, even when removing the outlier, the statistical significances of those models were low, R^2 of data fitting

and Q^2 of LOO cross-validation were lower than 0.75 and 0.5 respectively. This fact suggests that no reliable linear dependence exists between the studied biological activity and the TCI descriptors. In addition, Figure 4a depicts the approximate functional dependence of the descriptors present in the four-variable linear model and the IK_{Ca} channel-blocking activity. For obtaining this response curves a neural network monitoring scheme was used [21]. The variation of the predicted biological activity was monitored on changing the value of one input while keeping the remaining network inputs at a constant value. This procedure was repeated for all other network inputs. These response curves were also correlated with linear

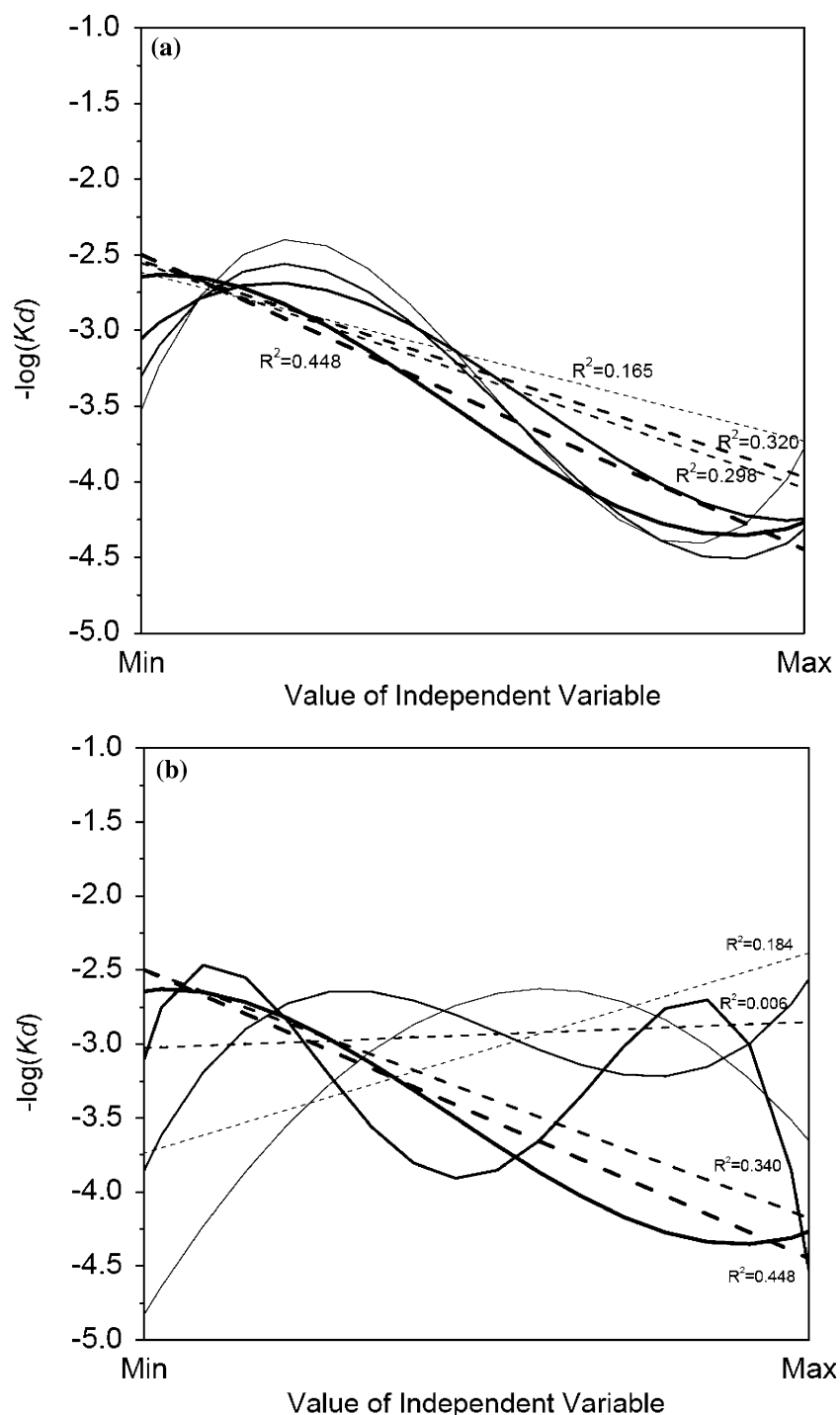


Figure 4. IK_{Ca} channel-blocking activity as a function of the individual TCIs descriptors for the best four-variable linear model (a) and the predictor BRANN 2 (b). The best fit linear models (dotted lines) and their R^2 values are also shown.

models as depicted by the straight dotted lines in the same plot. The low R^2 values of 0.448, 0.320, 0.298 and 0.165 for *GGI1*, *GGI2*, *GGI6* and *JGI6*

respectively corroborated that the relationships between the TCI descriptors and the studied biological activity is essentially non-linear.

Table 4. Statistics of the non-linear predictors of the IK_{Ca} channel-blocking activity of the clotrimazole analogs obtained by neural network training with descriptors selected by linear genetic search.

Descriptors in BRNN models	Number of hidden nodes	Data set without compound 10			
		R^2	RMSE	Q^2	RMSE _{cv}
<i>GGI5, JGI1, JGT</i>	2	0.859	0.328	0.613	0.482
	3	0.899	0.279	0.380	0.638
	4	0.960	0.176	0.053	1.111
<i>GGI1, GGI2, GGI6, JGI6</i>	2	0.876	0.309	0.638	0.471
	3	0.893	0.285	0.547	0.590
	4	0.910	0.262	0.391	0.635
	5	0.927	0.237	0.249	0.870
<i>GGI1, GGI6, JGI1, JGI2, JGI6</i>	3	0.956	0.181	0.390	0.801
	4	0.968	0.142	0.414	0.752
	5	0.968	0.142	0.565	0.621
	6	0.968	0.142	0.503	0.680

Optimum neural network predictors appear in bold letter.

Table 5. Statistics of the non-linear predictors of the IK_{Ca} channel-blocking activity of the clotrimazole analogs obtained by GNN approach.

Descriptors in BRNN models	Number of hidden nodes	Full data set	Data set without compound 10			
		R^2	R^2	RMSE	Q^2	RMSE _{cv}
<i>GGI1, GGI7, JGI2</i>	2	0.701	0.908	0.266	0.813	0.345
	3	0.704	0.912	0.259	0.595	0.482
	4	0.706	0.914	0.258	0.431	0.580
<i>GGI1, GGI7, JGI2, JGI3</i>	2	0.766	0.897	0.281	0.699	0.420
	3	0.763	0.970	0.151	0.901	0.265
	4	0.762	0.976	0.137	0.729	0.462
<i>GGI1, GGI3, GGI8, JGI2, JGI8</i>	5	0.762	0.976	0.137	0.710	0.481
	3	0.762	0.972	0.145	0.776	0.434
	4	0.760	0.976	0.135	0.778	0.402
	5	0.760	0.976	0.135	0.756	0.415
	6	0.760	0.976	0.135	0.744	0.424

Optimum neural network predictors appear in bold letter.

Regarding this, we tried to improve the poor performance of the selected subsets of descriptors by using such variables to train BRANNs with variable architectures. The statistics of these predictors, excluding the detected outlier, appear in Table 4. In this case, networks were able to achieve a remarkable better fit of the data ($R^2 > 0.9$) when comparing with the linear approach. However a remaining weakness of these models is its relative low predictive ability, illustrated by its relative low values of Q^2 of LOO cross-validation (< 0.7).

These results are expectable since several previous reports demonstrated that when ANNs were trained with variables selected by linear search routines, the networks largely overcame linear models by increasing data fitting but the predictors did not exhibit a remarkable improvement in predictive power [15, 17, 20, 46]. In such cases ANNs were able to learn the data very well but they were not stable enough predicting the activity of newer compounds presented to the networks. We can conclude that ANNs shown in Table 4 tended to overfit the data having so a limited

practical utility. Thus, a method that extracts relevant non-linear information from the data set is essential for having reliable neural network predictors.

GNNs simulations

Taking into account the highly non-linear dependence showed by the descriptors in the best linear models, we expected to find adequate combinations of three, four and five variables for reliably modeling of the *IKCa1* channel-blocking activity by means of the GNN approach. In spite of the unfit behavior of compound 10, inside the GNN framework networks were trained with full data set with the aim of looking for a non-linear model that fits well this clotrimazole analog. The implemented GA searches for the best fitted BRANN, in such a way that from one generation to another the algorithm tried to minimize the MSE of the networks (fitness function). By employing this approach instead a more complicated and time consuming cross-validation based fitness function we gain in CPU time and simplicity of the routine. Furthermore, we can devote our relatively small data set completely to train the networks. However, the use of the *MSE* fitness function could lead to undesirable well fitted but poor generalized networks as algorithm solutions. In this connection, we tried to avoid such results by two aspects: (1) keeping network architectures as simplest as possible (3-2-1, 4-2-1 and 5-3-1) inside the GA framework and (2) implementing Bayesian regulation in the network training function (Section 'Artificial neural networks').

Similarly to the linear model, removing compound 10 from the data set, due to its high residual value, remarkably improves the statistical parameters of the models. In addition to the best three BRANN predictors with 3-2-1, 4-2-1 and 5-3-1, architectures yielded by the GA routine, Table 5 shows the structures and statistics of several networks obtained by a screening process in which the number of hidden nodes in the initial predictors was increased with the aim of optimizing the reliability of the networks. It should be notice that parameters of data fitting (R^2 and RMSE) of all networks reported in Table 5 are in the same range that those showed in Table 4. However, it is noteworthy the improvement in the parameters of LOO cross-validation taking into account that

several predictors have Q^2 values higher than 0.7. Examining these results we found two outstanding predictors: the network with three inputs and two hidden nodes (BRANN 1) and the network with four inputs and three hidden nodes (BRANN 2) that exhibit Q^2 of cross-validation higher than 0.8. Those models also show the lower values of $RMSE_{cv}$.

Concerning the possibility of chance correlations, following the method used by So and Karplus [21], we performed a randomization test. Randomized values were given to the dependent variable (biological activities) and networks were trained using this randomized target and the real set of independent variables (molecular descriptors). By repeating this processes 500 times no correlation was found between Q^2 and R^2 values, similar to the results of So and Karplus [21].

Similarly as we did for the four descriptors giving the best four-variable MRA model, we studied the functional dependence between the *IKCa1* channel-blocking activity and the descriptors of the predictor BRANN 2. Figure 4b depicts the response curves obtained for *GGI1*, *GGI7*, *JGI2* and *JGI3*. As can be observed, when we linearly correlated such response curves (dotted lines) it was evidenced that the relationships between such descriptors and the modeled biological property is strictly non-linear as it is illustrated by the R^2 of 0.448, 0.340, 0.006 and 0.184 for *GGI1*, *GGI7*, *JGI2* and *JGI3* respectively. However, BRANNs were able to fit the convolute dependence between such descriptors and the *IKCa1* channel-blocking activities by means of a reliable non-linear relationship in such a way that an optimum adjustment of network weights and bias yielded a highly predictive network.

The plot of calculated vs. experimental *IKCa1* channel-blocking activity of the clotrimazole analogs using the predictor BRANN 2 is showed in Figure 5. As can be observed, the inclusion of compound 10 (marked with a circle) yielded a low fitted model (Figure 5a) but deleting this compound notably improves the statistical significance of the predictor (Figure 5b). These plots support considering compound 10 as an outlier in our models.

Furthermore, in order to settle the stability of our best predictor BRANN 2, we performed calculation of several randomly constructed training/test set partitions of 23 and 6 compounds respectively,

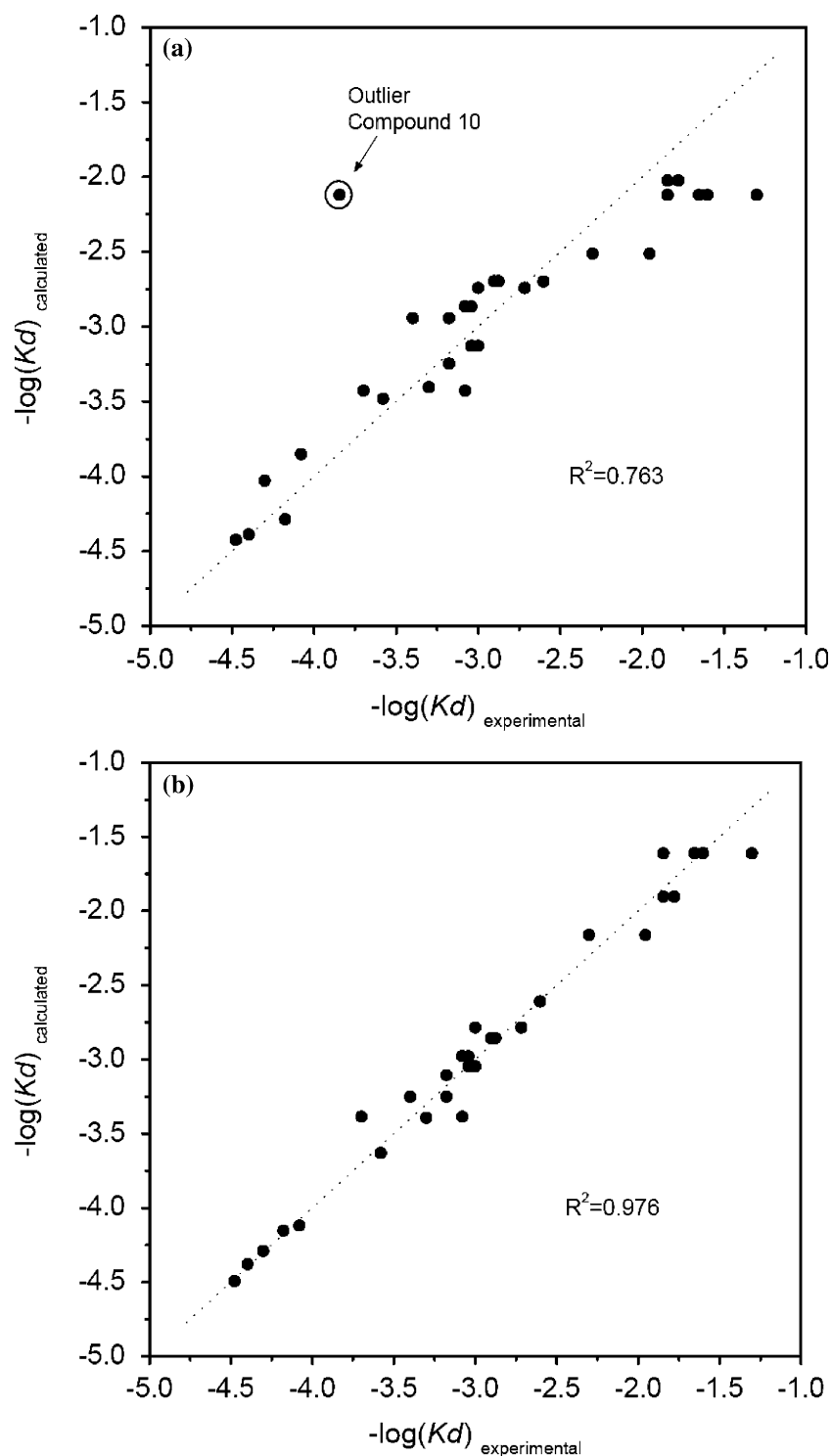


Figure 5. Plots of calculated vs. experimental IK_{Ca} channel-blocking activity of the clotrimazole analogs for data fitting with (a) and without (b) the outlier, compound 10. Compound 10 appears pointed with a black circle. The dotted lines are an ideal fit with the respective intercept and slope equal to zero and one.

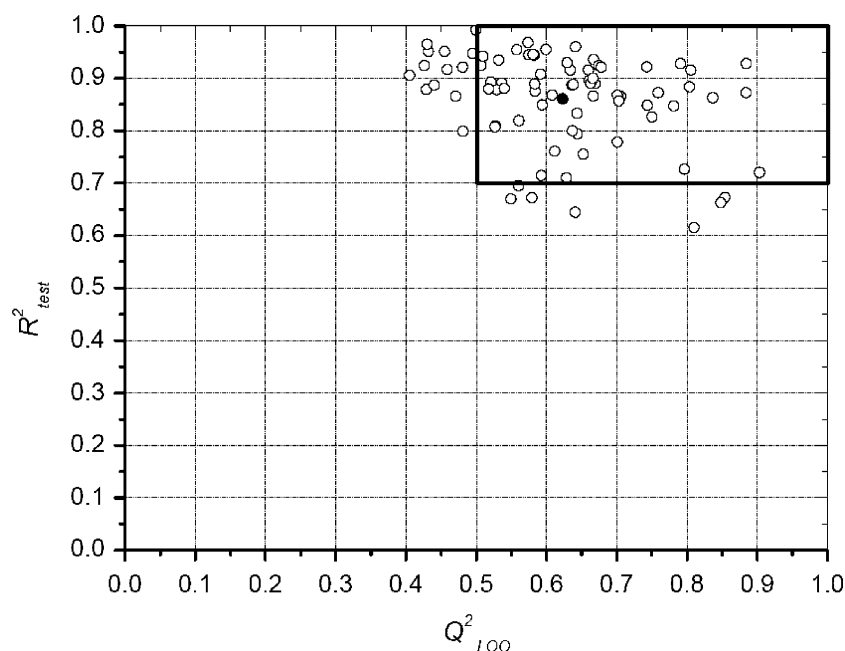


Figure 6. Plot of R^2 of test set (R^2_{test}) vs. Q^2 of LOO cross-validation of training set for 80 randomly constructed training/test sets partitions to settle best predictor stability. Black rectangle includes cases with R^2_{test} and Q^2 values over 0.7 and 0.5 respectively. Black dot represents mean values of $R^2_{\text{test}}=0.861$ and $Q^2=0.621$.

excluding the detected outlier. In this sense, Figure 6 depicts plots of R^2 of test set (R^2_{test}) vs. Q^2 of training set for 80 training/test set partitions made. As can be observed, the most of the cases appear at $R^2_{\text{test}} > 0.7$ and $Q^2 > 0.5$, specifically a 75% of the randomly constructed training and test sets (black rectangle in Figure 6). The mean values of $R^2_{\text{test}}=0.861$ and $Q^2=0.621$ for all the training/test set partitions, appear highlighted as a black dot in Figure 6. Regarding this, we consider our best model BRANN 2 reliable and stable enough, taking into account the criterion of Golbraikh and Tropsha [45] that a good QSAR model should have Q^2 values of internal LOO cross-validation over 0.5 and at the same time be able to predict external sets of compounds with high R^2 values.

Kohonen self-organizing map

With the aim of settling some structural features related to the activity of the studied compounds, variables in predictor BRANN 2 were used to obtain a KSOM of the triarylmethanes. Figure 7 depicts the 7×7 KSOM of the data, 19 of a total of 49 neurons were occupied. The majority of the neurons were occupied by two compounds at the

same time. Compounds with a similar range of activity were grouped into neighboring areas. The less active compounds were placed at the upper-left and bottom-right regions in the map. In turn, the most active compounds were located at the diagonal from the bottom-left to the upper-right regions of the map. Interestingly, compound 10 mismatched again to the rest of the data and this low active compound was misplaced in the same highly active neuron among the four most blocking compounds 1 (clotrimazole), 18 (TRAM-34), 19 and 30. Regarding this, the outlier behavior of this compound, previously observed in the quantitative linear and non-linear models, was corroborated by the KSOM.

Model's interpretation

In some previous QSAR reports, ion channel inhibitory activity had been successfully modeled by means of ANN approach [7, 23, 24]. Calcium channel antagonist activity of a large set of 1,4 dihydropyridines was modeled using a hybrid principal components (PC)-Neural Network approach, in which, among other methods, they used also a GA routine for selecting the best subset of

components for network training [23]. In spite of the differences between the type of ion channel studied in this work and the report of Hemmateenejad et al. [23], some comparison concerning the computational modeling can be addressed. Those authors achieved good results reporting R^2 for training, validation and test sets ranging from 0.90 to 0.94, 0.88 to 0.93 and 0.8 to 0.90 respectively. However, the best reported models included PCs that extracted information from different kinds of molecular descriptors thus even a shallow interpretation of the models is meaningless. Furthermore, a structure–activity relationship on the same type of calcium channel was reported by Takanaoka et al. [24]. Another set of 1,4 dihydropyridines was used for developing activity classification models by PC analysis (PCA) and ANNs. In this case the authors aimed to classify the inhibitors in three levels of antagonist activity by using a set of topological and quantum-chemical descriptors. However, their good results pointed out that the performance of the ANNs on the classification was equivalent to the classical PCA method.

In our approach, remarkable good results were obtained for modeling the IK_{Ca1} channel activity of some clotrimazole analogs. In addition to the goodness of the overall fitting of the data, our best

model exhibited a high stability in the calculation of several randomly constructed training/test set partitions (Figure 6). It is also noteworthy the accuracy of the predictor BRANN 2 to calculate the blocking activities of the five lowest active compounds (Figure 5b). Although the activity of the highest active inhibitor, known as TRAM–34 (compound 18), was predicted with one of the highest residual equals to -0.310 (Table 2). The predictor BRANN 2 yielded for this compound a $-\log(K_d)$ value of -1.611 representing $K_d = 41$ nM when the experimental value is 20 nM. However, a large error of 53 nM was also reported for this compound so the calculated value is thoroughly inside the experimental interval measured by Wulff et al. [6] in the original report.

At the light of the results of the KSOM map in Figure 7, some structural analysis can be addressed. Wulff et al. [6] pointed out that a halogen atom in orto position on a phenyl ring is the main feature for displaying a high IK_{Ca} blocking activity. Nevertheless, they also suggest that the occurrence of an imidazole group or a group of similar size, lipophilicity, and π -electron density, such as acetonitrile, pyrazole and tetrazole, are needed too for displaying a high activity. In the KSOM the low active compound 10 was misplaced in the same neuron that it is shared by the most active

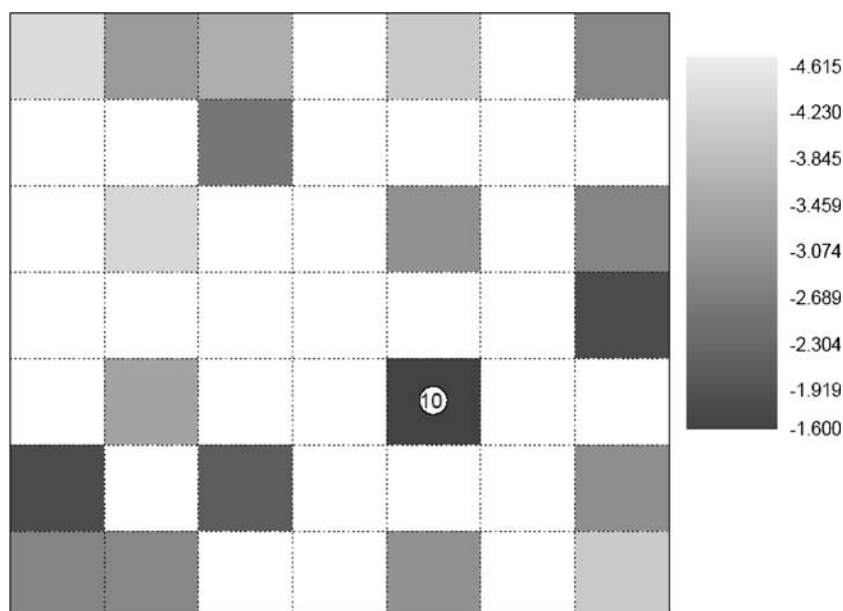


Figure 7. KSOM of the IK_{Ca} channel-blocking activity of the clotrimazole analogs. Number represents the outlier, compound 10, misplaced in the most active neuron. Activity legend is placed at the right-hand of the map.

compounds clotrimazole, TRAM-34, compounds 19 and 30. Although the 100-fold decrease in the reported blocking activity for compound 10 in comparison to clotrimazole, TRAM-34, compounds 19 and 30, is really astounding. Interestingly, the main difference between the outlier and such compounds is they present more than one nitrogen atom in the five-member heterocyclic ring (imidazole, pyrazole and tetrazole) (Table 2) whilst the outlier only had one nitrogen in such ring (pyrrole). Concerning that, we consider the blocker activity of this compound probably limited by some conformational effects than cannot be contained in the information gathered by the TCI descriptors we used.

On the other hand, compounds without a three phenyl basic structure (compounds 4 and 28) as well as compounds with an halogen in orto position but with bulky groups substituted on the phenyl rings (compounds 3, 22–27), were well located on less active neurons on the topological map. These results supported the hypothesis of Wulff et al. that such blockers are less effective since they lack of the adequate size for matching properly in IK_{Ca} channel.

Interpreting an ANN based QSAR model in terms of the specific contribution of substituents and other molecular features to the modeled activity is always a difficult task. TCI descriptors contain important information on the relationship between the compound structures and its activities by describing the molecular topology and the charge transfer through the molecule [9, 10]. Our data set do not vary appreciably in size and/or shape, mainly relative positions and substitutions of small groups change around the thiarylmethane structure. Then, electrotopological conformation is the main feature that it is varying from one compound to another. That is why, similarly to a previous report of our group, TCI-trained ANNs were able to explain approximately a 97% of the data variance [15]. The participation in our best model of descriptors of lag three and seven may be viewed in terms of association of blocking activity information content with structural fragments of such size. However, further deciphering of the information content of these descriptors is very complex as their computations involve integration of the structural fragments and due to this it is not possible to traverse backward from a higher state to a lower one [15].

Concluding remarks

Since biological phenomena are often complex by nature, in this work the inhibition of intermediate-conductance Ca^{2+} -activated K^{+} channel by a set of triarylmethanes was successfully modeled using a hybrid approach that combines GA and ANNs. TCI demonstrated to encode relevant structural information from the studied compounds that highly correlate in a non-linear way with its IK_{Ca} channel-blocking activity.

The implementation of Bayesian regularization on the network training function allowed obtaining robust models with optimum generalization qualities. GNN approach yielded a highly reliable predictor for the intermediate-conductance Ca^{2+} -activated K^{+} channel blocker activity of some triarylmethanes. Furthermore, the occurrence in the best predictor of charge indexes of lag three and seven suggests a high influence of molecular fragments of such sizes on the IK_{Ca} channel-blocking activity. Our results confirmed the usefulness of GA for feature extraction on QSAR studies and ANNs for optimum modeling of complex biological activities. Selected TCI descriptors were also enable to well distribute data set on a KSOM.

References

1. Gárdos, G., *Biophys. Acta.*, 30 (1958) 653.
2. Cook, N.S. and Quast, U., In *Potassium Channels*, Cook, N.S., Chichester, p. 70.
3. Haylett, D.G. and Jenkinson, D.H., In *Potassium Channels*, Cook, N.S., Chichester, 1990, p. 70.
4. Castle, N.A., *Perspect. Drug Discovery Des.*, 15 (1999) 131.
5. Vergara, C., LaTorre, R., Marrion, N.V. and Adelman, J.P., *Curr. Opin. Neurobiol.*, 8 (1998) 321.
6. Wulff, H., Miller, M.J., Hänsel, W., Grissmer, S., Cahalan, M.D. and Chandy, K.G., *Proc. Natl. Acad. Sci.*, 97 (2000) 8151.
7. Roxburgh, C.J., Ganellin, C.R., Athmani, S., Bisi, A., Quaglia, W., Benton, D.C.H., Shiner, M.A.R., Malik-Hall, M., Haylett, D.G. and Jenkinson, D.H., *J. Med. Chem.*, 44 (2001) 3244.
8. Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*. VCH, New York, 1993.
9. Gálvez, J., Garcia, R., Salabert, M.T. and Soler, R., *J. Chem. Inf. Comput. Sci.*, 34 (1994) 520.
10. Gálvez, J., Garcia-Domenech, R., de Julián-Ortiz, J.V. and Soler, R., *J. Chem. Inf. Comput. Sci.*, 35 (1995) 272.
11. Kios-Santamarina, I., Garcia-Domenech, R. and Gálvez, J., *Bioorg. Med. Chem. Lett.*, 18 (1998) 477.
12. Calabuig, C., Antón-Fos, G.M., Gálvez, J. and García-Domenech, R., *Int. J. Pharm.*, 278 (2004) 111.

13. González, M.P. and Terán, C., *Bull. Math. Biol.*, 66 (2004) 907.
14. González, M.P. and Terán, C., *Bioorg. Med. Chem. Lett.*, 14 (2004) 3077.
15. Fernández, M., Caballero, J., Helguera, A.M., Castro, E.A. and González, M.P., *Bioorg. Med. Chem.*, 13 (2005) 3269.
16. Fernández, M., Tundidor-Camba, A. and Caballero, J., *Mol. Simulat.*, 31 (2005) 575.
17. González, M.P., Caballero, J., Garriga, M., González, G., Helguera, A.M. and Fernández, M., *Bull. Math. Biol.* (2005) (in press).
18. González, M.P., Caballero, J., Tundidor-Camba, A., Helguera, A.M., Fernández, M., *Bioorg. Med. Chem.* DOI: 10.1016/j.bmc.2005.08.009.
19. Fernández, M. and Caballero, J., *Bioorg. Med. Chem.* DOI: 10.1016/j.bmc.2005.08.022.
20. Caballero, J. and Fernández, M., *J. Mol. Model.* DOI: 10.1007/s00894-005-0014-x.
21. So, S.S. and Karplus, M., *J. Med. Chem.*, 39 (1996) 1521.
22. So, S.S. and Karplus, M., *J. Med. Chem.*, 39 (1996) 5246.
23. Hemmateenejad, B., Akhond, M., Miri, R. and Shamsipur, M., *J. Chem. Inf. Comput. Sci.*, 43 (2003) 1328.
24. Takahata, Y., Costa, M.C.A. and Gaudio, A.C., *J. Chem. Inf. Comput. Sci.*, 43 (2003) 540.
25. Hemmateenejad, B., Safarpour, M.A., Miri, R. and Nesari, N., *J. Chem. Inf. Model.*, 45 (2005) 190.
26. Hawkins, D.M., *J. Chem. Inf. Comput. Sci.*, 44 (2004) 44.
27. Kier, L.B. and Hall, L.H. *Molecular Structure Descriptors: The Electrotopological State*. Academic Press, New York, 1999 .
28. Hall, L.H., Mohny, B. and Kier, L.B., *J. Chem. Inf. Comput. Sci.*, 31 (1991) 76.
29. Girones, X., Amat, L., Robert, D. and Carbo-Dorca, R., *J. Comput. Aided Mol. Des.*, 14 (2000) 477.
30. Todeschini, R. and Consonni, V. and Pavan, M., *DRA-GON*. version 2.1 (2003).
31. Holland, J.H. *Adaption in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, MI, 1975.
32. Cartwright, H.M. *Applications of Artificial Intelligence in Chemistry*. Oxford University Press, Oxford, 1993.
33. The MathWorks Inc. *MATLAB* version 7.0. (2004), www.mathworks.com.
34. The MathWorks Inc., *Genetic Algorithm and Direct Search Toolbox User's Guide for Use with MATLAB*. The Mathworks Inc, Massachusetts, 2004.
35. Hertz, J., Krogh, A. and Palmer, R.G. *Introduction to the Theory of Neural Computation*. Addison-Wesley Publishing Co, Redwood City, CA, 1991.
36. Kolmogorov, A.N., *SSSR*, 114 (1957) 953.
37. The MathWorks Inc., *Neural Network Toolbox User's Guide for Use with MATLAB*. The Mathworks Inc, Massachusetts, 2004.
38. Mackay, D.J.C., *Neural Comput*, 4 (1992) 415.
39. Burden, F.R. and Winkler, D.A., *J. Med. Chem.*, 42 (1999) 3183.
40. Burden, F.R. and Winkler, D.A., *Chem. Res. Toxicol.*, 13 (2000) 436.
41. Winkler, D.A. and Burden, F.R., *Biosilico*, 2 (2004) 104.
42. Polley, M.J., Winkler, D.A. and Burden, F.R., *J. Med. Chem.*, 47 (2004) 6230.
43. Kohonen, T., *Biol. Cybern.*, 43 (1982) 59.
44. Gasteiger, J. and Zupan, J., *Angew. Chem. Int. Ed. Engl.*, 32 (1995) 503.
45. Golbraikh, A. and Tropsha, A., *J. Mol. Graph. Model.*, 20 (2002) 269.
46. Zahouily, M., Bazoui, A.R., Sebt, S. and Zakarya, D., *J. Mol. Model.*, 8 (2002) 168.