

Modeling of farnesyltransferase inhibition by some thiol and non-thiol peptidomimetic inhibitors using genetic neural networks and RDF approaches

Maykel Pérez González,^{a,b} Julio Caballero,^{c,d} Alain Tundidor-Camba,^e
Aliuska Morales Helguera^{b,f} and Michael Fernández^{c,d,*}

^aUnit of Service, Drug Design Department, Experimental Sugar Cane Station “Villa Clara-Cienfuegos”,
Ranchuelo, Villa Clara, C.P. 53100, Cuba

^bChemical Bioactive Center, Central University of Las Villas, Santa Clara, Villa Clara, C.P. 54830, Cuba

^cMolecular Modeling Group, Center for Biotechnological Studies, University of Matanzas, Matanzas, C.P. 44740, Cuba

^dProbiotic Group, Center for Biotechnological Studies, University of Matanzas, Matanzas, C.P. 44740, Cuba

^eScientific Prospection Group, National Center for Scientific Researches (CNIC), PO Box 6880, Havana, Cuba

^fDepartment of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas,
Santa Clara, Villa Clara, C.P. 54830, Cuba

Received 31 May 2005; revised 1 August 2005; accepted 2 August 2005

Available online 26 September 2005

Abstract—Inhibition of farnesyltransferase (FT) enzyme by a set of 78 thiol and non-thiol peptidomimetic inhibitors was successfully modeled by a genetic neural network (GNN) approach, using radial distribution function descriptors. A linear model was unable to successfully fit the whole data set; however, the optimum Bayesian regularized neural network model described about 87% inhibitory activity variance with a relevant predictive power measured by q^2 values of leave-one-out and leave-group-out cross-validations of about 0.7. According to their activity levels, thiol and non-thiol inhibitors were well-distributed in a topological map, built with the inputs of the optimum non-linear predictor. Furthermore, descriptors in the GNN model suggested the occurrence of a strong dependence of FT inhibition on the molecular shape and size rather than on electronegativity or polarizability characteristics of the studied compounds.

© 2005 Elsevier Ltd. All rights reserved.

1. Introduction

In recent years, targeting farnesyltransferase (FT) enzyme has become a promising strategy in cancer therapy.¹ Transferring a farnesyl from farnesylpyrophosphate to the thiol of a cysteine side chain of protein residues, which bear the CAAX-tetrapeptide sequence (C: cysteine, A: aliphatic amino acid, and X: serine or methionine) at their C-terminus, is catalyzed by FT enzyme.² The inhibition of such an enzyme as a cancer therapy alternative is based on the fact that farnesylation is a pre-requisite for the transforming activity of oncogenic Ras, which is found in approximately 30% of all cancers in humans.³ However, recent evidence has been accumulating suggesting that

Ras may not be the only substrate that is involved in oncogenesis.² Attention has been shifted to RhoB, another member of the class of small GTPases involved in receptor trafficking.^{4,5} Although the mechanism by which FT inhibitors (FTIs) display their antiproliferative activity remains unsolved, the efficacy and low toxicity of such compounds have been demonstrated. Therefore, FT inhibition is considered a major emerging strategy in cancer treatment.⁶

Since the FT enzyme only recognizes and binds the last four C-terminal amino acids of the CAAX-consensus of substrate proteins, this tetrapeptide is used as a primary template for developing non-peptide FTIs.² The majority of those CAAX-peptidomimetics possess a free thiol group that coordinates the enzyme-bound zinc ion, as it has been shown for the native peptide substrate. However, non-thiol FTIs have also been developed recently taking into account several adverse effects associated with the thiol group.⁷ In such inhibitors, the coordination to the

Keywords: Farnesyltransferase; Neural networks; Genetic algorithm; Enzyme inhibition; QSAR.

* Corresponding author. Tel.: +53 45 26 1251; fax: +53 45 25 3101; e-mail addresses: michael.fernandez@umcc.cu; michael_llamosa@yahoo.com

zinc ion is usually accomplished by nitrogen-containing heterocycle;⁸ however, replacement of the heterocycles by non-metal-coordinating aryl residues can be carried out without losing too much of FT inhibitory activity.^{9,10} This fact has suggested the occurrence of at least one hitherto unknown aryl binding region on the FT active site.^{11,12}

Computational-based rational design of drugs has increased in the last decade. Most of those approaches are focused on quantitative structure–activity relationship (QSAR) studies, using different kinds of molecular descriptors for encoding chemical information.^{13–16} After computing a set of descriptors, multivariate linear or/and non-linear relationships are established between a reduced subset of variables and the inhibitory activity, leading to a mathematical model.

Since interactions between a chemical and a biological system are non-linear by nature, artificial neural network (ANN) methodology has been successfully applied in QSAR studies of biological activities yielding, in most of the cases, better results than multilinear regression analysis (MRA).^{16–23} Besides the non-linearity existing between biological activities and the computed molecular descriptors, another major problem arises when the number of calculated variable exceeds the number of compounds in the data set, so that one is dealing with an undetermined problem where undesirable overfitting can result.²¹ This problem can be handled by implementing a feature selection routine that determines which of the descriptors has a significant influence on the activity of a set of compounds. Genetic algorithm (GA), rather than forward or backward elimination procedures, 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.^{16–23}

In the context of *in silico* methods for modeling physico-chemical and biological properties of chemicals, the radial distribution function (RDF) approach has been introduced.²⁴ Successful application of this theoretical approach for deriving the 3D structure of organic molecules from their infrared spectra^{25,26} has inspired us to test and/or validate the RDF descriptors applicability in assessing discoveries of new drugs.

In this work, we employed GA for building linear and non-linear predictive models for the inhibition of FT by a data set of 78 CAAX-peptidomimetic inhibitors including 32 thiol and 46 non-thiol FTIs (Table 1). In a first approach, RDF descriptors were used for obtaining linear models of the studied property. In addition to a GA-based MRA, genetic neural network (GNN) approach was used for building an optimum neural network model with the RDF descriptors. To gain in performance both robustness of predictions and speed of computation ANNs, Bayesian regularization was implemented in a Levenberg–Marquardt algorithm for error minimization during supervised training of full-connected feed-forward ANNs. Furthermore, versatility of the ANNs was also used for mapping the FT inhibitory activities on a topological map using

competitive neural networks to address structural features related to the activity of the studied compounds.

2. Results and discussion

Taking into account that RDF descriptors encode information highly depend on molecular 3D structures, after structural optimization by semi-empirical method PM3, inhibitors were aligned with the CAAX-peptidomimetic portion from the crystal structure of a ternary complex of farnesyltransferase, farnesylpyrophosphate analogue, and *N*-acetyl-Cys-Val-Ile-selenoMetOH (PDB 1QBQ)²⁷ (Fig. 1), and it was established that there is no discordance among the optimized structures and the conformations they should adopt at the active site of the FT enzyme.

2.1. Multilinear regression analysis

As we previously pointed out, in a first approach, a MRA for the FT inhibitory activity of the studied compounds was achieved by means of the GA search routine. The symbols and definitions of the RDF descriptors in the models are given in Table 2. The model selection was subjected to the parsimony principle.²⁸ Then, we chose a function with high statistical significance but having as few descriptors as possible. The best QSAR model obtained is given below together with the statistical parameters of regression:

$$\begin{aligned} -\log(\text{IC}_{50}) = & 1.590 - 0.099 \cdot \text{RDF040u} - 0.102 \\ & \cdot \text{RDF140u} + 0.177 \cdot \text{RDF140m} \\ & + 0.077 \cdot \text{RDF090v} + 0.187 \\ & \cdot \text{RDF125v} + 0.067 \cdot \text{RDF050e} \\ & - 0.122 \cdot \text{RDF150p}, \end{aligned} \quad (1)$$

$$\begin{aligned} N = 78, \quad R^2 = 0.667, \quad S = 0.498, \quad F = 20.010, \\ p < 10^{-5}, \quad q_{\text{LOO}}^2 = 0.581, \quad S_{\text{LOO}} = 0.559, \\ q_{\text{LGO}}^2 = 0.577, \quad S_{\text{LGO}} = 0.534, \end{aligned}$$

where $-\log(\text{IC}_{50})$ is the studied property, N is the number of compounds included in the model, R^2 is the correlation coefficient, S is the standard deviation of the regression, F is the Fisher ratio, p is the significance of the variables in the model, q_{LOO}^2 and q_{LGO}^2 are the correlation coefficients of the LOO and LGO cross-validation, respectively, and S_{LOO} and S_{LGO} are the standard deviations of the LOO and LGO cross-validation, respectively. This seven-variable RDF QSAR model (MRA 1) explains only about 67% of inhibitory activities with a discrete predictive power measured by q^2 values of LOO and LGO > 0.5. As regards the poor statistical quality of the model MRA 1, we attempted to improve its reliability by removing some compounds having large residual activities as outliers.

An outlier to a QSAR is identified normally by having a large standard residual and can indicate the limits of

Table 1. Chemical structures of thiol and non-thiol FTIs, and experimental and predicted inhibitory activities by linear model MRA 1 and non-linear model BRANN 2

Compound	R1	-log(IC ₅₀) ^a			R1	-log(IC ₅₀) ^a			
		Exp.	MRA 1	BRANN 2		Exp.	MRA 1	BRANN 2	
1		3.187	3.681	3.207					
2	-CH ₃	2.119	2.513	2.490	16	X = -CH ₃	4.114	3.458	3.340
3		2.066	1.784	1.861	17	X = -Cl	3.979	3.327	3.740
4		2.745	2.890	2.976	18	X = -Br	4.119	4.106	4.187
5		3.155	2.986	3.222	19	X = -NO ₂	3.939	3.484	3.918
6		2.620	3.092	2.688	20	X = -CF ₃	3.140	3.819	2.992
7		3.260	2.824	3.126	21	X = -OCH ₃	3.264	3.256	3.488
8		2.321	2.419	2.611	22	X = -Ph	2.703	3.620	3.201
9		2.889	2.644	2.641					
10		2.848	3.443	3.053	23	X = 2,4-di-Cl	3.921	3.531	3.684
11		2.337	2.599	2.507	24	X = 3,4-di-Cl	3.319	3.398	3.490
12		3.420	3.861	3.506	25	X = 2,6-di-Cl	3.469	3.271	2.987
13		3.337	3.304	3.499	26	X = 3,5-di-Cl	3.272	2.993	2.746
14		3.398	3.617	3.231	27	X = 2,5-di-Cl	2.620	3.374	3.121
15		3.268	3.474	3.213	28	X = 2,3-di-Cl	3.377	3.495	3.137
				29	R = -CH ₃	4.071	3.632	3.885	
				30	R = -Ph	4.000	3.449	4.300	

Table 1. (continued)

R1	$-\log(\text{IC}_{50})^a$			R1	$-\log(\text{IC}_{50})^a$						
	Exp.	MRA 1	BRANN 2		Exp.	MRA 1	BRANN 2				
31		3.187	3.753	3.248	32		4.444	3.796	4.316		
R1	$-\log(\text{IC}_{50})^b$			R1	$-\log(\text{IC}_{50})^b$						
	Exp.	MRA 1	BRANN 2		Exp.	MRA 1	BRANN 2				
33		2.620	2.453	2.205	37		2.201	2.637	2.612		
34		3.939	3.072	3.803	38		2.000	2.702	2.197		
35		2.252	2.962	2.262	39		3.119	3.633	3.598		
36		2.824	3.344	3.459	40		3.959	3.888	3.785		
R2	$-\log(\text{IC}_{50})^c$			R2	$-\log(\text{IC}_{50})^c$						
	Exp.	MRA 1	BRANN 2		Exp.	MRA 1	BRANN 2				
41		3.910	3.822	4.085	53		4.585	5.019	4.743		

Table 1. (continued)

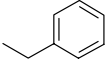
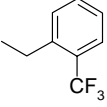
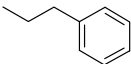
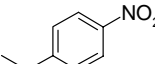
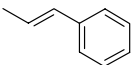
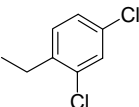
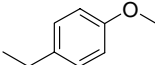
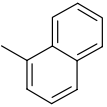
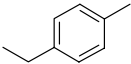
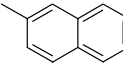
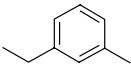
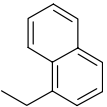
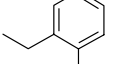
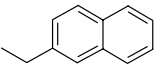
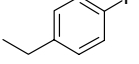
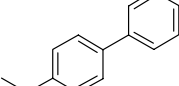
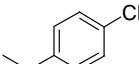
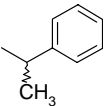
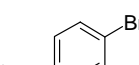
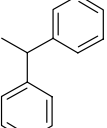
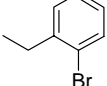
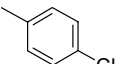
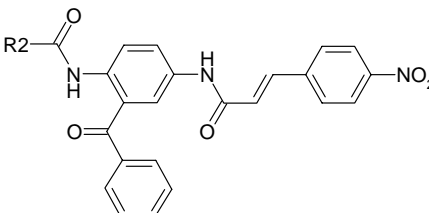
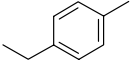
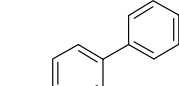
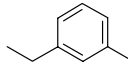
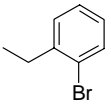
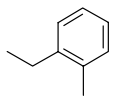
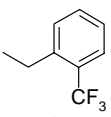
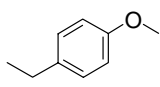
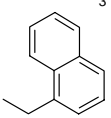
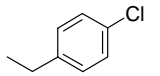
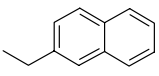
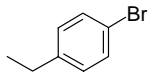
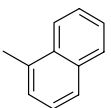
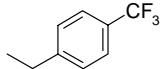
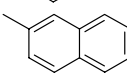
R2	-log(IC ₅₀) ^c			R2	-log(IC ₅₀) ^c		
	Exp.	MRA 1	BRANN 2		Exp.	MRA 1	BRANN 2
42 	5.222	3.978	4.767	54 	4.854	4.182	4.911
43 	4.398	4.324	4.435	55 	3.867	4.376	4.177
44 	5.301	4.035	4.864	56 	4.036	4.169	4.260
45 	4.456	4.407	4.315	57 	3.333	3.627	3.271
46 	4.456	4.245	4.827	58 	3.733	3.849	3.263
47 	4.237	4.481	4.304	59 	5.222	5.097	4.954
48 	3.686	4.013	4.226	60 	5.097	4.523	4.870
49 	4.824	4.302	4.551	61 	5.155	5.015	4.949
50 	3.991	4.182	4.177	62 	4.796	5.121	4.584
51 	4.301	4.276	4.097	63 	3.181	4.004	3.231
52 	4.824	5.003	4.771	64 	3.932	3.852	4.150
							
R2	-log(IC ₅₀) ^d			R2	-log(IC ₅₀) ^d		
	Exp.	MRA 1	BRANN 2		Exp.	MRA 1	BRANN 2
65 	3.629	3.161	3.788	72 	4.357	4.072	4.215

Table 1. (continued)

R2		-log(IC ₅₀) ^d			R2		-log(IC ₅₀) ^d		
		Exp.	MRA 1	BRANN 2			Exp.	MRA 1	BRANN 2
66		4.056	3.735	3.764	73		4.187	4.220	4.082
67		4.181	3.271	4.148	74		4.377	4.641	4.096
68		3.011	3.542	3.722	75		4.168	4.246	3.678
69		3.084	3.462	3.274	76		3.022	3.506	3.431
70		4.337	3.500	4.218	77		4.284	3.921	4.427
71		4.301	4.454	4.342	78		3.678	3.667	3.409

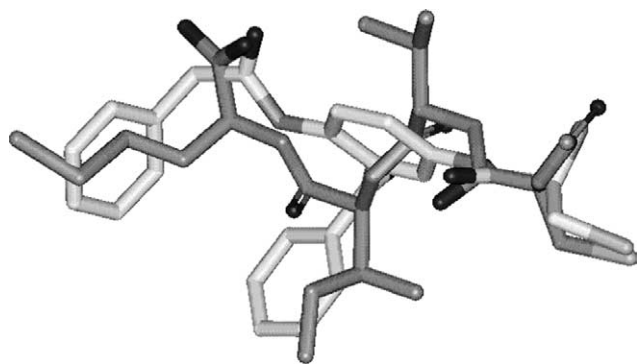
^a From Ref. 2.^b From Ref. 31.^c From Ref. 39.^d From Ref. 6.

Figure 1. Alignment between the farnesyltransferase inhibitor **1** and the enzyme-bound conformation of *N*-acetyl-Cys-Val-Ile-selenoMetOH (PDB 1QBQ).

applicability of QSAR models.²⁹ There are several reasons for their occurrence in QSAR studies; for example, chemicals might be acting by a mechanism different from that of the majority of the data set. It is also likely that outliers might be a result of a random experimental error that could be significant when analyzing the large data sets. Although it is acceptable to remove a small number of outliers from the QSAR,³⁰ it is noted that it is not acceptable to remove the outliers repeatedly from a QSAR analysis simply for improving a correlation. In the current work, the compounds **42** and **44** present large residuals (>1.0) and should be considered as outliers. At removal of these compounds from the training set, the next equation is obtained:

$$\begin{aligned}
 -\log(\text{IC}_{50}) = & 1.436 - 0.083 \cdot \text{RDF040u} - 0.106 \\
 & \cdot \text{RDF140u} + 0.191 \cdot \text{RDF140m} \\
 & + 0.060 \cdot \text{RDF090v} + 0.179 \\
 & \cdot \text{RDF125v} + 0.666 \cdot \text{RDF050e} \\
 & - 0.122 \cdot \text{RDF150p}, \quad (2)
 \end{aligned}$$

$$N = 76, \quad R^2 = 0.701, \quad S = 0.453, \quad F = 23.072,$$

$$p < 10^{-5}, \quad q_{\text{LOO}}^2 = 0.618, \quad S_{\text{LOO}} = 0.514,$$

$$q_{\text{LOO}}^2 = 0.616, \quad S_{\text{LOO}} = 0.498.$$

The reliability of the linear model increased when deleting the outliers, as reflected by the statistical quantities of Eq. 2. This linear model describes about 71% of data variance, and more importantly it shows a higher predictive power than model 1 (Eq. 1) measured by q^2 values of LOO and LGO > 0.6. However, the removed outliers, compounds **42** and **44**, are the most active FTIs in our data.

It is well-known that inhibition of FT involves coordination with a zinc ion in the enzyme active site for thiol inhibitors; however, non-thiol inhibitors present different main interactions involving matching to hydrophobic moieties on the enzyme active site cavity.^{6,31} Since different mechanisms should govern enzyme inhibition by each inhibitor type, we tried to obtain separate models for the 32 thiol and the 46 non-thiol FTIs forming

Table 2. Symbols and definitions of the RDF descriptors appearing in models MRA 1, MRA2, MRA3, MRA 4, and BRANN 2

Symbol	Definition
RDF025u	Radial distribution function at 2.5 Å/unweighted
RDF040u	Radial distribution function at 4.0 Å/unweighted
RDF065u	Radial distribution function at 6.5 Å/unweighted
RDF080u	Radial distribution function at 8.0 Å/unweighted
RDF140u	Radial distribution function at 14.0 Å/unweighted
RDF015m	Radial distribution function at 1.5 Å/weighted by atomic masses
RDF055m	Radial distribution function at 5.5 Å/weighted by atomic masses
RDF065m	Radial distribution function at 6.5 Å/weighted by atomic masses
RDF085m	Radial distribution function at 8.5 Å/weighted by atomic masses
RDF110m	Radial distribution function at 11.0 Å/weighted by atomic masses
RDF125m	Radial distribution function at 12.5 Å/weighted by atomic masses
RDF135m	Radial distribution function at 13.5 Å/weighted by atomic masses
RDF140m	Radial distribution function at 14.0 Å/weighted by atomic masses
RDF090v	Radial distribution function at 9.0 Å/weighted by van der Waals volumes
RDF110v	Radial distribution function at 9.0 Å/weighted by van der Waals volumes
RDF125v	Radial distribution function at 12.5 Å/weighted by van der Waals volumes
RDF150v	Radial distribution function at 15.0 Å/weighted by van der Waals volumes
RDF025e	Radial distribution function at 2.5 Å/weighted by Sanderson electronegativities
RDF035e	Radial distribution function at 3.5 Å/weighted by Sanderson electronegativities
RDF050e	Radial distribution function at 5.0 Å/weighted by Sanderson electronegativities
RDF070e	Radial distribution function at 7.0 Å/weighted by Sanderson electronegativities
RDF135e	Radial distribution function at 13.5 Å/weighted by Sanderson electronegativities
RDF155e	Radial distribution function at 15.5 Å/weighted by Sanderson electronegativities
RDF150p	Radial distribution function at 15.0 Å/weighted by atomic polarizabilities

our data set using the RDF descriptors. Similar to the full data set MRA model, two optimum linear models were built for the thiol FTIs, model MRA 3 (Eq. 3) and non-thiol FTIs, model MRA 4 (Eq. 4), having six and seven variables, respectively, and such equations are reported below:

$$\begin{aligned}
 -\log(\text{IC}_{50}) = & -4.434 - 0.205 \cdot \text{RDF150u} + 0.280 \\
 & \cdot \text{RDF015m} + 0.100 \cdot \text{RDF110m} \\
 & + 0.108 \cdot \text{RDF025e} - 0.060 \\
 & \cdot \text{RDF035e} - 0.137 \cdot \text{RDF135e}, \quad (3)
 \end{aligned}$$

$$\begin{aligned}
 N = 32, \quad R^2 = 0.884, \quad S = 0.234, \quad F = 31.702, \\
 p < 10^{-5}, \quad q_{\text{LOO}}^2 = 0.804, \quad S_{\text{LOO}} = 0.304, \\
 q_{\text{LOO}}^2 = 0.751, \quad S_{\text{LGO}} = 0.313,
 \end{aligned}$$

$$\begin{aligned}
 -\log(\text{IC}_{50}) = & 3.586 + 0.128 \cdot \text{RDF025u} - 0.152 \\
 & \cdot \text{RDF065u} - 0.243 \cdot \text{RDF080u} \\
 & + 0.184 \cdot \text{RDF125m} - 0.255 \\
 & \cdot \text{RDF150v} + 0.089 \cdot \text{RDF070e} \\
 & + 0.098 \cdot \text{RDF155e}, \quad (4)
 \end{aligned}$$

$$\begin{aligned}
 N = 46, \quad R^2 = 0.792, \quad S = 0.407, \quad F = 20.714, \\
 p < 10^{-5}, \quad q_{\text{LOO}}^2 = 0.703, \quad S_{\text{LOO}} = 0.487, \\
 q_{\text{LOO}}^2 = 0.700, \quad S_{\text{LOO}} = 0.448.
 \end{aligned}$$

As it can be observed, both models show higher statistical significances in comparison to linear models MRA 1 and MRA 2. Remarkably, model MRA 3 and MRA 4

have improved predictive powers exhibiting q^2 values >0.7 . This result pointed out that the more similar the mechanism of action of the chemicals on the fitted data, the more reliable the linear relationship that can be established between the RDF molecular descriptors and the FT inhibitory activities. Indeed, this fact constitutes a drawback of the linear QSAR studies that limited range of applications. The linear relationship here obtained for the whole data is unable to completely solve the convolute dependences between molecular descriptors and the biological activities that appear in the multifactor and partially hitherto unknown mechanisms of inhibition of the FT enzyme.

2.2. Genetic neural network analysis

Since biological interactions are non-linear by nature, an aim of this work was to find a reliable non-linear model for the inhibition of farnesyltransferase. Taking into account the poor statistical significance of model MRA 1, we expected to find adequate combinations of seven variables for reliable modeling of the FT inhibitory activity by means of the GNN approach using the calculated RDF descriptors. In spite of the unfit behavior of compounds **42** and **44** in the MRA analysis, inside the GNN framework, networks were trained with the whole data set with the aim of looking for a non-linear model that fits all the compounds under investigation well. 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 a whole data set to train the networks. However, the

use of a MSE fitness function could lead to undesirable well-fitted albeit poorly generalized networks as algorithm solutions. In this connection, we tried to avoid such results by following two aspects: 1) keeping network architectures as simplest as possible (7-3-1) inside the GA framework and 2) implementing Bayesian regularization in the network training function (Section 3.4).

In addition to the best seven-input BRANN predictor with a 7-3-1 architecture yielded by the GA routine, Table 3 shows the structures and statistics of several networks obtained by a screening process in which the number of hidden nodes in the initial predictor was varied with the aim of optimizing the reliability of the

networks. Symbols and definitions of the descriptors are given in Table 2. As can be observed, three hidden node network was the optimum predictor exhibiting R^2 , q^2 of LOO and q^2 of LGO values of 0.874, 0.701, and 0.674, respectively. Increasing the number of neurons in the hidden layer causes a gradual decrease in predictive power of the BRANNs, previously observed by Burden and Winkler.³² On the contrary, two hidden nodes are not sufficient for reliably solving the non-linear relationship between RDF descriptors and the activity. As reported in Table 3, the number of optimum parameters yielded by the Bayesian regularization was asymptotic, with a maximum number of optimum parameters equal to 45. However, the best

Table 3. Statistics of the non-linear predictors for the FT inhibitory activity of the thiol and non-thiol FTIs. Optimum neural network predictor appears in boldface

Descriptors	BRANN model	Hidd. nod.	Num. par.	Opt. par.	R^2	S	q^2_{LOO}	S_{LOO}	q^2_{LGO}	S_{LGO}
RDF055m	1	2	19	12	0.582	0.530	0.386	0.647	0.423	0.632
RDF065m	2	3	28	25	0.874	0.291	0.701	0.450	0.674	0.470
RDF085m	3	4	37	31	0.894	0.268	0.601	0.530	0.605	0.532
RDF125m	4	5	46	37	0.918	0.236	0.654	0.492	0.620	0.530
RDF135m	5	6	55	41	0.923	0.221	0.553	0.588	0.511	0.632
RDF140m	6	7	64	45	0.943	0.197	0.491	0.634	0.475	0.645
RDF110v	7	8	73	45	0.940	0.202	0.489	0.632	0.472	0.641

Hidd. nod. represents the number of hidden nodes, Num. par. represents the number of neural network parameters, Opt. par. represents the optimum number of neural network parameters yielded by the Bayesian regularization, R^2 is the square correlation coefficient for the data fitting in the model, q^2_{LOO} and q^2_{LGO} are the square correlation coefficients of the predictions in the LOO and LGO cross-validation process, S is the standard deviation of the data fitting in the model, and S_{LOO} and S_{LGO} are the standard deviations of the predictions in the LOO and LGO cross-validations.

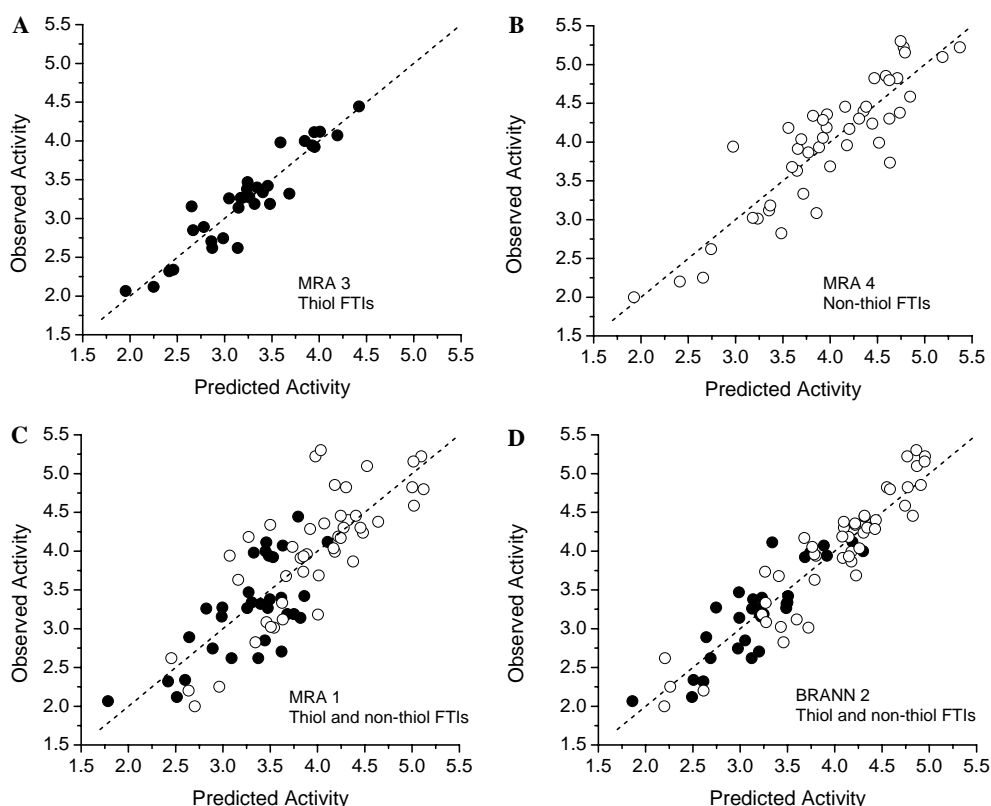


Figure 2. Plots of observed versus predicted $-\log(\text{IC}_{50})$ FT inhibitory activity of thiol (●) and non-thiol (○) FTIs according to linear models MRA 3 (A), MRA 4 (B), and MRA 1 (C) including thiol inhibitors, non-thiol inhibitors and whole data set, respectively, and non-linear model BRANN 2 (D) including the whole data set. The dotted lines are an ideal fit, with the respective intercept and slope equal to 0 and 1.

predictor with three hidden nodes exhibits an optimum number of parameters equal to 25 for an initial number of parameters equal to 28.

It should be noted that statistical parameters of data fitting for predictor BRANN 2 are of higher quality than the statistics reported for MRA models for the whole data set. Furthermore, it is noteworthy to observe an improvement in the parameters of LOO and LGO cross-validations, taking into account that this optimum non-linear predictor containing the whole data set has LOO and LGO q^2 values of about 0.7, whilst linear models MRA 1 and MRA 2 have poorer q^2 values of about 0.6.

Figure 2 shows plots of the observed versus predicted FT inhibitory activity of MRA models 1, 3, and 4 as well as for non-linear predictor BRANN 2. As can be observed in Figures 2A and B, an MRA approach is able to yield two well-fitted models for thiol and non-thiol inhibitors separately. However, Figure 2C confirms that this linear method poorly fits the whole data set. On the other hand, the plot in Figure 2D depicts that the neural network is able to fit the whole data set with accuracy higher than those of the MRA models. Meanwhile, predictor BRANN 2 describes approximately 90% FT inhibitory activity variance, while the MRA approach describes only about 70% of the whole data set and even deleting two outliers.

2.3. 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 topological map of the FT inhibitory activity. Figure 3 shows the

10 × 10 KSOM of the data; 63 of a total of 100 neurons were occupied. As it is observed, compounds with a similar range of activity were grouped into neighboring areas. The less active compounds are grouped in the upper-left diagonal half region, whilst the most active compounds are grouped in the lower-right diagonal half region. A total of eight neurons were classified as ‘conflictive neurons’ in which compounds were misallocated, and such neurons are pointed out in the map by a cross. Some structural similarities can be addressed among compounds allocated in neighboring neurons. Low active compounds, including thiol derivatives with small and hydrophobic substituents at position R1, compounds 2, 3, 4, 8, 9, 10, and 31, appear at the upper-left zone of the map (zone 1). On the other hand, in the lower-middle zone (zone 2) are located highly active non-thiol compounds 41, 42, 43, 44, 45, 46, 47, 49, 50, 62, and 64, which present as common feature a monosubstituted phenyl or benzyl substituents at position R2. At the same time, compounds with bulkier R2 substituents, compounds 52, 53, 55, 56, 59, 60, 61, and 72, but having a similar range of high inhibitory activity as compounds in zone 2 are grouped at the lower-right zone (zone 3).

2.4. Models’ interpretation

Little has been done on the modeling of FT inhibition by the QSAR approach. However, Polley et al. recently reported a broad Bayesian neural network QSAR study on a large data set of about 2000 FTIs³³ in which the importance of subtle hydrophobic effects in the binding site was reflected. In this paper, the authors made a comprehensive review of the ‘state of the art’ concerning FT inhibition modeling. They mentioned that the most of the previous QSAR works on FTIs are focused on small

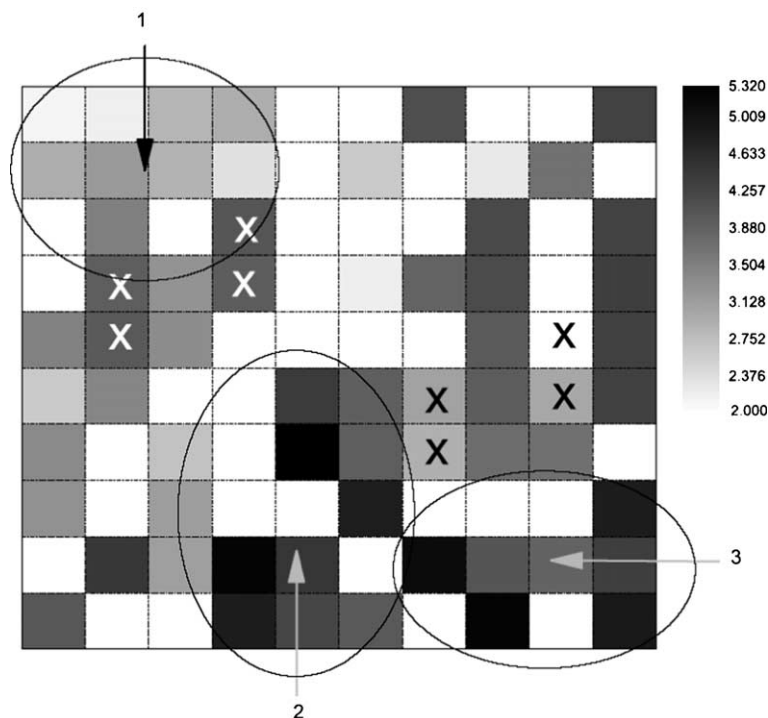


Figure 3. Kohonen self-organizing map of the $-\log(\text{IC}_{50})$ FT inhibitory activity of thiol and non-thiol FTIs. Crosses mean conflictive neurons. Circles depict zones 1, 2, and 3 in the map grouping compounds with common structural features having similar range of inhibitory activities.

data sets, having highly congeneric series of molecules^{34–37} or reported models that are qualitative rather than quantitative such as the TOPS-MODE model reported by Estrada et al. for the anticancer activity of FTIs.³⁸

Our results well agree with the report of Polley et al., in which they showed that modeling of FT inhibition using MRA yielded models of lower quality than with the BRANN they used. In this connection, similar to their results, we obtained a linear model describing about 70% of whole data set variance, whilst our best BRANN model is able to describe about a 90% but having a slightly lower predictive power with q^2 values of about 0.70 in comparison with the value of 0.76 reported by Polley et al. for their BRANN model. However, our non-linear predictor overcomes in predictive power the previous linear models cited by those authors with q^2 values of about 0.5–0.6. This fact strongly supports the thesis pointed out by Polley et al.³³ suggesting that FT inhibition has a nonlinear component.

Interpreting a QSAR model in terms of the specific contribution of substituents and other molecular features to the modeled activity is always a difficult task. In this paper, RDF descriptors in linear model MRA 1 suggest the occurrence of some linear dependence between the inhibitory activity of the thiol and non-thiol compounds, and the 3D molecular distribution of mass, van der Waals volume, polarizability, and electronegativity calculated at radii ranging from 4.0 to 15.0 Å from the geometrical center of each molecule. Since the maximum molecular radius of the compounds in our data set is about 15.0 Å, we conclude that model MRA 1 includes contributions from both inner and outer parts of the inhibitor molecules. This fact well agrees with docking studies performed on FTIs in which the peptidomimetic inhibitor should match into a multiresidue cavity resembling the C-terminal amino acids of the CAAX-consensus of substrate proteins.² The GA-based MRA extracted the linear contributions to the FT inhibition of mass, van der Waals volume, polarizability, and electronegativity distributions. However, beyond the interpretation made, the linear model obtained showed poor predictive power, even when two outliers were removed. The fact that better behaviors were achieved when modeling the inhibition by thiol and non-thiol compounds separately suggests that the complexity of the inhibition of FT enzyme is out of the range of a successful application of the MRA approach.

On the contrary, in this study, ANNs provided a closer approximation to the biological phenomenon by yielding a more reliable model. The optimum non-linear model also includes contributions from the inner to outer parts of the inhibitor structure. Interestingly, predictor BRANN 2 includes RDF descriptors calculated at radii ranging from 5.5 to 14.0 Å (Table 3), very much similar to the RDF descriptors in the linear model MRA 1. However, a noteworthy but main difference arises. Among the seven descriptors in the non-linear model, one is weighted by the van der Waals volume, while the other six descriptors appear to be weighted just by atomic masses. This result contrasts with the fact that the linear model includes contributions of all the atomic properties here tested. The

neural network is able to establish a reliable non-linear dependence between descriptors just encoding the size and shape of the studied molecules and their inhibitory activities. This fact strongly suggests that the main features controlling the inhibition of FT are the molecular size and shape of the inhibitor rather than the electronegativity and/or the polarizability properties of the whole molecule or a specific substituent.

To put some light on the meaning of RDF descriptors, here found relevant for non-linear modeling of the FT inhibition, Figure 4 shows an approximate representation of the six mass-weighted descriptors for the most (Fig. 4A) and less (Fig. 4B) active compounds in our data set. As can be observed, the main difference between these compounds is the mass distribution in the outer parts of the molecule, specifically at radii ranging from 12.0 to 14.0 Å. This result well agrees with docking studies on non-thiol FTIs in which the ability of the hydrophobic outsider substituents of the inhibitor for properly matching into a complex multi-binding pocket is the main interaction governing its inhibitory efficiency.^{6,31}

In a general way, our results support previous reports on FT inhibition showing that the ability for coordinating the zinc ion in the enzyme active site is relegated to a secondary place according to our non-linear model. In this regard, the occurrence of a favorable hydrophobic matching on the active site cavity is more important than the interaction with the metal ion for displaying and enhanced inhibitory activity.

2.5. Concluding remarks

Since biological phenomena are complex by nature, in this work the inhibition of FT enzyme by a set of 78 thiol and non-thiol inhibitors was successfully modeled using a hybrid approach that combines GA and ANNs. An MRA approach is unable to successfully solve the modeled activity. However, RDF descriptors have been demonstrated to encode relevant non-linear structural information. In this sense, ANN was able to describe about 87% inhibitory activity variance with a relevant predictive power measured by q^2 values of LOO and LGO cross-validation of about 0.7. According to their activity levels, thiol and non-thiol FTIs were well-distributed in a topological map built with the seven inputs of the optimum ANN. Furthermore, the appearance in the neural network predictor of RDF descriptors weighted by mass and van der Waals volume suggested the occurrence of a strong dependence of the FT inhibition on the molecular shape and size rather than from electronegativity or polarizability characteristics of the studied compounds.

3. Materials and methods

3.1. The radial distribution function approach

The 3D coordinates of the atoms of molecules can be transformed into a structure code that has a fixed number of descriptors, irrespective of the size of a molecule. This task is performed by a structure coding technique referred

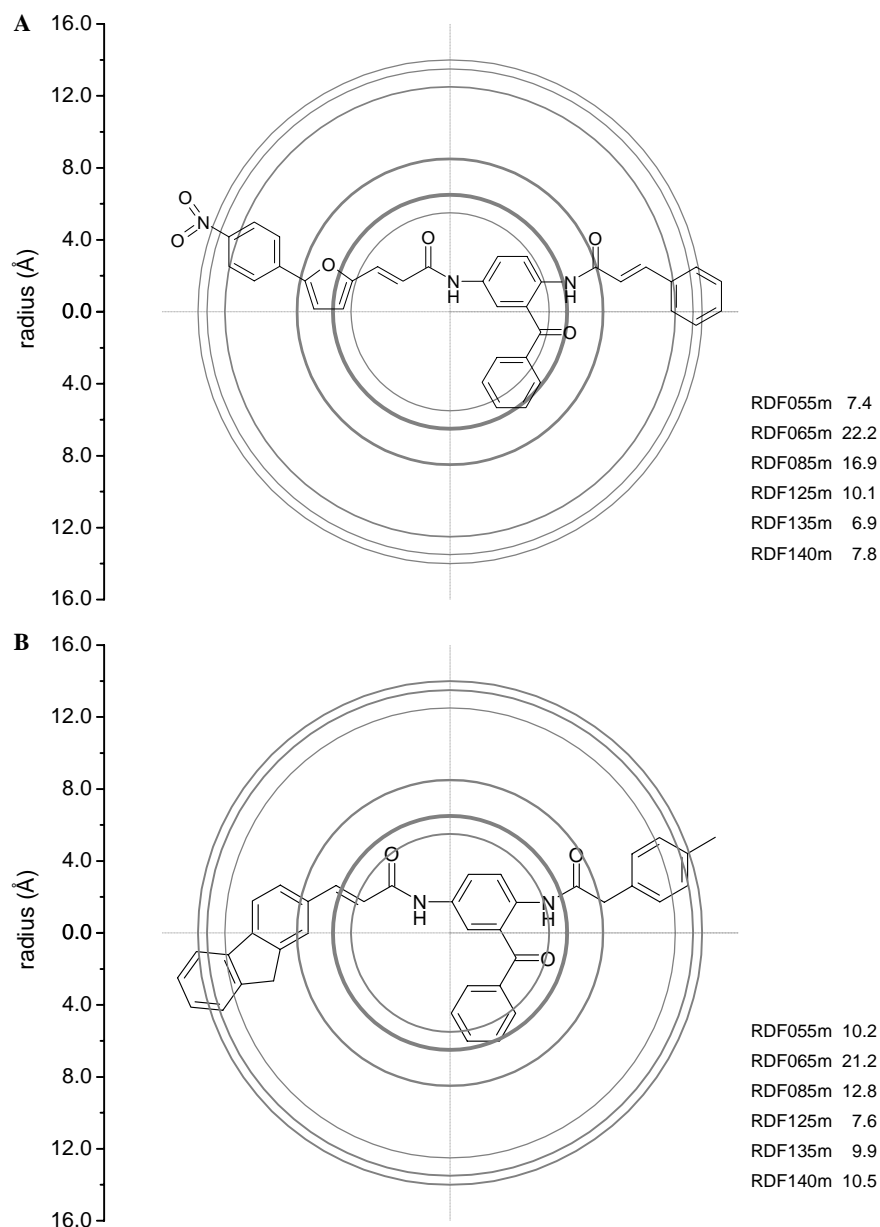


Figure 4. Graphical representation of the six mass-weighted RDF descriptors of the optimum predictor BRANN 2 for the highest (A) and the lowest (B) active FTIs. Thickness of the circle line represents the relative values of mass distributions at 5.5, 6.5, 8.5, 12.5, 13.5, and 14.0 Å according to the calculated RDF descriptors.

to as radial distribution function code (RDF code).^{24,25} In general, there are some prerequisites for a structural code:

- independence from the number of atoms, that is, the size of a molecule,
- unambiguity regarding the three-dimensional arrangement of the atoms, and
- invariance against translation and rotation of the entire molecule.

Formally, the radial distribution function of an ensemble of N atoms can be interpreted as the probability distribution to find an atom in a spherical volume of radius r .²⁶ The equation represents the radial distribution function code as it is used in this investigation:

$$g(r) = f \sum_{i=1}^{N-1} \sum_{j>1}^N A_i A_j e^{-B(r-r_{ij})^2}, \quad (5)$$

where f is a scaling factor and N is the number of atoms. By including characteristic atomic properties A of the atoms i and j , the RDF codes can be used in different tasks to fit the requirements of the information to be represented. The exponential term contains the distance r_{ij} between the atoms i and j , and the smoothing parameter B that defines the probability distribution of the individual distances. $g(r)$ was calculated at a number of discrete points with defined intervals. The atomic properties A_i and A_j used in this equation enable the discrimination of the atoms of a molecule for almost any property that can be attributed to an atom. Such distri-

bution function provides, besides information about interatomic distances in a whole molecule, the opportunity to gain access to other valuable information, for example, bond distances, ring types, planar and non-planar systems, and atom types. This fact is the most valuable consideration for a computer-assisted code elucidation. The radial distribution function in this form meets the entire requirement mentioned above, especially invariance against linear translations.

3.2. Data set and molecular descriptors

In the present study, a data set of 78 FTIs for which their activities are reported in the literature^{2,6,31,39} was used. In such reports, the inhibitory activity was determined using the fluorescence enhancement assay as described by Pompliano et al.⁴⁰ The assay employed yeast FT fused to glutathione *S*-transferase at the N-terminus of the β -subunit.⁴⁰ Farnesylpyrophosphate and the dansylated pentapeptide Ds-GlyCysValLeuSer were used as substrates. On farnesylation of the cysteine thiol, the dansyl residue is placed in a lipophilic environment, resulting in an enhancement of fluorescence at 505 nm used to monitor the enzyme reaction. Molecular structures, numbering of the substituents, and activities of the FTIs are given in Table 1. IC₅₀ refers to the millimolar concentration of the compound required for 50% inhibition of the enzyme activity.

In this way, we carry out geometry optimization calculations for each compound of this study using the quantum chemical semi-empirical method PM3⁴¹ included in Mopac 6.0 computer software.⁴² Dragon⁴³ computer software was employed to calculate the RDF molecular descriptors at radius ranging from 1.0 to 15.5 Å with radius increments of 0.5 Å. As weighting properties we tried all available properties in the Dragon software (atomic masses, atomic van der Waals volumes, atomic Sanderson electronegativities and atomic polarizabilities) in such a way that a total of 150 descriptors were computed.^{25,26}

3.3. Variable selections

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.^{17–23}

Since 150 RDF molecular descriptors were available for the QSAR analysis and only a subset of them is statistically significant in terms of correlation with biological activities, deriving an optimal QSAR model through variable selection needs to be addressed. In this sense, linear and non-linear GA searches were carried out for building the optimum linear and non-linear models. GAs are stochastic optimization methods that have been inspired by evolutionary principles. The distinctive aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores different regions in parameter space.¹⁸ The first step is to create a population of models. These models mate with each other, mutate,

crossover, reproduce, and then evolve through successive generations toward an optimum solution.

The GA implemented in this paper is a version of the So and Karplus algorithm¹⁷ that was previously reported by us²³ and was programmed within the Matlab environment⁴⁴ using Genetic Algorithm and Neural Networks Tool Boxes.^{45,46}

3.4. Regularized-artificial neural networks

In contrast to common statistical methods, Artificial Neural Networks (ANNs) are not restricted to linear correlations or linear subspaces.⁴⁷ They take into account non-linear structures and structures of arbitrarily shaped clusters or curved manifolds. 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.⁴⁸ As biological phenomena are considered non-linear by nature, the ANN technique has been successfully applied to discover the possible existence of non-linear relationships between biological activity and molecular descriptors that are ignored for the linear approach.⁴⁹

Typically, neural network training aims to reduce the mean square errors of the network $F = \text{MSE}$. Regularization involves modifying the performance function, usually known as cost function (F). It is possible to improve generalization if an additional term is added:

$$F = \beta \times \text{MSE} + \alpha \times \text{MSW}, \quad (6)$$

where MSW is the sum of squares of the network weights and biases, and α and β are objective function parameters. The relative size of the objective function parameters dictates the emphasis for training, getting a smoother network response. MacKay's bayesian regularization automatically sets the correct values for the objective function parameters,⁵⁰ in this sense the regularization is optimized.

Bayesian regularization overcomes the remaining deficiencies of neural networks.⁵⁰ By using Bayesian regularization, robust models well matched to the data and able to make accurate predictions can be obtained. Since the algorithm automatically regularizes the training process usually non validation set recurs so all the data set can be devoted to train the network.³² The Bayesian neural net has the potential to give models that are relatively independent of neural network architecture, above a minimum architecture, and the Bayesian regularization method estimates the number of effective parameters. Bayesian regularized ANNs (BRANNs) have been successfully used on the QSAR studies of biological activities and specifically in drug discovery.^{32,51,52}

Our BRANNs are classical back-propagation neural nets that incorporate the Bayesian regularization algorithm for finding the optimum weights. The Bayesian

regularization takes place within the Levenberg–Marquardt algorithm⁴⁶ implemented in Matlab environment. The input and output values were normalized prior network training.

3.5. Self-organizing maps

To settle structural similarities among the FTIs a Kohonen self organizing map (KSOM) was built. Kohonen⁵³ introduced a neural network model that generates KSOMs. In such maps, molecules with similar descriptor vectors are projected into the same or closely adjacent neurons.⁵⁴ These networks have been widely used for addressing structural similarities among chemical data sets.⁵⁵

In this work, KSOMs were implemented in a 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.⁴⁶

3.6. Validation of the models

Linear and non-linear models obtained were validated by calculating q^2 values. The q^2 values are calculated from ‘leave-one-out’ (LOO) and ‘leave-group-out’ (LGO) cross-validation process. One data point or a group of data points (specifically 6 data points for whole data set models and 4 data points for partial data set models) are removed from the data set and the model recalculated; the predicted values for those points are then compared to its experimental values. 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 . The q^2 values can be considered a measure of the predictive power of a regression equation: whereas R^2 can always be increased artificially by adding more parameters (descriptors), q^2 decreases if a model is overparameterized,²⁸ and is therefore a more meaningful summary statistic for QSAR models.

Acknowledgments

The authors would like to acknowledge to Professor Martin Schlitzer and Professor David Winkler for providing valuable information regarding FTIs and BRANNs, respectively. The useful comments of the anonymous referees that greatly helped to improve the quality of the manuscript are also gratefully acknowledged.

References and notes

- Sebti, S. M.; Hamilton, A. D. *Drug Discovery Today* **1998**, 3, 26.
- Sakowski, J.; Böhm, M.; Sattler, I.; Dahse, H.-M.; Schlitzer, M. *J. Med. Chem.* **2001**, 44, 2886.
- Zhang, F. L.; Casey, P. J. *Annu. Rev. Biochem.* **1996**, 65, 241.
- Du, W.; Lebowitz, P. F.; Prendergast, G. C. *Mol. Cell Biol.* **1999**, 19, 183.
- Prendergast, G. C. *Curr. Opin. Cell Biol.* **2000**, 12, 166.
- Sakowski, J.; Sattler, I.; Schlitzer, M. *Bioorg. Med. Chem.* **2002**, 10, 233.
- Reynolds, J. E. F. Ed. *Martindale The Extra Pharmacopeia*, 31st ed.; Royal Pharmaceutical Society of Great Britain: London, 1996; p 821.
- Hunt, J. T.; Lee, V. G.; Leftheris, K.; Seizinger, B.; Carboni, J.; Mabius, J.; Ricca, C.; Yan, N.; Manne, V. *J. Med. Chem.* **1996**, 39, 353.
- O'Connor, S. J.; Barr, K. J.; Wang, L.; Sorensen, B. K.; Tasker, A. S.; Sham, H.; Ng, A.-C.; Cohen, J.; Devine, E.; Cherian, S.; Saeed, B.; Zhang, H.; Lee, J. Y.; Warner, R.; Tahir, S.; Kovar, P.; Ewing, P.; Alder, J.; Mitten, M.; Leal, J.; Marsh, K.; Bauch, J.; Hoffman, D. J.; Sebti, S. M.; Rosenberg, S. H. *J. Med. Chem.* **1999**, 42, 3701.
- Augeri, D. J.; Janowick, D.; Kalvin, D.; Sullivan, G.; Larsen, J.; Dickman, D.; Ding, H.; Cohen, J.; Lee, J.; Warner, R.; Kovar, P.; Cherian, S.; Saeed, B.; Zhang, H.; Tahir, S.; Ng, S.-C.; Sham, H.; Rosenberg, S. H. *Bioorg. Med. Chem. Lett.* **1999**, 9, 1069.
- Breslin, M. J.; deSolms, J.; Giuliani, E. A.; Stokker, G. E.; Graham, S. L.; Pompliano, D. L.; Mosser, S. D.; Hamilton, K. A.; Hutchinson, J. H. *Bioorg. Med. Chem. Lett.* **1998**, 8, 3311.
- Ciccarone, T. M.; MacTough, S. C.; Williams, T. M.; Dinsmore, C. J.; O'Neill, T. J.; Shah, D.; Culberson, J. C.; Koblan, K. S.; Kohl, N. E.; Gibbs, J. B.; Oliff, A. I.; Graham, S. L.; Hartman, G. D. *Bioorg. Med. Chem. Lett.* **1999**, 9, 1991.
- González, M. P.; Morales, A. H. *J. Comput. Aided Mol. Des.* **2003**, 17, 665.
- González, M. P.; Morales, A. H.; González-Díaz, H. A. TOPS-MODE approach to predict permeability coefficients. *Polymer* **2004**, 45, 2073.
- González, M. P.; Terán, C. A TOPS-MODE approach to predict adenosine kinase inhibition. *Bioorg. Med. Chem. Lett.* **2004**, 14, 3077.
- Fernández, M.; Caballero, J.; Morales, A. H.; Castro, E. A.; González, M. P. *Bioorg. Med. Chem.* **2005**, 13, 3269.
- So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, 39, 1521.
- So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, 39, 5246.
- Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1328.
- Takahata, Y.; Costa, M. C. A.; Gaudio, A. C. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 540.
- Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Nesari, N. *J. Chem. Inf. Model.* **2005**, 45, 190.
- González, M. P.; Caballero, J.; Garriga, M.; González, G.; Morales, A. H.; Fernández, M. *Bull. Math. Biol.*, doi:10.1007/S00894-005-0014-X.
- Caballero, J.; Fernández, M. *J. Mol. Mod.* doi:10.1007/S00894-005-0014-X.
- Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 1030.
- Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V.; Fresenius, J. *Anal. Chem.* **1997**, 359, 50.
- Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *Vib. Spectrosc.* **1999**, 19, 151.
- Strickland, C. L.; Windsor, W. T.; Syto, R.; Wang, L.; Bond, R.; Wu, R.; Schwartz, J.; Le, H. V.; Beese, L. S.; Weber, P. C. *Biochemistry* **1998**, 37, 16601.
- Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 1.
- Lipnick, R. L. *Sci. Total Environ.* **1991**, 109, 131.

30. Devillers, J.; Lipnick, R. L. Practical applications of regression analysis in environmental QSAR studies. In *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and toxicology*; Karcher, K., Devillers, J., Eds.; Kluwer: Dordrecht, 1990, pp 129–143.
31. Mitsch, A.; Böhm, M.; Wißner, P.; Sattler, I.; Schlitzer, M. *Bioorg. Med. Chem.* **2002**, *10*, 2657.
32. Burden, F. R.; Winkler, D. A. *J. Med. Chem.* **1999**, *42*, 3183.
33. Polley, M. J.; Winkler, D. A.; Burden, F. R. *J. Med. Chem.* **2004**, *47*, 6230.
34. Giraud, E.; Luttmann, C.; Lavelle, F.; Riou, J.-F.; Mailliet, P.; Laoui, A. *J. Med. Chem.* **2000**, *43*, 1807.
35. Wan, S.; Yi, X.; Guo, Z. *Yaoxue Xuebao* **2002**, *37*, 257.
36. Wan, S.; Yi, X.; Guo, Z. *Yaoxue Xuebao* **2001**, *36*, 423.
37. Sung, N.-D.; Yu, S.-J.; Myung, P.-K.; Kwon, B.-M. *Han'guk Nonghwa Hakhoechi* **2000**, *43*, 95.
38. Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. *J. Med. Chem.* **2000**, *43*, 1975.
39. Kettler, K.; Sakowski, J.; Silber, K.; Sattler, I.; Klebe, G.; Schlitzer, M. *Bioorg. Med. Chem.* **2003**, *11*, 1521.
40. Pompliano, D. L.; Gomez, R. P.; Anthony, N. J. *J. Am. Chem. Soc.* **1992**, *114*, 7945.
41. Stewart, J. J. P. *J. Comp. Chem.* **1989**, *10*, 210.
42. MOPAC version 6.0. Frank J. Seiler Research Laboratory, US Air Force Academy, Colorado Springs, CO, 1993.
43. Todeschini, R.; Consonni, V.; Pavan, M. **(2002)** Dragon Software version 2.1.
44. Matlab version 7.0. The MathWorks, Inc., 2004.
45. The MathWorks Inc. Genetic algorithm and direct search toolbox user's guide for use with MATLAB, The Mathworks Inc., Massachusetts, 2004.
46. The MathWorks Inc. Neural network toolbox user's guide for use with MATLAB, The Mathworks Inc., Massachusetts, 2004.
47. Sumpter, B. G.; Getino, C.; Noid, D. W. *Annu. Rev. Phys. Chem.* **1994**, *45*, 439.
48. Kolmogorov, A. N. *Dokl. Akad. Nauk SSSR* **1957**, *114*, 953.
49. Winkler, D. A. *Mol. Biotech.* **2004**, *27*, 139.
50. Mackay, D. J. C. *Neural Comput.* **1992**, *4*, 415.
51. Burden, F. R.; Winkler, D. A. *Chem. Res. Toxicol.* **2000**, *13*, 436.
52. Winkler, D. A.; Burden, F. R. *Biosilico* **2004**, *2*, 104.
53. Kohonen, T. *Biol. Cybern.* **1982**, *43*, 59.
54. Gasteiger, J.; Zupan, J. *Angew. Chem., Int. Ed. Engl.* **1995**, *32*, 503.
55. Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, 1999.