

This article was downloaded by:

On: 14 January 2011

Access details: *Access Details: Free Access*

Publisher *Taylor & Francis*

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Molecular Simulation

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t713644482>

2D Autocorrelation modeling of the activity of trihalobenzocycloheptapyridine analogues as farnesyl protein transferase inhibitors

M. Fernández^{ab}; A. Tundidor-Camba^c; J. M. Caballero^{ab}

^a Molecular Modeling Group, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba ^b Probiotic Group, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba ^c Scientific Prospection Group, National Center for Scientific Researches (CNIC), Havana, Cuba

To cite this Article Fernández, M. , Tundidor-Camba, A. and Caballero, J. M.(2005) '2D Autocorrelation modeling of the activity of trihalobenzocycloheptapyridine analogues as farnesyl protein transferase inhibitors', *Molecular Simulation*, 31: 8, 575 — 584

To link to this Article: DOI: 10.1080/08927020500134144

URL: <http://dx.doi.org/10.1080/08927020500134144>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

2D Autocorrelation modeling of the activity of trihalobenzocycloheptapyridine analogues as farnesyl protein transferase inhibitors

M. FERNÁNDEZ^{†‡}, A. TUNDIDOR-CAMBA[¶] and J. M. CABALLERO^{†‡*}

[†]Molecular Modeling Group, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba

[‡]Probiotic Group, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba

[¶]Scientific Prospection Group, National Center for Scientific Researches (CNIC), P.O. Box 6880, Havana, Cuba

(Received February 2005; in final form February 2005)

The inhibitory activity towards farnesyl protein transferase enzyme (FPT) of 49 piperidine substituted trihalobenzocycloheptapyridine analogues (thBCHPs) has been successfully modelled using 2D spatial autocorrelation vectors. Predictive linear and non-linear models were obtained by forward stepwise multilinear regression analysis (MRA) and artificial neural network (ANN) approaches, respectively. A variable selection routine that selected relevant non-linear information from the data set was employed prior to networks training. The MRA model, using three descriptors, was able to explain about 68% data variance. The model showed a linear dependence between the inhibitory activities and autocorrelation coefficients weighted by van der Waals volumes and atomic polarizabilities on the inhibitors molecules. The non-linear approach preserve several characteristics described for the linear one. Three descriptors were selected encoding the same atomic properties, but the new ones were able to explain about 92% data variance. In addition, the ANN model had higher predictive power. Furthermore, inhibitors were well distributed regarding its activity levels in a Kohonen self-organizing map (SOM) built using the input variables of the best neural network.

Keywords: QSAR analysis; Feed-forward neural network analysis; Self organizing maps; Farnesyl protein transferase inhibitors

1 Introduction

The development of cancer therapeutics in the last few years has taken new dimensions since modern biological techniques open the way to understand key cellular processes at the individual protein level [1]. Growth factor signalling pathways are among the first systems to have been investigated successfully. In this sense, the potential for therapeutic intervention at several different levels between the cell membrane and the nucleus has become evident and a variety of molecular biological and “small molecule” tools are under investigation.

One of the key signalling system is the Ras pathway. Mutated forms of Ras, which are constitutively active, are found in approximately 30% of all cancers in man. Ras proteins play a central role in the signal transduction cascades controlling these processes. Most interest in modifying the action of the Ras oncogene has been focused on inhibition of the farnesyl protein transferase enzyme

(FPT) [2]. In order to exert its functional effects, Ras has to be docked into the cell membrane. The cytosolic protein has to be modified at the C-terminus by addition of a lipophilic “tail” (farnesyl pyro phosphate: FPP) which then anchors it into the cell membrane. FPT recognizes and binds only the last four C-terminal amino acids of the CAAX-consensus sequence (C, cysteine; A, aliphatic amino acid; X, serine or methionine) of its substrate proteins; this tetrapeptide is therefore a primary template for the development of non-peptide farnesyltransferase inhibitors [3].

Early attempts to discover inhibitors of FPT focused on modifications of the isoprenoid and CAAX polypeptide substrates of the enzyme. While potent FPP-derived inhibitors have been discovered, most attention has been paid to analogues of the CAAX peptide. They have been considered peptides based on the CAAX sequence containing a free thiol group [4,5]. Peptidomimetics FPT inhibitors that are non-thiol peptides have also been reported [6]. However, the peptidic nature or the presence of a free thiol

*Corresponding author. Tel: +53-45-26-1251. Fax: +53-45-25-3101. Email: . E-mail: julio.caballero@umcc.cu; . E-mail: jmcr77@yahoo.com

group in these FPT inhibitors may have disadvantages in the development of such compounds as therapeutic agents. In recent years the number of non-peptidic, non-thiol-containing selective FPT inhibitors is increased [7–10].

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 modelling biological activities of natural and synthetic chemicals [11]. Graph-theoretical and topological methods are included in the most QSAR studies. Among these methods, 2D spatial autocorrelations has been successfully used for modelling logP-values [12], biological activities [13], for pharmaceutical [14] and toxicological research [15].

Only a few papers have reported QSAR studies on FPT inhibitors. Giraud *et al.* [16] applied a system of multiple techniques coupled with handpicked thermodynamic, physical and topological descriptors to build a discriminatory model that could separate actives from inactives over highly analogous compounds. On the other hand, Estrada *et al.* [17] used a diverse training set of anticancer compounds to compute a fragment-based QSAR model that could be subsequently employed to compute the probability of a particular compound being active against cancer. Polley *et al.* [18] employed Bayesian regularized ANN to generate a QSAR model on a large FPT inhibitors data set using molecular descriptors.

In this work autocorrelation vectors were used for encoding structural information from piperidine substituted trihalobenzocycloheptapyridine analogues (thBCHPs), and linear and non-linear models of the FPT inhibitory activity were built using multivariate-linear regression analysis (MRA) and ANNs. A comparative study was developed according to the results of data fitting and the predictive power of the models measured by cross-validation technique. The versatility of ANNs was used also for mapping the thBCHPs inhibitory activities on a topological map using competitive neural networks.

2 Method

2.1 Spatial autocorrelation 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 [19].

Autocorrelation vectors have several useful properties. First, a substantial reduction in data can be achieved by limiting the topological distance, l . Second, the autocorrelation coefficients are independent of the original atom numberings, so they are canonical. And thirdly, the length of the correlation vector is independent of the size of the molecule [19].

For the autocorrelation vectors, H-depleted molecular structure is represented as a graph G and physico-chemical properties of atoms as real values assigned to the vertices of G (table 1). These descriptors can be obtained by summing up the products of certain properties of two atoms, located at given topological distances or spatial lag in G . Three spatial autocorrelation vectors were employed for modelling the inhibitory activity:

Moran's index [20]:

$$I(p_k, l) = \frac{N}{2L} \frac{\sum_{ij} \delta_{ij}(p_{ki} - \bar{p}_k)(p_{kj} - \bar{p}_k)}{\sum_i (p_{ki} - \bar{p}_k)} \quad (1)$$

Geary's coefficient [21]:

$$c(p_k, l) = \frac{(N-1)}{4L} \frac{\sum_{ij} \delta_{ij}(p_{ki} - \bar{p}_k)(p_{kj} - \bar{p}_k)}{\sum_i (p_{ki} - \bar{p}_k)} \quad (2)$$



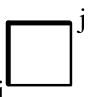
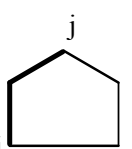
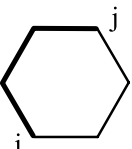

Broto-Moreau's autocorrelation coefficient [22]:

$$A(p_k, l) = \sum_i \delta_{ij} p_{ki} p_{kj} \quad (3)$$

where $I(p_k, l)$, $c(p_k, l)$ and $A(p_k, l)$ are Moran's index, Geary's coefficient and Broto-Moreau's autocorrelation coefficient at spatial lag l , respectively; p_{ki} and p_{kj} are the values of property k of atom i and j , respectively; \bar{p}_k is the average value of property k and $\delta(l, d_{ij})$ is a Dirac-delta function defined as

$$\delta(l, d_{ij}) = \begin{cases} 1 & \text{if } d_{ij} = l \\ 0 & \text{if } d_{ij} \neq l \end{cases} \quad (4)$$

Table 1. Representation of different molecular graphs G and topological distances or spatial lags d_{ij} .

Molecular Graphs G						
d_{ij}	1	1	2	2	3	4

where d_{ij} is the topological distance or spatial lag between atoms i and j .

Spatial autocorrelation measures the level of interdependence between properties, and the nature and strength of that interdependence. It may be classified as either positive or negative. In a positive case all similar values appear together, while a negative spatial autocorrelation has dissimilar values appearing in close association [20,21]. In a molecule, Moran's and Geary's spatial autocorrelation analysis tests whether the value of an atomic property at one atom in the molecular structure is independent of the values of the property at neighbouring atoms. If dependence exists, the property is said to exhibit spatial autocorrelation. Moreau and Broto first applied autocorrelation function to the topology of molecular structures [22]. The autocorrelation vectors represent the degree of similarity between molecules.

A data matrix is generated with the spatial autocorrelation vectors calculated for each compound. Afterwards, dimensionality reduction methods were employed for selecting the most relevant vector components for building linear and neural network models.

2.2 Data sets and models

Inhibitory activities against FPT enzyme and molecular structures of 30 thBCHPs were taken from the literature [23]. Compounds activities were reported as their ability to inhibit the transfer of [^3H]-farnesyl from farnesyl diphosphate to H-Ras-CVLS, a process that is mediated by FPT. IC_{50} refers to the nanomolar concentration of the compound required for 50% inhibition of the enzyme activity [23]. Molecular structure and biological activities are summarized in table 2. Activities reported like lower or higher than threshold values were taken equal to the threshold value.

Prior to molecular descriptor calculations, 3D structures of the studied compounds were geometrically optimized using semi-empirical quantum-chemical method PM3 [24] implemented in MOPAC 6.0 [25] computer software.

Dragon [26] computer software was used for calculating unweighted and weighted Moran, Geary and Broto-Moreau 2D-autocorrelation vectors. As weighting properties we tried atomic masses, atomic Van der Waals volumes, atomic Sanderson electronegativities and atomic polarizabilities. Autocorrelation vectors were calculated at spatial lags l ranging from 1 up to 8.

The total number of computed descriptors was 96. Descriptors with constant values were discarded. For the remaining descriptors pairwise correlation analysis was performed in order to reduce, in a first step, the collinearity and correlation between descriptors. The procedure consists of the elimination of the descriptor with lower variance from each pair of descriptors with the modulus of the pair correlation coefficients higher than a predefined value R_{\max} (0.90). Afterwards, the number of remained descriptors was 37.

2.3 Forward stepwise multilinear regression analysis

The most significant parameters for the multilinear regression analysis (MRA) model were identified from the data set using forward stepwise regression method [27], where the independent variables are individually added or deleted from the model at each step of the regression depending on the Fisher ratio values selected to enter and to remove until the "best" model is obtained. Statistical analysis and data exploration was carrying out using the Statistica version 6.0 [28] computer software. Examining the regression coefficients, the standard deviations, the significances and the number of variables in the equation determined the quality of the model.

2.4 Feed-forward neural network approach

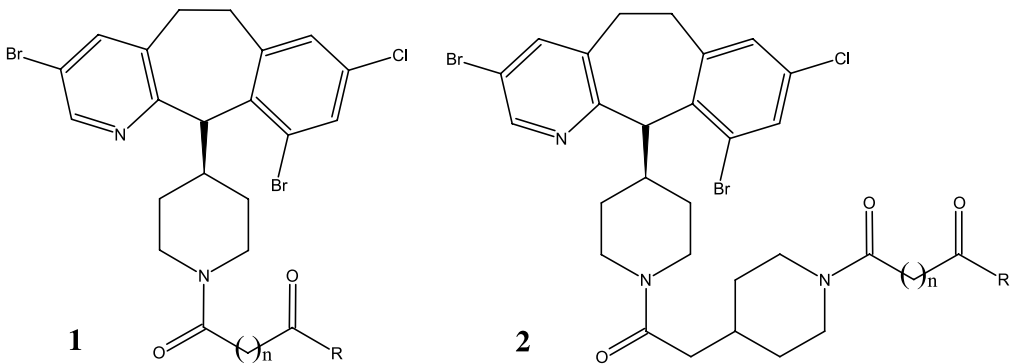
ANNs are computer-based models in which a number of processing elements, also called neurons, units, or nodes are interconnected by links in a netlike structure forming "layers" [29,30]. A variable value is assigned to every neuron. The neurons can be one of three different kinds. The input neurons receive their values from independent variables, 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.

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 error. 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 [31].

Choosing the adequate descriptors for non-linear QSAR studies is difficult because there are no absolute rules that govern this choice. Recently, evolutionary algorithms and specifically genetic algorithms have been used for variable selection problems combined to ANNs [32]. In this work, a neural network feature selection procedure that extracts non-linear information from the data set was employed for data dimensionality reduction before network training. In this regard, neuro-genetic input selection routine (NGISR) of Statistica Neural Networks package from Statistica 6.0 computer software was used. This tool

Table 2. Structures of trihalobenzocycloheptapyridine analogues and experimental and predicted inhibitory activities by forward stepwise multilinear regression analysis and artificial neural network models against farnesyl protein transferase.



Compound	n	R	log(10 ⁶ /IC ₅₀)		
			Exp	MRA	ANN
1a	1	ONa	3.92	4.06	3.94
1b	1	NH ₂	4.17	4.55	4.13
1c	2	OCH ₃	5.04	4.31	5.03
1d	2	ONa	3.57	4.17	3.51
1e	2	NH ₂	4.85	4.69	4.81
1f	3	ONa	4.36	4.47	4.63
1g	3	NH ₂	4.96	4.86	4.79
1h	4	OCH ₃	3.82	4.66	3.81
1i	4	ONa	4.77	4.62	4.91
1j	4	NH ₂	5.11	4.98	5.08
1k	5	OCH ₃	4.96	4.75	4.72
1l	5	ONa	5.04	4.74	5.20
1m	5	NH ₂	5.48	5.07	5.24
2a	1	OEt	5.52	5.25	5.31
2b	1	NH ₂	5.48	5.51	5.50
2c	2	OCH ₃	5.29	5.22	5.39
2d	2	ONa	5.92	5.25	5.48
2e	2	NH ₂	5.60	5.54	5.68
2f	3	OEt	5.32	5.25	5.42
2g	3	ONa	5.96	5.41	5.52
2h	3	NH ₂	5.41	5.63	5.48
2i	4	OCH ₃	5.28	5.41	5.54
2j	4	ONa	5.44	5.45	5.58
2k	4	NH ₂	5.46	5.68	5.53
2l	5	OCH ₃	5.16	5.44	5.49
2m	5	ONa	5.85	5.50	5.64
2n	5	NH ₂	5.48	5.70	5.62

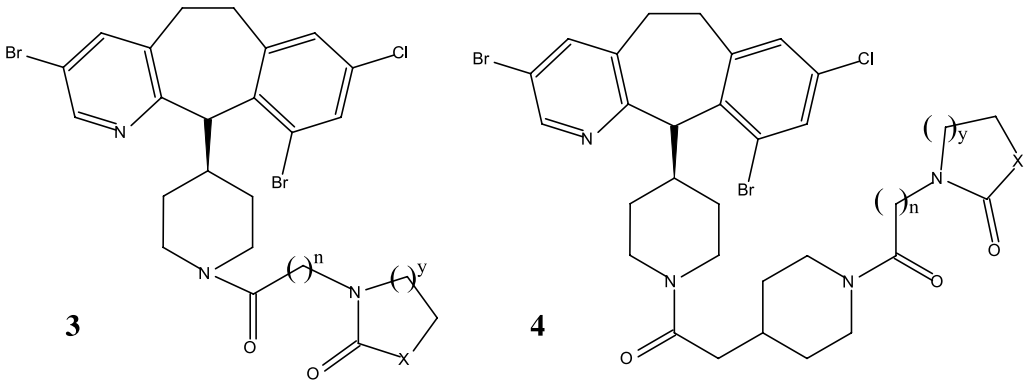


Table 2 (Continued)

Compound	<i>n</i>	<i>y</i>	<i>X</i>	$\log(10^6/IC_{50})$		
				Exp	MRA	ANN
3a	1	1	CH ₂	5.40	5.15	5.49
3b	2	1	CH ₂	4.66	4.66	4.74
3c	3	1	CH ₂	4.07	4.53	4.12
3d	1	2	CH ₂	5.52	5.34	5.44
3e	2	2	CH ₂	4.72	4.87	4.69
3f	3	2	CH ₂	4.10	4.71	4.13
3g	1	1	NH	4.64	4.88	4.68
3h	2	1	NH	4.72	4.51	4.73
3i	3	1	NH	4.80	4.39	4.81
3j	1	2	NH	5.52	4.97	5.35
3k	2	2	NH	4.70	4.61	4.80
3l	3	2	NH	4.25	4.48	4.21
4a	1	1	CH ₂	5.49	5.90	5.59
4b	2	1	CH ₂	5.52	5.47	5.46
4c	3	1	CH ₂	5.38	5.30	5.23
4d	1	2	CH ₂	5.85	6.00	5.81
4e	2	2	CH ₂	5.44	5.58	5.53
4f	3	2	CH ₂	5.28	5.40	5.19
4g	2	1	NH	5.13	5.36	5.30
4h	1	2	NH	5.60	5.72	5.61
4i	2	2	NH	5.57	5.39	5.46
4j	3	2	NH	5.09	5.23	5.25

combines the algorithms of genetic algorithms, probabilistic and generalized regression neural networks to automatically search for optimal combinations of input variables [33].

Feed-forward networks had 3 and 1 neurons in the input and output layers corresponding to independent and dependent variables, respectively. The architecture was optimized varying the number of neurons in the hidden layer. Training functions updated weights and bias values according to gradient descent momentum and an adaptive learning rate. Network training function parameters were optimized by varying both learning rate and momentum from 0.01 to 0.99.

Matlab version 6.5 [34] was used for implementing fully connected, three-layer, feed-forward computational neural networks with back-propagation training. In these nets, the transfer function of input and output layers was linear, and the hidden layer had neurons with a hyperbolic tangent transfer function. Network training was stopped when the minimum gradient of 0.001 was reached and then adjusted network weight and bias were stored.

2.5 Self-organizing maps

In order to settle structural similarities among the thBCHPs, a Kohonen self organizing map (SOM) was built. The autocorrelation descriptors selected by NGISR were used for unsupervised training of 9×9 neuron map. Kohonen [35] introduced a neural network model that generates a SOM. 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 [36]. Such maps of surface properties have been used for comparing wide variety of biologically active compounds [37].

$$out_{c_s} \leftarrow \min \left[\sum_{i=1}^m (X_{si} - w_{ij})^2 \right] \quad (5)$$

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_{ij} , most similar to the input variables (equation 5). 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, c_s . Finally, a molecule is projected into that neuron of the network with weights that come closest to the description of the molecule by the autocorrelation vector.

It should be noticed that the criterion embedded in equation 5 for determining the winning neuron for a molecule basically constitutes the measure determining the similarity of molecular structures. Molecules with similar autocorrelation vectors, X_s , are projected into the same or closely adjacent neurons. SOM were implemented in Matlab 6.5, neurons were initially located at a grid topology. The ordering phase was developed in 1000 steps with 0.9 learning rate until tuning neighbourhood 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.

2.6 Model validation

Models were validated by calculating Q^2 values. The Q^2 values are calculated from “leave-one-out” (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 .

$$Q^2 = 1 - \frac{\sum_{i=1}^N (Y_i - A_i)^2}{\sum_{i=1}^N (Y_i - \bar{A}_i)^2} \quad (6)$$

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 .

The Q^2 values can be considered a measure of the predictive power of a model: Whereas R^2 can always be increased artificially by adding more parameters (descriptors or neurons), Q^2 decreases if a model is over-parameterized [38], and is therefore a more meaningful summary statistic for predictive models.

3 Results and discussion

3.1 Multilinear regression analysis

2D autocorrelation descriptors were used for obtaining, in a first approach, a MRA model for the inhibitory activities of thBCHPs against FPT enzyme with acceptable statistic significance and predictive power (equation 7). Following the principle of parsimony [38] we choose a three variable model as the “best” model.

$$\begin{aligned} \log(10^6/\text{IC}_{50}) = & 8.779 \times \text{MATS1v} - 12.126 \\ & \times \text{GATS5v} - 4.686 \times \text{GATS6p} \\ & + 21.963 \end{aligned} \quad (7)$$

$$N = 49 \quad R = 0.828 \quad S = 0.335 \quad F = 32.735$$

$$p < 10^{-5} \quad Q^2 = 0.629 \quad S_{cv} = 0.364$$

In equation 7, N is the number of compounds included in the model, R is the correlation coefficient, S is the standard deviation of the regression, F is the Fisher ratio, Q^2 is the correlation coefficient of the cross-validation, p is the significance of the variables in the model and S_{cv} is the standard deviation of the cross-validation.

Inhibitory activities of the thBCHPs predicted by the linear model appear in table 2. The plot of experimental $\log(10^6/\text{IC}_{50})$ versus calculated $\log(10^6/\text{IC}_{50})$ is given in figure 1A. This model is able to explain about 68% data variance and more important it is quite stable to the

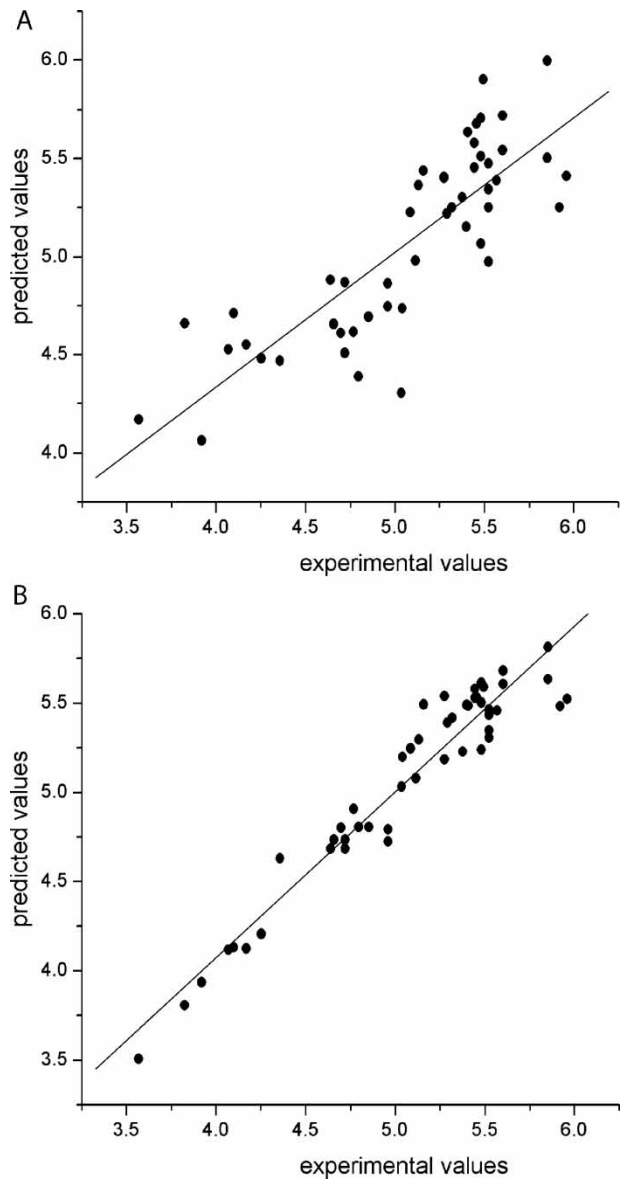


Figure 1. Experimental and predicted values from MRA (A) and ANN (B).

Table 3. Symbols of the descriptors selected by forward stepwise multilinear regression analysis and neuro-genetic input selection routine and their definitions.

Variable*	Forward stepwise multilinear regression analysis
MATS1v	Moran autocorrelation of lag 1/weighted by atomic van der Waals volumes
GATS5v	Geary autocorrelation of lag 5/weighted by atomic van der Waals volumes
GATS6p	Geary autocorrelation of lag 6/weighted by atomic polarizabilities
Variable*	Neuro-genetic input selection routine
GATS2v	Geary autocorrelation of lag 2/weighted by atomic van der Waals volumes
MATS7v	Moran autocorrelation of lag 7/weighted by atomic van der Waals volumes
MATS5p	Moran autocorrelation of lag 5/weighted by atomic polarizabilities

*The definition of the terms appears largely explained in reference 39.

inclusion–exclusion of compounds as measured by the correlation coefficient ($Q^2 > 0.5$). Variables in the model correspond to Moran's and Geary's spatial autocorrelation coefficients weighted by atomic van der Waals volumes and atomic polarizabilities [39] (table 3). These autocorrelation descriptors represent the degree of similarity between inhibitor molecules based on such properties at spatial lags 1, 5 and 6, respectively.

3.2 Feed-forward neural network approach

Since biological interactions are non-linear by nature; the main goal of this work was to train ANNs for modelling the inhibitory activities against FPT enzyme of thBCHPs. Choosing the optimum architecture for networks is always a difficult task, in our work we followed the criterion that $1.80 < \rho < 2.2$ [$\rho = (\text{number of data points in the training set})/(\text{number of adjustable weights and bias controlled by the network})$] [40]. In this sense, the networks architecture was fixed in 3-5-1 ($\rho = 1.88$). Network inputs and outputs were normalized before training processes. For learning rate and momentum were assigned 0.9 and 0.02 values, respectively. Training was carried out for 2000 epochs, the global minimum was selected among ten replicas in all cases.

The selection of the optimum variable subset among a large number of descriptors for fitting a model is a key question in modelling processes. A lot of reports described the use of MRA for dimensionality reduction in NN modelling [41]. Recently, several novel approaches that attempt to select variables that gather non-linear information have been published. The most of these methods combine genetics algorithm and different ANN approaches [42,43]. In this work, we used the NGISR approach implemented in Statistica Neural Networks package (see "Materials and Methods" section) for selecting a second variable subset enable of retaining non-linear information. This method is a feature selection routine based on a neuro-genetic algorithm which reduces data dimensionality by removing redundant information.

Applying the above mentioned method the data was reduced to 13 descriptors. All possible combinations of three variables, within this reduced data set, were tested for training feed-forward ANNs, using the 3-5-1 architecture, the best 10 models were selected considering R . Afterwards, the best model was selected applying LOO cross-validation, taking into account the Q^2 value. Finally, a subset of three descriptors was achieved. The descriptors and their definitions are in table 3. The non-linear model (ANN) was generated using this subset. Cross-correlation analysis showed that all pairwise correlations were ≤ 0.467 between this variables, also indicating a low colinearity (see table 4).

The five hidden neurons were chosen to maintain ρ [40] between 1.8 and 2.2. To verify this condition we have also performed a trial by taking three to six neurons in the hidden layer (ρ between 3.06 and 1.58) and it was found that the five hidden neurons give the best results as given

Table 4. Correlation matrix of the descriptors selected by neuro-genetic input selection routine.

	<i>GATS2v</i>	<i>MATS7v</i>	<i>MATS5p</i>
<i>GATS2v</i>	1.000		
<i>MATS7v</i>	0.027	1.000	
<i>MATS5p</i>	0.323	0.467	1.000

in table 5. Fitting of the data, stated in R and MSE values, was improved with the increment of hidden neurons. However, Q^2 of LOO cross-validation was increased until a maximum of 0.684 for five neurons, but beyond this value Q^2 began to decrease. This result confirmed as better as the network fits the data worse is its predictive power. Since an important feature of a QSAR model is its ability for making predictions we consider five neurons as optimum value for generating the optimum model. There was not an appreciable influence varying both learning rate and momentum from 0.01 to 0.99, therefore initial conditions were kept. The predicted inhibitory activities for this model are reported in table 2. The plot in figure 1B indicates that there is a significant correlation between experimental and calculated values of $\log(10^6/IC_{50})$ for the ANN model.

Like in MRA model, Moran's and Geary's spatial autocorrelation coefficients weighted by atomic van der Waals volumes and atomic polarizabilities are present in the ANN model. Other similarities are evidenced. Both models had these features in common:

- Spatial autocorrelation coefficient weighted by atomic van der Waals volumes of short lag (1v and 2v in MRA and ANN, respectively).
- Spatial autocorrelation coefficient weighted by atomic van der Waals volumes of large lag (5v and 7v in MRA and ANN, respectively).
- Spatial autocorrelation coefficient weighted by atomic polarizabilities of large lag (6p and 5p in MRA and ANN, respectively).

However, non-linear model overcomes the linear one obtained in this work; in certain way tunes the linear relationship. The network fitted the data with a higher R^2 being able to describe about 92% of data variance in comparison with 68% the linear model. Moreover, the network was able to predict the inhibitory activity of unknown compounds with higher accuracy. Its higher Q^2 of LOO cross-validation of 0.684 emphasized that ANN

Table 5. Variation of statistic parameters with number of hidden neurons in ANN model.

<i>Hidden neurons</i>	<i>R</i>	<i>MSE</i>	<i>Q</i> ²	<i>S_{CV}</i>
3	0.955	0.087	0.621	0.376
4	0.958	0.081	0.625	0.350
5	0.962	0.068	0.684	0.319
6	0.971	0.056	0.636	0.392

has higher predictive power. Neural network approach proved to be more reliable than forward stepwise MRA method. Both models reveal that it exists dependence between inhibitory activities of the FPT inhibitors and the spatial autocorrelations of atomic volumes and polarizabilities on the inhibitor structure, but the neural network approach senses a non-linear dependence.

The identified descriptors have van der Waals volume and polarizability as physicochemical weighting component in them indicating their influence on the inhibitory activity. The biggest difference among the interpretation of two models is evidenced in the shortest-lag term weighted by volumes that appears in the linear model but not in the non-linear one. Lag-one descriptors have particular characteristics. They do not differentiate between linear and branched structural templates. Meanwhile, lag-two descriptors are sensible to branched topological forms. The participation of descriptors of lag five, six and seven may be viewed in terms of association of activity information content with the five, six and seven centered structural fragments. 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.

Several kinds of inhibitors are reported in the last years able to dock the active site of FPT [44]. Published crystal structures have revealed that peptide-mimetic FPT inhibitors may block the peptide substrate site or can occupy part of the peptide site and the exit

groove in a manner similar to the farnesylated peptide product. In piperidine substituted thBCHPs the piperidine is directed in the hydrophobic shallow channel away from the core of FPT [45] interacting with aromatic residues (Trp102 β , Trp106 β , Tyr361 β). Substitutions in piperidine can be tolerated modulating the inhibitory activity. Our QSAR study suggests that size and polarizability of atoms are responsible for this effect.

3.3 Kohonen self-organizing map

Variables selected by NGISR approach were used to obtain a SOM of the inhibitory activity of the cytokinin-derived compounds. We built a 9×9 Kohonen SOM. Figure 2 depicts the KNN map of the data, 38 of a total of 81 neurons were occupied. Eleven neurons were occupied by two compounds at the same time.

As it is observed, compounds with a similar range of activities were grouped into neighboring areas. Noteworthy, thBCHPs with low inhibitory activities were placed in adjacent lowly active neurons at the upper-right zone. On the other hand, the most active derivatives were placed at adjacent highly active neurons through the left and lower areas. Only two compounds are outside: Compound **1c** is located into lowly active neurons neighborhood and its logarithmic activity is above 5; similarly compound **3g** is located into highly active neurons neighbourhood and its logarithmic activity is below 5. Both compounds can be proposed as outliers, since they are not well classified by the selected autocorrelation descriptors.

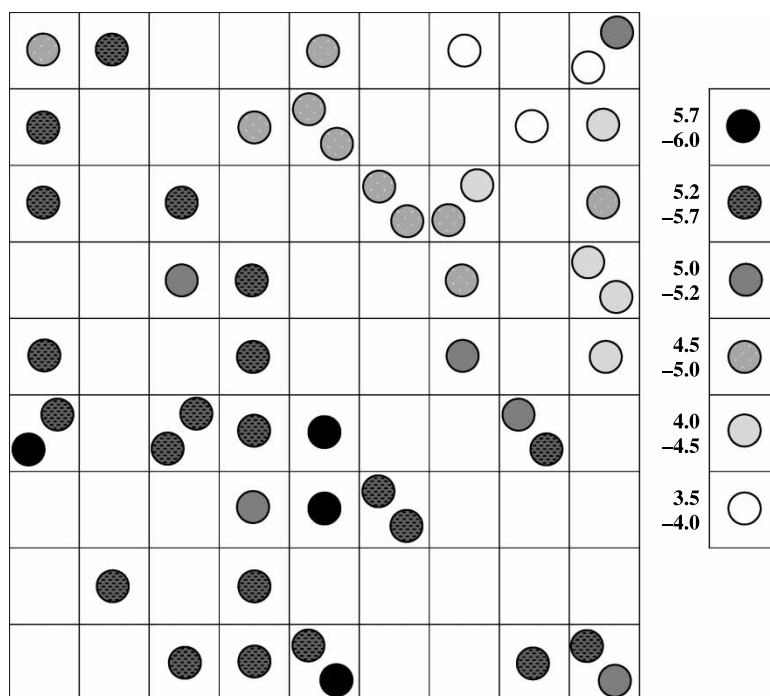


Figure 2. KNN map for the data set using descriptors from non-linear model. Squares and circles denoted neurons and compounds, respectively. Circles at right decode the ranges of inhibitory activities (\log_{10}^6/IC_{50}).

4 Conclusions

Biological phenomena are complex by nature. In this work the inhibitory activity against FPT of a set of piperidine substituted trihalobenzocycloheptapyridine compounds was successfully modelled using MRA and ANN. 2D spatial autocorrelation descriptors were used for encoding structural information of the studied compounds. Neural network approach showed to overcome linear model by having higher correlation coefficient and predictive power. A non-linear dependence between thBCHPs inhibitory activities and the spatial autocorrelations of atomic properties on the inhibitor structure were found. Selected autocorrelation descriptors were also able to well distribute data set on a Kohonen SOM.

References

- [1] F.T. Boyle, G.F. Costello. Cancer therapy: A move to the molecular level. *Chem. Soc. Rev.*, **27**, 251 (1998).
- [2] J.E. Buss, J.C. Marsters Jr. Farnesyl transferase inhibitors: The successes and surprises of a new class of potential cancer chemotherapeutics. *Chem. Biol.*, **2**, 787 (1995).
- [3] S.M. Sehti, A.D. Hamilton. New approaches to anticancer drug design based on the inhibition of farnesyltransferase. *Drug Discov. Today*, **3**, 26 (1998).
- [4] A.M. Garcia, C. Rowell, K. Ackermann, J.J. Kowalczyk, M.D. Lewis. Peptidomimetic inhibitors of Ras farnesylation and function in whole cells. *J. Biol. Chem.*, **268**, 18415 (1993).
- [5] N.E. Kohl, C.A. Omer, M.W. Conner, N.J. Anthony, J.P. Davide, S.J. de Solms, E.A. Giuliani, R.P. Gomez, S.L. Graham, K. Hamilton, L.K. Handt, G.D. Hartman, K.S. Koblan, A.M. Kral, P.J. Miller, S.D. Mosser, T.J. O'Neill, E. Rands, M.D. Schaber, J.B. Gibbs, A. Oliff. Inhibition of farnesyltransferase induces regression of mammary and salivary carcinomas in Ras transgenic mice. *Nat. Med.*, **1**, 792 (1995).
- [6] N.J. Anthony, R.P. Gomez, M.D. Schaber, S.D. Mosser, K.A. Hamilton, T.J. O'Neill, K.S. Koblan, S.L. Graham, G.D. Hartman, D. Shah, E. Rands, N.E. Kohl, J.B. Gibbs, A.I. Oliff. Design and *in vivo* analysis of potent nonthiol inhibitors of farnesyl protein transferase. *J. Med. Chem.*, **42**, 3356 (1999).
- [7] K. Kettler, J. Sakowski, K. Silber, I. Sattler, G. Klebe, M. Schlitzer. Non-thiol farnesyltransferase inhibitors: N-(4-acylamino-3-benzoylphenyl)-3-[5-(4-nitrophenyl)-2-furyl]acrylic acid amides. *Bioorg. Med. Chem.*, **11**, 1521 (2003).
- [8] H. Lee, J. Lee, Y. Shin, W. Jung, J.H. Kim, K. Park, S. Ro, H.H. Chung, J.S. Koh. 3-Aryl-4-aryloyl-1-(1H-imidazol-5-yl)methylpyrrole, a novel class of farnesyltransferase inhibitors. *Bioorg. Med. Chem. Lett.*, **11**, 2963 (2001).
- [9] A. Afonso, J. Weinstein, J. Kelly, R. Wolin, S.B. Rosenblum, M. Connolly, T. Guzi, L. James, D. Carr, R. Patton, W.R. Bishop, P. Kirschmeier, M. Liu, L. Heimark, K.J. Chen, A.A. Nomeir. Analogues of 1-(3, 10-Dibromo-8-chloro-6, 11-dihydro-5H-benzo[5, 6]-cyclohepta[1, 2-b]pyridin-11-yl)piperidine as inhibitors of farnesyl protein transferase. *Bioorg. Med. Chem.*, **7**, 1845 (1999).
- [10] C.J. Dinsmore, M.J. Bogusky, J.C. Culbertson, J.M. Bergman, C.F. Homnick, C.B. Zartman, S.D. Mosser, M.D. Schaber, R.G. Robinson, K.S. Koblan, H.E. Huber, S.L. Graham, G.D. Hartman, J.R. Huff, T.M. Williams. Conformational restriction of flexible ligands guided by the transferred NOE experiment: Potent macrocyclic inhibitors of farnesyltransferase. *J. Am. Chem. Soc.*, **123**, 2107 (2001).
- [11] H. Kubinyi. *QSAR: Hansch Analysis and Related Approaches*, VCH, New York (1993).
- [12] J. Devillers, D. Domine. Comparison of reliability of log *P* values calculated from a group contribution approach and from the autocorrelation method. *SAR QSAR Environ. Res.*, **7**, 195 (1997).
- [13] G. Moreau, P. Broto. Autocorrelation of molecular structures: Application to SAR studies. *Nouv. J. Chim.*, **4**, 757 (1980).
- [14] M. Wagener, J. Sadowski, J. Gasteiger. Autocorrelation of molecular properties for modelling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.*, **117**, 7769 (1995).
- [15] J. Devillers. Autocorrelation descriptors for modelling (eco) toxicological endpoints. In *Topological Indices and Related Descriptors in QSAR and QSPR*, J. Devillers, A.T. Balaban (Eds.), pp. 595–612, Gordon and Breach Science Publishers, Amsterdam (1999).
- [16] E. Giraud, C. Luttmann, F. Lavelle, J.F. Riou, P. Mailliet, A. Laoui. Multivariate data analysis using D-optimal designs, partial least squares, and response surface modeling: A directional approach for the analysis of farnesyltransferase inhibitors. *J. Med. Chem.*, **43**, 1807 (2000).
- [17] E. Estrada, E. Uriarte, A. Montero, M. Teixeira, L. Santana, E. De Clercq. A novel approach for the virtual screening and rational design of anticancer compounds. *J. Med. Chem.*, **43**, 1975 (2000).
- [18] M.J. Polley, D.A. Winkler, F.R. Burden. Broad-based quantitative structure–activity relationship modeling of potency and selectivity of farnesyltransferase inhibitors using a Bayesian regularized neural network. *J. Med. Chem.*, **47**, 6230 (2004).
- [19] H. Bauknecht, A. Zell, A.H. Bayer, P. Levi, M. Wagener, J. Sadowski, J. Gasteiger. Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: Dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.*, **36**, 1205 (1996).
- [20] P.A.P. Moran. Notes on continuous stochastic processes. *Biometrika*, **37**, 17 (1950).
- [21] R.F. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, **5**, 115 (1954).
- [22] G. Moreau, P. Broto. Autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.*, **4**, 359 (1980).
- [23] F.G. Njoroge, B. Vibulbhan, P. Pinto, C.L. Strickland, W.R. Bishop, P. Kirschmeier, V. Girijavallabhan, A.K. Ganguly. Trihalobenzocycloheptapyridine analogues of SCH 66336 as potent inhibitors of farnesyl protein transferase. *Bioorg. Med. Chem.*, **11**, 139 (2003).
- [24] J.J.P. Stewart. Optimization of parameters for semi-empirical methods I-method. *J. Comp. Chem.*, **10**, 210 (1989).
- [25] MOPAC version 6.0. Frank J. Seiler Research Laboratory, U.S. Air Force academy (1993).
- [26] DRAGON. version 2.1.R. Todeschini, V. Consonni, M. Pavan (2002).
- [27] R.B. Kowalski, S. Wold. Pattern recognition in chemistry. In *Handbook of Statistics*, P.R. Krishnaiah, L.N. Kanal (Eds.), pp. 673–697, North Holland Publishing Company, Amsterdam (1982).
- [28] (2001). STATISTICA (data analysis software system), version 6. StatSoft, Inc.
- [29] J. Zupan, J. Gasteiger. Neural networks: A new method for solving chemical problems or just a passing faze? *Anal. Chim. Acta*, **248**, 1 (1991).
- [30] T. Aoyama, Y. Suzuki, H. Ichikawa. Neural networks applied to structure–activity relationships. *J. Med. Chem.*, **33**, 905 (1990).
- [31] A.N. Kolmogorov. *Dokl. Akad. Nauk SSSR*, **114**, 953 (1957).
- [32] S.S. So, M. Karplus. Genetic neural networks for quantitative structure–activity relationships: Improvement and application of benzodiazepine affinity for benzodiazepine/GABA A receptor. *J. Med. Chem.*, **39**, 5246 (1996).
- [33] StatSoft, Inc. Electronic Statistics Textbook. Tulsa, OK: web: <http://www.statsoft.com/textbook/stathome.html> (2004).
- [34] Matlab 6.5 The Math Works, Inc. (2002).
- [35] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, **43**, 59 (1982).
- [36] J. Gasteiger, J. Zupan. Neural networks in chemistry. *Angew. Chem. Int. Ed. Engl.*, **32**, 503 (1995).
- [37] J. Gasteiger, X. Li. Mapping the electrostatic potential of muscarinic and nicotinic agonists with artificial neural networks. *Angew. Chem. Int. Ed. Engl.*, **33**, 643 (1994).
- [38] D.M. Hawkins. The problem of overfitting. *J. Chem. Inf. Comput. Sci.*, **44**, 1 (2004).
- [39] R. Todeschini, V. Consonni. *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim (2000).
- [40] S. So, W.G. Richards. Application of neural networks: Quantitative structure–activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl) pyrimidines as DHFR inhibitors. *J. Med. Chem.*, **35**, 3201 (1992).

- [41] M. Zahouily, A.R. Bazoui, S. Sebt, D. Zakarya. Structure–cytotoxicity relationships for a series of HEPT derivatives. *J. Mol. Model.*, **8**, 168 (2002).
- [42] A. Yasri, D. Hartsough. Toward an optimal procedure for variable selection and QSAR model building. *J. Chem. Inf. Comput. Sci.*, **41**, 1218 (2001).
- [43] R. Vanyúr, K. Héberger, J. Jakus. Prediction of anti-HIV-1 activity of a series of tetrapyrrole molecules. *J. Chem. Inf. Comput. Sci.*, **43**, 1829 (2003).
- [44] I.M. Bell. Inhibitors of farnesyltransferase: A rational approach to cancer chemotherapy? *J. Med. Chem.*, **47**, 1869 (2004).
- [45] C.L. Strickland, P.T. Weber, W.T. Windsor, Z. Wu, H.V. Le, M.M. Albanese, C.S. Alvarez, D. Cesarz, J. del Rosario, J. Deskus, A.K. Mallams, F.G. Njoroge, J.J. Piwinski, S. Remiszewski, R.R. Rossman, A.G. Taveras, B.V. Vibulbhan, R.J. Doll, V.M. Girijavallabhan, A.K. Ganguly. Tricyclic farnesyl protein transferase inhibitors: Crystallographic and calorimetric studies of structure–activity relationships. *J. Med. Chem.*, **42**, 2125 (1999).