Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 14 (2006) 4137–4150
Bioorganic & Medicinal Chemistry

Linear and nonlinear QSAR study of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives as matrix metalloproteinase inhibitors

Michael Fernández,[a] Julio Caballero[a,*] and Alain Tundidor-Camba[b]

[a]*Molecular Modeling Group, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba*
[b]*Scientific Prospection Group, National Center for Scientific Researches (CNIC), PO Box 6880, Havana, Cuba*

Received 30 October 2005; revised 26 January 2006; accepted 30 January 2006
Available online 28 February 2006

Abstract—The inhibitory activity (IC_{50}) toward matrix metalloproteinases (MMP-1, MMP-2, MMP-3, MMP-9, and MMP-13) of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives (HPSAAs) has been successfully modeled using 2D autocorrelation descriptors. The relevant molecular descriptors were selected by linear and nonlinear genetic algorithm (GA) feature selection using multiple linear regression (MLR) and Bayesian-regularized neural network (BRANN) approaches, respectively. The quality of the models was evaluated by means of cross-validation experiments and the best results correspond to nonlinear ones ($Q^2 > 0.7$ for all models). Despite the high correlation between the studied compound IC_{50} values, the 2D autocorrelation space brings different descriptors for each MMP inhibition. On the basis of these results, these models contain useful molecular information about the ligand specificity for MMP S'_1, S_1, and S'_2 pockets.

1. Introduction

Matrix metalloproteinases (MMPs) constitute a family of structurally related zinc-containing endopeptidases that are involved in the degradation of the macromolecular components in the extracellular matrix (ECM) of connective tissue.[1] Normally, the homeostasis of MMPs is maintained by tissue inhibitors, but in cancer progression, control over MMP activity is lost.[2] Indeed, MMPs have been identified as factors that promote tumor progression, since MMP family members collectively degrade all structural components of the ECM, which leads to the destruction of the matrix barriers surrounding the tumor, permitting invasion into surrounding connective tissues, entry and exit from blood vessels, and metastasis to distant organs.[3]

The MMP family includes fibroblast collagenase (MMP-1), gelatinase A (MMP-2), stromelysin-1 (MMP-3), matrilysin (MMP-7), neutrophil collagenase (MMP-8), gelatinase B (MMP-9), stromelysin-2 (MMP-10), stromelysin-3 (MMP-11), collagenase-3 (MMP-13), and several membrane-type MMPs. Members of the MMP family (with the exception of MMP-7) share structural features including propeptide, catalytic, and hemopexin domains.

The MMP inhibitors have been considered as potential therapeutics for cancer and other diseases.[4] A broad-spectrum of peptidic or nonpeptidic structures bearing a zinc-binding ligand (e.g., carboxylic or hydroxamic acids) have been reported in recent years,[5–9] some of which reached an advanced clinical trial. It has been identified that MMP family members contribute to different stages of tumor progression, therefore, the design of selective MMP inhibitors is desired. Since MMP family members share the main structural characteristics, the design of selective MMP inhibitors is a difficult task. The establishment of the differences in the inhibitors has been approached by exploring the differences in the MMP active sites. Recently, the number of available high-resolution X-ray crystal structures of MMP-inhibitor complexes has dramatically increased. This structural information has become an important tool in designing selective potential inhibitors. Besides, the availability of combinatorial chemistry and computa-

Keywords: QSAR analysis; MMP inhibitors; Bayesian-regularized Genetic Neural Networks; 2D Autocorrelation space.
* Corresponding author. Tel.: +53 45 26 1251; fax: +53 45 25 3101; e-mail: jmcr77@yahoo.com

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.01.072

tional resources has accelerated the drug design process. Molecular dynamics and docking-type techniques have helped to explore the structural differences of MMP family members and their interactions with MMP inhibitors.[10] In addition, quantitative structure–activity relationship (QSAR) studies have been successfully applied for modeling activities of MMP inhibitors.[11–15]

In this work, we carried out a QSAR study using 2D topological information. 2D Autocorrelation pool was used for encoding structural information from a set of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives (HPSAAs) (the chemical structures are shown in Table 1) and the relevant information that relates the topological features of these compounds with their inhibitory activities against several MMP family members (MMP-1, MMP-2, MMP-3, MMP-9, and MMP-13) was extracted by linear and nonlinear genetic algorithm (GA) feature selection (logarithmic experimental activities ($\log(10^6/IC_{50})$) are shown in Table 2). The results were analyzed in order to assess the 2D autocorrelation subspaces able to establish HPSAA molecular structure–MMP inhibition relationships.

2. Theory

2.1. Genetic algorithm

Genetic algorithms (GAs) are governed by biological evolution rules.[16] They 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.[17] The first step is to create a population of N individuals. Each individual encodes the same number of randomly chosen descriptors. The fitness of each individual in this generation is determined. In the second step, a fraction of children of the next generation is produced by crossover (crossover children) and the rest by mutation (mutation children) from the parents on the basis of their scaled fitness scores. The new offspring contains characteristics from two or one of its parents. GA method is especially useful when the finding of global optimum is harmed by the presence of many local optima in a complex response hypersurface. GA succeeds whilst the presence of local optima makes direct optimization methods unreliable and the exhaustive search is impossible for high dimensionality problems.

2.2. Bayesian-regularized Genetic Neural Networks (BRGNN)

An artificial neural network (ANN) is a layered structure consisting of computing units named neurons and connections between neurons named synapses.[18] Neurons of the input layer are associated with independent variables or input variables; while the output neurons are associated with response variables or output variables. Synapses are oriented connections linking neurons from the input layer to neurons of the hidden layer and neurons from the hidden layer to output neurons. The strength of the synapse from neuron i to neuron j is determined by means of a real value w_{ij}, named weight. Furthermore, each neuron j from the hidden layer, and eventually the output neuron, are associated with a real value b_j named the neuron's bias and with a nonlinear function, named the transfer or activation function.

Typically, ANN training aims to reduce the mean square errors (MSE) of the network ($F = \mathrm{MSE}$). When the number of neurons of the hidden layers tends to increase, ANNs tend to reduce F, but the network loses its ability to generalize. Network has memorized the training examples, but it has not learned to generalize to new situations, it means network overfits the data.

For overcoming the deficiencies of ANN Bayesian-regularization techniques have been successfully applied in the context of both regression and classification problems.[19] This involves modifying the performance function F. It is possible to improve generalization if an additional term is added.

$$F = \beta \times \mathrm{MSE} + \alpha \times \mathrm{MSW}, \quad (1)$$

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,[19] in this sense the regularization is optimized.

Bayesian regularization solves some of the well-known problems of back-propagated ANNs. BRANNs have the potential to give models which are relatively independent of neural network architecture above a minimum architecture. All available training data can be devoted to the model and potentially lengthy validation processes can be avoided. The Bayesian regularization estimates the number of effective parameters which are lower than the number of weights. The concerns about overfitting and overtraining are eliminated so that the production of a definitive and reproducible model is attained. In addition, they are faster than standard neural networks.

Bayesian regularization diminishes the inherent complexity of neural networks, being governed by Occam's Razor;[20] in this sense, it produces predictors that are robust and well matched to the data. For this reason, BRANNs are considered as accurate predictors for QSAR analysis.[21,22] The joining with GA feature selection (Bayesian-regularized Genetic Neural Networks: BRGNN) increases the possibilities of BRANNs for modeling as we indicated in previous works.[23–26]

3. Results and discussion

The selective inhibition of MMP family members is an arduous task since the topology and the nature of the residues in the enzyme active site are highly conserved

among the different MMPs. Correlation matrix (Table 3) shows that inhibitory activities of HPSAAs employed in this study against five MMP family members are highly related to each other. More in detail, the activities against collagenases (MMP-1 and MMP-13) and gelatinases (MMP-2 and MMP-9) correlate to a large extend ($R^2 > 0.7$), however, they are less related to activities against stromelysin- 1 (MMP-3) ($R^2 < 0.7$). In other respect, the activities against MMP-1 are poor for these compounds ($\log(10^6/IC_{50})$ between 1.5 and 4.8) as was indicated for the authors.[7–9]

3.1. Multiple linear regression approach

Linear correlations were developed by means of MLR models for inhibitory activities of HPSAAs against five MMP family members with acceptable statistical significances and predictive power (Eqs. 2–6).

MLR-MMP-1:

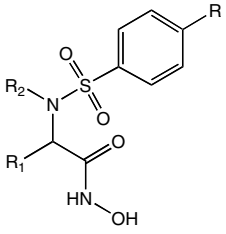
$$\log(10^6/IC_{50}) = -166.804 \times MATS4m - 51.519 \times MATS8m - 13.020 \times MATS3v + 4.817 \times GATS1e - 6.913 \times GATS2e + 219.442 \quad (2)$$

$$N = 26; \quad R^2 = 0.834; \quad S = 0.383; \quad p < 10^{-5}$$

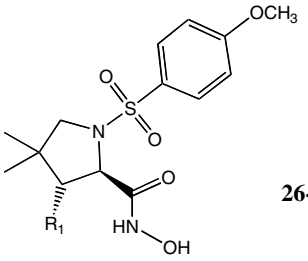
$$Q^2_{LOO} = 0.745; \quad S_{CV\,LOO} = 0.421;$$

$$Q^2_{L3O} = 0.708; \quad S_{CV\,L3O} = 0.455$$

Table 1. Structural features of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives (HPSAAs)



1-21 **R=MeO-**
22,23 **R=Ph**
24 **R=Br**
25 **R=PhO-**



26-32

Compound[a]	R_1	R_2
1	3-Pyridinylmethyl	Isobutyl
2	2-(Benzylsulfanyl)ethyl	2-(Benzhydrylamino)-2-oxoethyl
3	2-(Benzylsulfanyl)ethyl	Isobutyl
4	2-[([1,1′-Biphenyl]-4-ylmethyl)sulfanyl]ethyl	Isobutyl
5	2-{[4-(Benzyloxy)benzyl]sulfanyl}ethyl	Isobutyl
6	Ethyl	Isobutyl
7	2-(Benzylsulfanyl)ethyl	2-(Benzylamino)-2-oxoethyl
8	2-(Benzylsulfanyl)ethyl	2-Oxo-2-[(3-pyridinylmethyl)amino]ethyl
9	2-(Benzylsulfanyl)ethyl	2-[(Cyclohexylmethyl)amino]-2-oxoethyl
10	2-(Benzylsulfanyl)ethyl	2-{[Di(2-pyridinyl)methyl]amino}-2-oxoethyl
11	2-(Benzylsulfanyl)ethyl	2-[(Dicyclohexylmethyl)amino]-2-oxoethyl
12	2-(Benzylsulfanyl)ethyl	2-[(4-Methoxybenzyl)amino]-2-oxoethyl
13	2-(Benzylsulfanyl)ethyl	2-[(3,5-Dimethoxybenzyl)amino]-2-oxoethyl
14	2-(Benzylsulfanyl)ethyl	2-Oxo-2-[(2,4,6-trimethoxybenzyl)amino]ethyl
15	2-(Benzylsulfanyl)ethyl	2-(Cyclohexylamino)-2-oxoethyl
16	2-(Benzylsulfanyl)ethyl	2-(4-Morpholinyl)-2-oxoethyl
17	2-(Benzylsulfonyl)ethyl	Isobutyl
18	2-[(3-Methoxybenzyl)sulfanyl]ethyl	Isobutyl
19	2-[(3-Pyridinylmethyl)sulfanyl]ethyl	Isobutyl
20	2-[(3-Thienylmethyl)sulfanyl]ethyl	Isobutyl
21	2-[(2,3,4,5,6-Pentafluorobenzyl)sulfanyl]ethyl	Isobutyl
22	2-(Methylsulfanyl)ethyl	Isobutyl
23	2-(Benzylsulfanyl)ethyl	Isobutyl
24	2-(Benzylsulfanyl)ethyl	Isobutyl
25	2-(Benzylsulfanyl)ethyl	Isobutyl
26	(Benzylsulfanyl)methyl	—
27	Vinyl	—
28	Hydroxymethyl	—
29	[(2-Phenylethyl)sulfanyl]methyl	—
30	[(4-Methoxybenzyl)sulfanyl]methyl	—
31	(Benzyloxy)methyl	—
32[b]	1-Hydroxy-2-(phenylsulfanyl)ethyl	—

[a] Compounds **1**–**6**, **22**, and **23** are from Ref. 7; **26**–**32** are from Ref. 8, and **7**–**21**, **24** and **25** are from Ref. 9.
[b] Racemic pair.

Table 2. Experimental and predicted $\log(10^6/IC_{50})$ by linear and nonlinear models

Compound	MMP-1					MMP-2					MMP-3					MMP-9					MMP-13				
	Exp	Predictions		Rel errors		Exp	Predictions		Rel errors		Exp	Predictions		Rel errors		Exp	Predictions		Rel errors		Exp	Predictions		Rel errors	
		Lin	Nonlin	Lin	Nonlin		Lin	Nonlin	Lin	Nonlin		Lin	Nonlin	Lin	Nonlin		Lin	Nonlin	Lin	Nonlin		Lin	Nonlin	Lin	Nonlin
1	4.02	3.43	4.10	0.15	0.02	4.82	4.62	4.68	0.04	0.03	4.85	4.56	4.72	0.06	0.03	5.00	4.82	4.94	0.04	0.01	4.92	4.47	4.53	0.09	0.08
2	—	—	—	—	—	5.64	5.83	5.65	0.03	0.00	5.43	5.60	5.46	0.03	0.01	6.30	6.29	6.16	0.00	0.02	5.35	6.02	5.62	0.13	0.05
3	3.98	4.08	3.98	0.03	0.00	6.15	5.86	5.79	0.05	0.06	6.15	5.53	5.73	0.10	0.07	5.60	6.08	6.00	0.09	0.07	4.92	5.41	5.10	0.10	0.04
4	—	—	—	—	—	4.52	5.11	5.14	0.13	0.14	4.70	4.95	4.49	0.05	0.04	6.70	6.03	6.35	0.10	0.05	—	—	—	—	—
5	—	—	—	—	—	4.31	4.96	4.64	0.15	0.08	4.54	5.31	4.85	0.17	0.07	5.10	5.41	5.28	0.06	0.04	4.70	5.09	4.65	0.08	0.01
6	3.42	3.27	3.32	0.04	0.03	4.28	4.22	4.44	0.01	0.04	4.35	4.47	4.21	0.03	0.03	5.15	5.08	4.88	0.01	0.05	4.02	4.08	4.24	0.01	0.05
7	4.29	4.95	4.69	0.15	0.09	6.15	6.27	6.16	0.02	0.00	5.80	6.03	5.89	0.04	0.02	6.70	6.85	6.49	0.02	0.03	6.30	5.92	6.12	0.06	0.03
8	4.70	4.73	4.63	0.01	0.01	5.92	6.10	5.80	0.03	0.02	5.66	6.01	5.61	0.06	0.01	6.70	6.78	6.57	0.01	0.02	5.96	5.73	5.99	0.04	0.01
9	—	—	—	—	—	6.52	6.10	6.52	0.06	0.00	—	—	—	—	—	8.00	7.20	7.47	0.10	0.07	—	—	—	—	—
10	4.41	4.48	4.50	0.02	0.02	6.40	5.69	6.22	0.11	0.03	5.96	5.36	5.82	0.10	0.02	6.70	6.52	6.87	0.03	0.03	6.40	5.66	6.14	0.12	0.04
11	—	—	—	—	—	5.05	5.00	5.54	0.01	0.10	—	—	—	—	—	5.73	6.15	6.31	0.07	0.10	—	—	—	—	—
12	4.72	4.55	4.49	0.04	0.05	6.15	5.84	6.26	0.05	0.02	6.05	5.78	5.99	0.04	0.01	6.70	6.42	6.69	0.04	0.00	6.15	5.69	6.18	0.07	0.00
13	4.44	4.26	4.40	0.04	0.01	5.80	5.77	5.61	0.01	0.03	5.33	4.95	5.26	0.07	0.01	6.22	5.76	6.25	0.07	0.00	5.64	5.66	5.65	0.00	0.00
14	4.31	4.27	4.24	0.01	0.02	5.92	5.62	5.82	0.05	0.02	5.60	5.47	5.57	0.02	0.01	6.52	6.60	6.36	0.01	0.02	5.80	5.57	5.81	0.04	0.00
15	4.80	4.37	4.55	0.09	0.05	6.00	6.19	5.97	0.03	0.01	5.80	4.83	5.65	0.17	0.03	6.70	7.10	6.88	0.06	0.03	6.10	5.73	6.03	0.06	0.01
16	3.27	3.56	3.28	0.09	0.00	4.80	5.44	4.82	0.13	0.00	3.71	4.62	3.99	0.25	0.08	4.54	5.57	4.97	0.23	0.09	4.64	5.27	4.88	0.14	0.05
17	3.40	3.41	3.48	0.00	0.02	5.37	5.27	5.70	0.02	0.06	4.99	5.41	5.13	0.08	0.03	6.05	6.41	5.85	0.06	0.03	4.93	4.81	4.82	0.02	0.02
18	3.53	3.85	3.66	0.09	0.04	5.42	5.52	5.44	0.02	0.00	5.24	4.96	5.55	0.05	0.06	5.49	5.10	5.80	0.07	0.06	5.28	5.36	5.31	0.02	0.01
19	3.90	3.71	3.92	0.05	0.01	5.20	5.64	4.98	0.08	0.04	5.21	5.24	5.36	0.01	0.03	5.52	5.77	5.20	0.05	0.06	4.79	5.10	4.51	0.06	0.06
20	3.73	3.39	3.50	0.09	0.06	5.49	5.28	5.20	0.04	0.05	5.27	5.14	5.20	0.02	0.01	6.00	5.73	5.60	0.04	0.07	5.29	5.24	5.09	0.01	0.04
21	1.71	1.60	1.72	0.06	0.01	2.74	2.78	2.91	0.01	0.06	3.15	3.03	3.12	0.04	0.01	3.83	3.75	3.88	0.02	0.01	3.39	3.37	3.80	0.01	0.12
22	3.81	3.84	3.83	0.01	0.01	6.15	5.91	6.00	0.04	0.02	4.43	4.34	4.44	0.02	0.00	6.70	6.29	6.70	0.06	0.00	5.52	5.25	5.45	0.05	0.01
23	—	—	—	—	—	5.39	5.89	5.54	0.09	0.03	4.05	4.42	4.06	0.09	0.00	5.60	6.09	5.98	0.09	0.07	5.42	5.65	5.35	0.04	0.01
24	3.79	3.37	3.66	0.11	0.03	4.80	4.79	5.22	0.00	0.09	3.82	4.01	3.98	0.05	0.04	5.37	5.59	5.54	0.04	0.03	4.85	4.99	5.09	0.03	0.05
25	2.84	3.62	2.83	0.27	0.00	5.80	5.20	5.61	0.10	0.03	5.38	5.36	5.23	0.00	0.03	6.30	5.86	6.27	0.07	0.00	5.49	5.14	5.16	0.06	0.06
26	3.70	3.62	3.68	0.02	0.01	5.79	5.66	5.21	0.02	0.10	5.17	5.21	5.00	0.01	0.03	6.05	5.86	5.97	0.03	0.01	5.26	4.91	5.07	0.07	0.04
27	1.59	2.33	1.63	0.47	0.03	3.36	3.92	3.31	0.17	0.01	3.53	3.56	3.71	0.01	0.05	3.86	4.24	3.88	0.10	0.01	3.28	3.49	3.18	0.06	0.03
28	2.06	2.28	2.09	0.11	0.01	3.84	3.65	3.78	0.05	0.02	3.87	4.07	3.91	0.05	0.01	4.30	5.10	4.46	0.19	0.04	3.66	3.74	3.87	0.02	0.06
29	3.18	3.16	3.26	0.01	0.03	4.81	5.17	5.01	0.07	0.04	5.19	5.21	5.32	0.00	0.03	5.55	6.04	5.70	0.09	0.03	4.77	4.59	4.82	0.04	0.01
30	3.34	3.51	3.46	0.05	0.04	5.34	5.46	5.07	0.02	0.05	4.78	4.81	4.59	0.01	0.04	5.36	5.24	5.61	0.02	0.05	4.65	4.86	4.99	0.05	0.07
31	2.79	2.62	2.84	0.06	0.02	4.49	4.59	4.39	0.02	0.02	5.00	4.74	4.94	0.05	0.01	5.04	4.58	4.89	0.09	0.03	4.35	4.47	4.27	0.03	0.02
32	3.73	3.46	3.72	0.07	0.00	5.20	4.63	5.22	0.11	0.00	5.13	5.05	5.28	0.02	0.03	5.74	5.29	5.65	0.08	0.02	5.09	5.04	5.04	0.01	0.01

MLR-MMP-2:

$$\log(10^6/IC_{50}) = -0.042 \times ATS2v - 110.925 \times MATS4m - 55.421 \times MATS8m - 8.431 \times MATS3v + 2.379 \times GATS6e + 5.227 \times GATS2p + 162.799 \quad (3)$$

$N = 32; \quad R^2 = 0.808; \quad S = 0.420; \quad p < 10^{-5}$

$Q^2_{LOO} = 0.721 \quad S_{CV\,LOO} = 0.454$

$Q^2_{L3O} = 0.697 \quad S_{CV\,L3O} = 0.479$

MLR-MMP-3:

$$\log(10^6/IC_{50}) = -112.340 \times MATS4m - 15.666 \times MATS3v + 8.716 \times MATS6e + 2.379 \times GATS6e - 1.457 \times GATS8e + 7.200 \times GATS2p + 104.088 \quad (4)$$

$N = 30; \quad R^2 = 0.750; \quad S = 0.431; \quad p < 10^{-5}$

$Q^2_{LOO} = 0.581; \quad S_{CV\,LOO} = 0.516;$

$Q^2_{L3O} = 0.544 \quad S_{CV\,L3O} = 0.548$

MLR-MMP-9:

$$\log(10^6/IC_{50}) = 109.844 \times MATS2m - 51.167 \times MATS4m + 8.380 \times MATS2v + 13.240 \times MATS6e + 7.864 \times GATS6e + 13.015 \times GATS2p - 72.742 \quad (5)$$

$N = 32; \quad R^2 = 0.767; \quad S = 0.478; \quad p < 10^{-5}$

$Q^2_{LOO} = 0.644; \quad S_{CV\,LOO} = 0.544;$

$Q^2_{L3O} = 0.605; \quad S_{CV\,L3O} = 0.597$

MLR-MMP-13:

$$\log(10^6/IC_{50}) = -86.944 \times MATS4m - 33.782 \times MATS8m - 5.375 \times MATS3v + 8.188 \times MATS1p + 1.103 \times GATS5e + 122.338 \quad (6)$$

$N = 29; \quad R^2 = 0.787; \quad S = 0.397; \quad p < 10^{-5}$

$Q^2_{LOO} = 0.703; \quad S_{CV\,LOO} = 0.423;$

$Q^2_{L3O} = 0.692; \quad S_{CV\,L3O} = 0.434$

In Eqs. 2–6, N is the number of compounds included in the models, R^2 are the square of correlation coefficients, S is the standard deviation of the regressions, p is the significance of the variables in the models, Q^2_{LOO} and $S_{CV\,LOO}$ are the correlation coefficients and standard deviations of the LOO cross-validation, respectively, and Q^2_{L3O} and $S_{CV\,L3O}$ are the correlation coefficients and standard deviations of the L3O cross-validation, respectively. Plots of predicted versus experimental $\log(10^6/IC_{50})$ values for the linear models are shown in Figure 1. The correlation of each one of the descriptors in these equations with each other was calculated. The resulted correlation matrix is represented in Table 4. In general, there is no significant intercorrelation between the linear GA selected descriptors. We found some colinear pairs ($R^2 > 0.7$): MATS2v-GATS2p (MLR-MMP-9), GATS1e-GATS2e (MLR-MMP-1), and ATS2v-MATS8m (MLR-MMP-2).

The orthogonalization process proposed by Randić[27] was carried out for statistical interpretation of the models containing correlated indices. The idea of orthogonalization is to remove effects of linear interdependence of the various descriptors to be used in setting up experimental correlations. The high correlation among different descriptors obscures the real effect of them since it results in highly unstable regression coefficients. Randić's method of orthogonalization allows establishing the real relative importance of a descriptor without statistical changes in the model. The results for models containing colinear pairs (MLR-MMP-1, MLR-MMP-2, and MLR-MMP-9) are shown in Eqs. 7–9:

Table 3. Correlation matrix for the *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide activities against MMP family members

	MMP-1	MMP-2	MMP-3	MMP-9	MMP-13
MMP-1	1				
MMP-2	0.724	1			
MMP-3	0.592	0.685	1		
MMP-9	0.703	0.748	0.604	1	
MMP-13	0.777	0.851	0.612	0.873	1

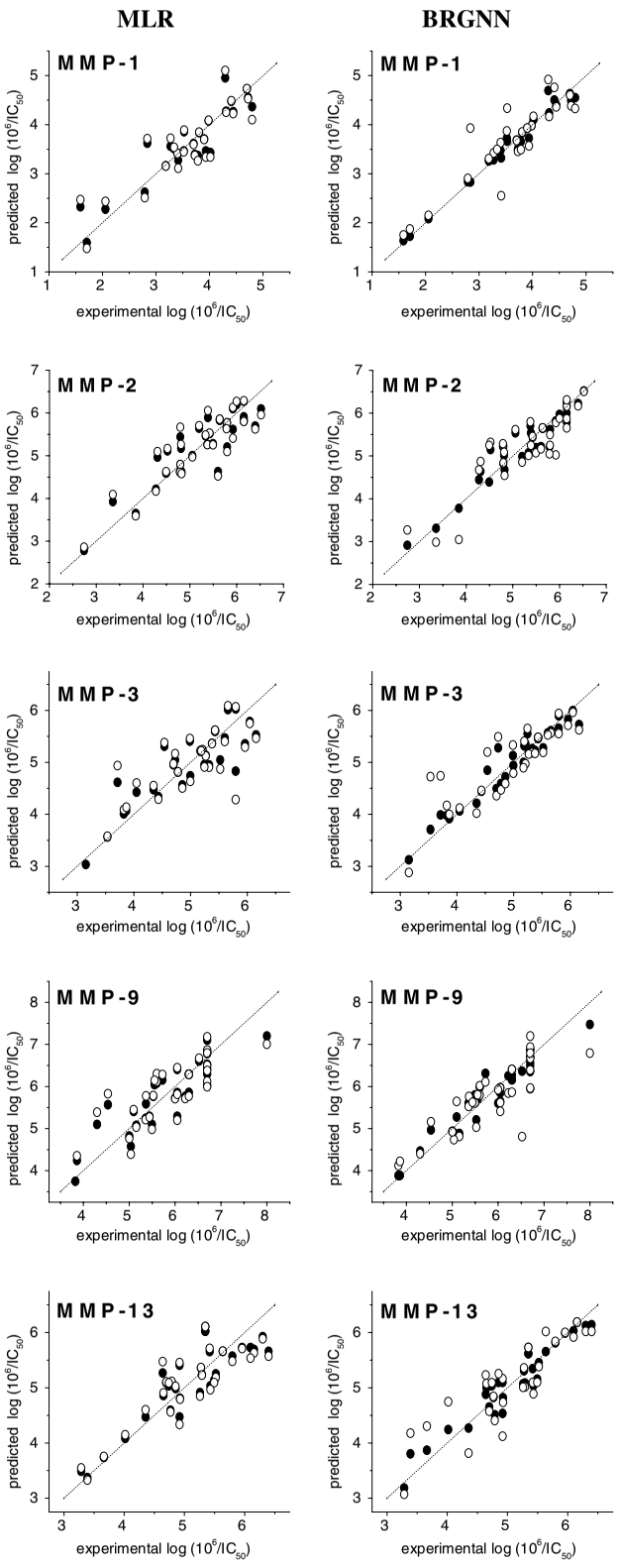


Figure 1. Plot of predicted versus experimental $\log(10^6/IC_{50})$ values for MMP inhibition by *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives using linear (left) and nonlinear (right) models. (•) Training predictions; (○) LOO cross-validation predictions.

Table 4. Correlation matrix for descriptors selected by linear GA[a]

	ATS2v	MATS2m	MATS4m	MATS8m	MATS2v	MATS3v	MATS6e	MATS1p	GATS1e	GATS2e	GATS5e	GATS6e	GATS8e	GATS2p
ATS2v	1													
MATS2m	0.209	1												
MATS4m	0.039	0.001	1											
MATS8m	**0.709**	0.228	0.010	1										
MATS2v	0.418	0.127	0.092	0.440	1									
MATS3v	0.212	0.008	0.564	0.168	0.538	1								
MATS6e	0.372	0.061	0.153	0.563	0.391	0.316	1							
MATS1p	0.212	0.186	0.121	0.196	0.247	0.001	0.009	1						
GATS1e	0.013	0.010	0.036	0.010	0.179	0.019	0.075	0.117	1					
GATS2e	0.007	0.019	0.115	0.005	0.191	0.010	0.046	0.309	**0.720**	1				
GATS5e	0.051	0.104	0.011	0.019	0.069	0.016	0.015	0.018	0.243	0.228	1			
GATS6e	0.097	0.127	0.242	0.093	0.168	0.386	0.422	0.033	0.049	0.094	0.151	1		
GATS8e	0.334	0.370	0.129	0.563	0.537	0.350	0.447	0.083	0.025	0.018	0.036	0.049	1	
GATS2p	0.420	0.160	0.090	0.408	**0.836**	0.503	0.268	0.292	0.165	0.131	0.009	0.178	0.508	1

[a] High correlations ($R^2 > 0.7$) are indicated in boldface.

MLR-MMP-1:

$$\log(10^6/IC_{50}) = -79.680 \times O[MATS4m] - 61.543 \times O[MATS8m] - 5.561 \times O[MATS3v] + 4.817 \times O[GATS1e] - 4.981 \times O[GATS2e] + 65.454 \quad (7)$$

MLR-MMP-2:

$$\log(10^6/IC_{50}) = -0.049 \times O[ATS2v] - 110.925 \times O[MATS4m] - 32.131 \times O[MATS8m] - 2.739 \times O[MATS3v] + 4.427 \times O[GATS6e] + 4.716 \times O[GATS2p] + 37.443 \quad (8)$$

MLR-MMP-9:

$$\log(10^6/IC_{50}) = 91.121 \times O[MATS2m] - 96.892 \times O[MATS4m] + 10.538 \times O[MATS2v] + 13.240 \times O[MATS6e] + 6.716 \times O[GATS6e] + 6.278 \times O[GATS2p] - 0.631 \quad (9)$$

Furthermore, in Table 5 we give the values of the mean effect of each variable in these models before and after the orthogonalization process. This analysis corrected some correlation coefficients, however, the positive and negative effects were kept as in the original models.

An additional test was made to see if both members of each correlated pair are needed in significant models.[28] The results are given in Table 6. If one selected descriptor of each of the two pairs is deleted from the initial basis set of linear models, their quality declines: individual removal of GATS2e and MATS8m from MLR-MMP-1 and MLR-MMP-2, respectively, cannot be tolerated, while removal of GATS1e in MLR-MMP-1, or removal of ATS2v in MLR-MMP-2, is marginally acceptable. Furthermore, the exclusion of MATS2v or GATS2p damages to a great extent the predictive ability of MLR-MMP-9 model. Overall, it appears that each member of the correlated pairs individually provides unique information to the models.

Inhibitory activities (IC_{50}) of the HPSAAs against the MMP family members predicted by the linear models appear in Table 2. MLR models were able to explain data variance and were quite stable to the inclusion–exclusion of compounds as measured by LOO and L3O correlation coefficients ($Q^2 > 0.5$). In this sense, the less reliable model was MLR-MMP-3 according to its Q^2 value.

The high correlation between inhibitory activities was reflected in some similarities between the linear models:

Table 5. Mean effect of each variable included in models MLR-MMP-1, MLR-MMP-2, and MLR-MMP-9 before and after orthogonalization

MLR-MMP-1			MLR-MMP-2			MLR-MMP-9		
Index	No-orthog	Orthog	Index	No-orthog	Orthog	Index	No-orthog	Orthog
MATS4m	−1.273	−0.608	ATS2v	−0.425	−0.495	MATS2m	0.493	0.409
MATS8m	−0.761	−0.909	MATS4m	−0.764	−0.764	MATS4m	−0.352	−0.667
MATS3v	−0.832	−0.355	MATS8m	−0.859	−0.498	MATS2v	0.674	0.848
GATS1e	0.235	0.235	MATS3v	−0.518	−0.168	MATS6e	1.071	1.071
GATS2e	−0.691	−0.498	GATS6e	0.365	0.679	GATS6e	1.206	1.030
			GATS2p	0.478	0.431	GATS2p	1.191	0.574

Table 6. R^2 values and LOO cross-validation analysis (Q^2 and S_{CV}) for models generated by excluding correlated descriptors

Model	Variable excluded	Total no. of variables in model	R^2	Q^2	S_{CV}
MLR-MMP-1	None	5	0.834	0.745	0.421
MLR-MMP-1	GATS1e	4	0.812	0.726	0.436
MLR-MMP-1	GATS2e	4	0.702	0.537	0.572
MLR-MMP-2	None	6	0.808	0.721	0.454
MLR-MMP-2	ATS2v	5	0.745	0.638	0.516
MLR-MMP-2	MATS8m	5	0.551	0.273	0.749
MLR-MMP-9	None	6	0.767	0.644	0.544
MLR-MMP-9	MATS2v	5	0.711	0.553	0.628
MLR-MMP-9	GATS2p	5	0.612	0.373	0.864
BRGNN-MMP-2	None	6	0.908	0.742	0.441
BRGNN-MMP-2	MATS5v	5	0.884	0.701	0.485
BRGNN-MMP-2	MATS5p	5	0.752	0.220	0.777
BRGNN-MMP-9	None	6	0.910	0.771	0.424
BRGNN-MMP-9	MATS2v	5	0.876	0.397	0.703
BRGNN-MMP-9	GATS3v	5	0.865	0.592	0.569

- All MLR models contain MATS4m as a negative influence.
- MATS3v has a negative influence in the inhibition of MMP-1, MMP-2, MMP-3, and MMP-13.
- MATS8m has a negative influence in the inhibition of MMP-1, MMP-2, and MMP-13.
- GATS6e and GATS2p have positive influences in the inhibition of MMP-2, MMP-3, and MMP-9.

The common features expressed in the 2D autocorrelation space represent the similarities between MMPs relevant for the interaction with HPSAA inhibitors. By contrast, the peculiarities must contain information about the relevant differences for selective MMP inhibitions. In this sense, the more showy dissimilarities are summed up:

- The activities against collagenases (MMP-1 and MMP-13) coincide in MATS4m, MATS8m and MATS3v descriptors. They were differentiated by atomic Sanderson electronegativity and polarizability weighted terms. MLR-MMP-1 has short lag atomic Sanderson electronegativity weighted terms (GATS1e and GATS2e), while MLR-MMP-13 has large lag atomic Sanderson electronegativity and short lag atomic polarizability weighted terms (GATS5e and MATS1p).
- Models of activities against MMP-2, MMP-3, and MMP-9 coincide in MATS4m, GATS6e, and GATS2p descriptors. In addition, these models have short lag atomic van der Waals volume weighted terms (ATS2v and MATS3v for MLR-MMP-2, MATS3v for MLR-MMP-3, and MATS2v for MLR-MMP-9). They were mainly differentiated by atomic mass and Sanderson electronegativity weighted terms. MLR-MMP-2 has large lag atomic mass weighted term MATS8m, MLR-MMP-3 has large lag atomic Sanderson electronegativity weighted terms MATS6e and GATS8e, while MLR-MMP-9 has short lag atomic mass and large lag atomic Sanderson electronegativity weighted terms MATS2m and MATS6e.

3.2. Bayesian-regularized Genetic Neural Network approach

Despite the agreeable results found by GA combined with MLR analysis, we carried out an additional nonlinear search for exploring other possibilities. Recently, we proposed the BRGNN approach[23] which surpassed the limits of the linear solutions when modeling of inhibitory activities was performed.[24,25] This can be ascribed to the facilities of ANNs for approximating complex relations by hyperbolic tangent transfer function employment. The assistance of Bayesian-regularization brings stability and avoids overfitting effects when nonlinear GA search is developed. In our current application, ANN architectures were varied testing different quantities of neurons in hidden layers.

We explored two alternatives for selecting the best models in the GA search (see Section 5). By means of alternative A, the best models were selected according to the MSE of data fitting and posterior cross-validation. By contrast, in alternative B the validation is performed through the model development step. Alternative A is computationally much less expensive than the alternative B that required calculation of cross-validation-related neural networks. We did not find significant differences in results employing both alternatives. Nearly all the optimal solutions that were revealed using Q^2-value criterion were also discovered by alternative A for all MMP family members. Furthermore, the much less expansive alternative A leads to the same best solution (higher Q^2) for each MMP inhibitory activity.

Standard back-propagated genetic neural networks using residual error of the training set as fitness function usually yield models which are optimal for the training data but they do not have good predictive abilities.[29] For this reason, a variety of fitness functions which are proportional to the residual error of the test set,[29–32] or even the cross-validation set from the neural network simulations[29,33] are commonly reported as better options. The selection of the model based on a single validation set may cause that predictors perform well on a particular external set, but there is no guarantee that the same results may be achieved on another. In this sense, this criterion can bring guileful conclusions. For example, it can happen that several outliers, by pure coincidence, are out of the test set, in which case, the validation error will be small. Otherwise, cross-validation is a too CPU expensive process which also introduces additional instability to the general GA search. In our BRGNN approach, we expected good results using training set residual error as fitness function because of Bayesian-regularization advantages (see Section 2 and references therein).

The descriptors and statistics for BRGNN models are depicted in Table 7 and plots of predicted versus experimental $\log(10^6/IC_{50})$ values are shown in Figure 1. The best models included six variables and contain two nodes in the hidden layer. The number of optimum parameters yielded by the Bayesian regularization was 14 or 15 in all cases. BRGNN statistics reveal that neural network approaches surpass the results for linear models in regard to fitness and predictive capacity ($Q^2_{LOO} > 0.7$ in all cases). Inhibitory activities ($\log(10^6/IC_{50})$) of the HPSAAs against the MMP family members predicted by the nonlinear models appear in Table 2. Similarly to the variables selected by linear GA, there is no significant intercorrelation between the nonlinear selected descriptors for IC_{50} modeling, as it is seen in Table 8. We found some colinear pairs ($R^2 > 0.7$): MATS2v-GATS3v (BRGNN-MMP-9) and MATS5v-MATS5p (BRGNN-MMP-2). As it was done for linear models, it was verified if both members of each correlated pair are needed in significant models.[28] The results are given in Table 6. If one selected descriptor of each of the two pairs is deleted from the initial basis set of nonlinear models, their quality declines: individual removal of MATS5p from BRGNN-MMP-2 cannot be tolerated, while removal of MATS5v is marginally acceptable. Furthermore, the exclusion of MATS2v or GATS3v damages to a great extent the predictive ability

Table 7. Descriptors and statistics for BRGNN models[a]

Models	Variables	*n*	Num par	Opt par	R^2	*S*	*p*	LOO		L3O[b]	
	$\log(10^6/IC_{50})$							Q^2	S_{CV}	Q^2	S_{CV}
BRGNN-MMP1	ATS2v, MATS5m, MATS7m, GATS1v, GATS1e, GATS4p	26	17	15	0.972	0.138	$<10^{-5}$	0.792	0.390	0.719	0.462
BRGNN-MMP2	MATS5m, MATS5v, MATS5p, GATS4v, GATS7v, GATS7p	32	17	14	0.908	0.262	$<10^{-5}$	0.742	0.441	0.733	0.455
BRGNN-MMP3	ATS2v, MATS1m, MATS6m, MATS6e, GATS1v, GATS5v	30	17	15	0.939	0.191	$<10^{-5}$	0.721	0.404	0.688	0.430
BRGNN-MMP9	MATS6m, MATS2v, MATS1p, GATS3v, GATS7v, GATS8v	32	17	15	0.916	0.259	$<10^{-5}$	0.771	0.424	0.713	0.477
BRGNN-MMP13	ATS2v, MATS4m, MATS7v, MATS1p, MATS5p, GATS7p	29	17	14	0.922	0.219	$<10^{-5}$	0.735	0.403	0.725	0.410

[a] 6-2-1 architecture was employed in all models. Num par represents the number of neural network parameters; Opt par represents the optimum number of neural network parameters yielded by the Bayesian regularization.

[b] Average from five cross-validation experiments.

Table 8. Correlation matrix for descriptors selected by nonlinear GA[a]

	ATS2v	MATS1m	MATS4m	MATS5m	MATS6m	MATS7m	MATS2v	MATS5v	MATS7v	MATS6e	MATS1p	MATS5p	GATS1v	GATS3v	GATS4v	GATS5v	GATS7v	GATS8v	GATS1e	GATS4p	GATS7p
ATS2v	1																				
MATS1m	0.026	1																			
MATS4m	0.039	0.000	1																		
MATS5m	0.261	0.259	0.027	1																	
MATS6m	0.339	0.078	0.029	0.105	1																
MATS7m	0.229	0.009	0.097	0.152	0.478	1															
MATS2v	0.418	0.020	0.092	0.002	0.158	0.177	1														
MATS5v	0.399	0.020	0.000	0.443	0.603	0.320	0.192	1													
MATS7v	0.491	0.021	0.067	0.278	0.345	0.045	0.146	0.382	1												
MATS6e	0.372	0.015	0.153	0.208	0.337	0.579	0.391	0.439	0.108	1											
MATS1p	0.212	0.019	0.121	0.020	0.036	0.006	0.247	0.037	0.324	0.009	1										
MATS5p	0.536	0.086	0.073	0.404	0.459	0.459	0.455	**0.731**	0.312	0.636	0.085	1									
GATS1v	0.285	0.286	0.016	0.082	0.026	0.011	0.275	0.083	0.314	0.050	0.511	0.162	1								
GATS3v	0.335	0.042	0.467	0.002	0.094	0.202	**0.727**	0.096	0.023	0.359	0.040	0.442	0.097	1							
GATS4v	0.249	0.055	0.097	0.625	0.292	0.370	0.000	0.397	0.116	0.333	0.031	0.369	0.002	0.031	1						
GATS5v	0.372	0.051	0.001	0.495	0.521	0.337	0.214	**0.926**	0.373	0.457	0.062	**0.785**	0.162	0.110	0.370	1					
GATS7v	0.508	0.029	0.017	0.247	0.341	0.059	0.263	0.353	**0.929**	0.159	0.358	0.398	0.378	0.097	0.100	0.391	1				
GATS8v	0.037	0.002	0.478	0.030	0.000	0.003	0.186	0.000	0.041	0.028	0.010	0.043	0.002	0.394	0.000	0.003	0.001	1			
GATS1e	0.013	0.006	0.036	0.242	0.035	0.091	0.179	0.052	0.000	0.075	0.117	0.025	0.057	0.049	0.513	0.041	0.011	0.051	1		
GATS4p	0.273	0.103	0.377	0.367	0.234	0.372	0.070	0.321	0.061	0.506	0.047	0.472	0.005	0.254	**0.785**	0.293	0.084	0.082	0.317	1	
GATS7p	0.034	0.037	0.114	0.002	0.005	0.254	0.025	0.003	0.315	0.050	0.285	0.001	0.230	0.002	0.085	0.000	0.389	0.000	0.149	0.090	1

[a] High correlations ($R^2 > 0.7$) are indicated in boldface.

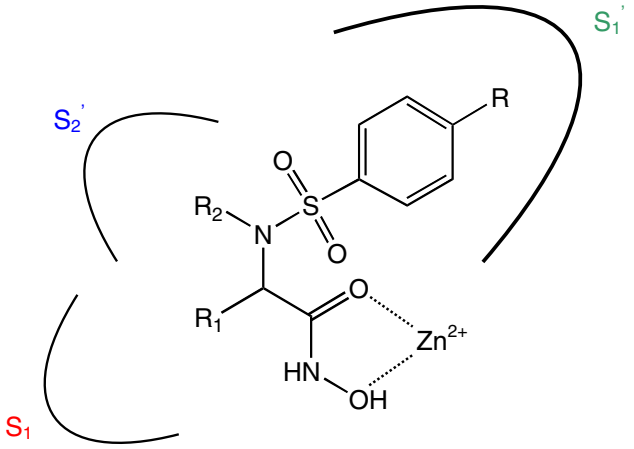
of BRGNN-MMP-9 model. Overall, it appears that each member of the correlated pairs individually provides unique information to the nonlinear model.

Despite the high correlation between pairs MATS5v-GATS5v and MATS7v-GATS7v ($R^2 > 0.9$), these descriptors do not share a common nonlinear model, therefore, they represent the same molecular information related to inhibition of different MMPs.

Unlike the models found by linear QSAR analysis, the nonlinear ones adjust the structure-relationships by quite unequal spaces. In this sense, on the basis of the great reliability that they show statistically, expressed by means of LOO and L3O experiments, we consider the nonlinear solutions as best-suited for analyzing the relevant differences for selective MMP inhibitions. The inspection of BRGNN models reveals quite a few coincidences: BRGNN-MMP-2 and BRGNN-MMP-13 models share three descriptors (MATS5p, GATS7p, and colinear pair MATS7v-GATS7v), however, BRGNN-MMP-1 and BRGNN-MMP-9 do not share any descriptor.

3.3. The linear and nonlinear evidences

By our approach, the differences between models have been interpreted as the differences between the relevant molecular information for having a certain inhibitory activity. Our QSAR study is based on previous structure–activity relationship study in which authors modified functional groups at the P'_1, P_1, and P'_2 sites (R, R_1 and R_2 in Fig. 2) of the inhibitors as functional probes for S'_1, S_1, and S'_2 subsites of MMP family members.[7–9] These authors searched for selectivity in accordance with the differences between these pockets. In consequence, the 2D autocorrelation spaces extracted by linear and nonlinear GA can be interpreted as the relevant features for the inhibitor-pocket interactions and the different spaces are caused by the differences between MMP S'_1, S_1, and S'_2 pockets.



MMP-1	SPFDGPGGNLAHAF....LHRVAAHE
MMP-2	YPFDGKDGLLAHAF....LFLVAAHE
MMP-3	YPFDGPGNVLAHAY....LFLVAAHE
MMP-9	YPFDGKDGLLAHAF....LFLVAAHE
MMP-13	YPFDG PSGLLAHAF....LFLVAAHE

Figure 2. Position of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives inside MMP active site. Comparison of amino acids sequences of MMPs. Colored letters indicate the amino acids of S'_1 (●), S_1 (●) and S'_2 (●) pockets that contribute to the ligand specificity.

The success of HPSAAs as MMP inhibitors lies in hydroxamic acid moiety chelating with the Zn^{2+} atom and sulfonamide group-related hydrogen bonds. The occupation of the pockets allows modulating the selectivity by steric, hydrophobic, and electronic differences among MMP family members. S'_1 subsite in MMPs is the most well-defined area of binding and consists of a hydrophobic pocket which varies in depth for the different MMPs.[34] MMP-1 has a characteristic Arg in its S'_1 subsite (Fig. 2). The long side chain of the Arg extends to the bottom of the S'_1 subsite and forms a rather shallow pocket. In other MMPs, this Arg is replaced by a Leu. A few mutations of amino acids occurred at the S_1 pocket. Similarly, Tyr and Phe were found in MMP-2, MMP-9, and MMP-13 (Fig. 2). Tyr is replaced by Ser in MMP-1 and Phe is replaced by Tyr in MMP-3. S'_2 subsite is essentially hydrophobic. MMP-2, MMP-9, and MMP-13 share Gly-Leu (Fig. 2), which is replaced by Gly-Asn in MMP-1 and by Asn-Val in MMP-3.[35–37] These differences, which are mainly related to the steric, hydrophobicity and electronic availability of these side chains, reflect on inhibitor binding and enzymatic activity.

In recent publications, Gupta et al.[11–15] studied the effect of several molecular properties in IC_{50} activities of MMP inhibitors using linear QSAR. This study evidenced linear dependences between molecular properties and inhibitory activities using hydrophobicity-related descriptors ($^1\chi^v$ or $\log P$) and electrotopological state (E-state) indices. As a result, authors assess the importance of electrostatic interactions and hydrophobicity as the key features which drive the inhibitor-enzyme affinities.

Unlike the previous reports, we used a varied pool of 2D autocorrelation descriptors in order to search the relevant structural information.[38] 2D Autocorrelation space can be readily derived directly from the molecular structures without any experimental effort. The computation of these descriptors involves the summations of different autocorrelation functions corresponding to the different fragment lengths and leads to different autocorrelation vectors corresponding to the lengths of the structural fragments. Bearing in mind this aspect, the interpretation of 2D autocorrelation descriptors is uneasy. On behalf of a greater applicability, physicochemical properties were inserted as weighting components. As a result, these descriptors address the topology of the structure or parts thereof in association with selected physicochemical properties. At this level, it is suitable to clarify that this analysis can not offer the specific positions of the atoms since 2D descriptors encode a global and dimension-limited information.

In order to interpret our results, we evaluated the relevance of the physicochemical properties. For this,

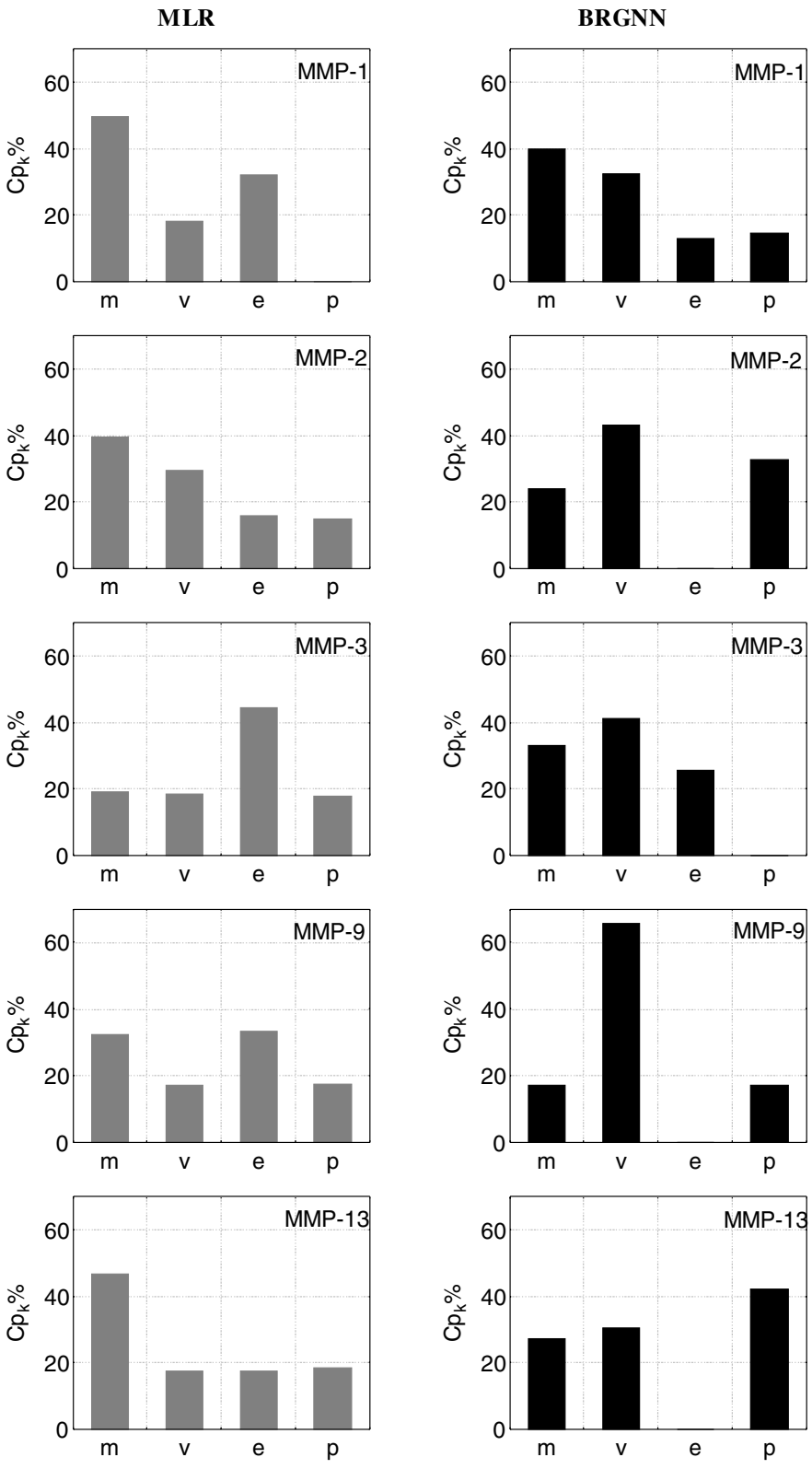


Figure 3. Contribution of atomic properties (C_{p_k}) from linear (left) and nonlinear (right) models.

we chose to estimate the relative contributions of each descriptor on the MMP inhibitory activity. The contribution of descriptors was estimated for MLR and BRGNN models. The descriptor under study was removed from the model and mean of the absolute deviation values Δm_i between the observed and estimated value for all compounds was calculated. Finally, the contribution C_i[39] of descriptor i is given by:

$$C_i = \frac{100 \times \Delta m_i}{\sum \Delta m_i} \tag{10}$$

The contributions (C_i) of descriptors weighted by the same physicochemical property are added up, in this way the contribution C_{p_k} is obtained (p_k = m, v, e, and p). The results are given in Figure 3.

The linear approach leads to the following results:

- Atomic masses have high contributions in the inhibition of all MMPs ($C_m > 30\%$), except for MMP-3.
- Inhibition of MMP-1, MMP-3, and MMP-9 is greatly influenced by atomic Sanderson electronegativities ($C_e > 30\%$).
- Inhibition of MMP-9 has the same contributions of atomic masses and Sanderson electronegativities ($C_m = 32\%$ and $C_e = 33\%$).
- Atomic Sanderson electronegativities are the key features in the inhibition of MMP-3 ($C_e = 44\%$).
- In general, the influence of atomic van der Waals volumes and polarizabilities is poor ($C_v < 30\%$ and $C_v < 20\%$). Indeed, inhibition of MMP-1 is not influenced by atomic polarizabilities.

The nonlinear models are the result of the exploration of more complex relations. They bring more reliable conclusions in accordance with the cross-validation experiments:

- Atomic van der Waals volumes have high contributions in the inhibition of all MMPs ($C_v > 30\%$).
- Atomic van der Waals volumes are the key features in the inhibition of MMP-2, MMP-3, and MMP-9 ($C_v = 43\%$, $C_v = 41\%$, and $C_v = 65\%$).
- Atomic masses are the key features in the inhibition of MMP-1 ($C_m = 40\%$).
- Atomic polarizabilities are the key features in the inhibition of MMP-13 ($C_p = 42\%$).
- Inhibition of MMP-3 is not influenced by atomic polarizabilities.
- Atomic Sanderson electronegativities only influence inhibition of MMP-1 and MMP-3 ($C_e = 13\%$ and $C_e = 26\%$).

From the analysis of linear and nonlinear models, we can extract the main features relevant for the design of selective inhibitors. The primacy of atomic van der Waals volumes and atomic masses in our models points out the pertinence of the shape and atomic constitution of HPSAAs to occupy the MMP pockets. Otherwise, the nonlinear analysis suggests that MMP-13 inhibition can be modulated by polarizable residues. Finally, the linear models overestimated the effects of the atomic Sanderson electronegativities. According to nonlinear evidences, this effect must be considered in MMP-1 and MMP-3, which can be caused by the additional electrostatic interactions that provoke the changes in the S_1 and S_2' subsites for these MMPs (Fig. 2).

4. Conclusions

MMPs have been recognized as promising targets for cancer therapy on the basis of their massive up-regulation in malignant tissues and their unique ability to degrade all components of the extracellular matrix. Although many inhibitors of them have been reported and the high resolution X-ray crystal structures of MMP-inhibitor complexes have emerged to the scientific world, the selective inhibition of the MMP family members keeps on being a difficult task.

In this work, linear and nonlinear QSAR models were developed for relating inhibitor structural features with their biological activity against several MMP family members. The structural information was numerically encoded as 2D autocorrelation descriptors. Objective feature selection routine, which combined the genetic algorithm with MLR or BRANN fitness evaluator, was used to develop linear and nonlinear models. Nonlinear models overcome the linear results according to cross-validatory experiments.

Based on these results, we have successfully carried out a QSAR study in which the interactions of inhibitor substituents with the S_1', S_1, and S_2' pockets of MMP active sites were characterized in order to gain the knowledge about the essential features for increasing the selectivity. We conclude that HPSAA anchoring in MMP active sites is guided by spatial and hydrophobic-related effects. According to our results, the electronic effects have a poor contribution, but increase their importance in MMP-1 and MMP-3 inhibition.

Our results corroborate that the employment of 2D autocorrelation descriptors is extremely useful in QSAR studies giving simple correlations between the molecular structures and biological activities.

5. Experimental

5.1. Data sets: source and prior preparation

Inhibitions of MMP-1, MMP-2, MMP-3, MMP-9, and MMP-13 (IC_{50}) for 32 HPSAAs were taken from the literature.[7–9] For modeling, IC_{50} activities were converted in logarithmic activities $\log(10^6/IC_{50})$, where 10^6 guarantees that logarithmic activities range between 1 and 9. The chemical structures are shown in Table 1 and experimental activities ($\log(10^6/IC_{50})$) are shown in Table 2. The activity parameters IC_{50} (nM) are measures of inhibitory activity and refer to the nanomolar concentration of the MMP inhibitors leading to 50% inhibition of the MMP. Prior to molecular descriptor calculations, 3D structures of the studied compounds were geometrically optimized using the semiempirical quantum-chemical method PM3[40] implemented in the MOPAC 6.0[41] computer software.

5.2. Variable selections

A data matrix was 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 nonlinear models.

Three spatial autocorrelation vectors were employed for modeling the inhibitory activities: Broto-Moreau's autocorrelation coefficients,[42] Moran's indices,[43] and Geary's coefficients.[44] Spatial autocorrelation measures the level of interdependence between properties, and the nature and strength of that interdependence. 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 neighboring atoms. If dependence exists, the property is said to exhibit spatial autocorrelation. The autocorrelation vectors represent the degree of similarity between molecules. The Dragon[45] software was used for calculating unweighted and weighted Broto-Moreau, Moran, and Geary 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 colinearity 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 ($R^2_{max} = 0.95$). Afterwards, the number of remained descriptors was 50.

Since many molecular descriptors were available for QSAR analysis and only a reduced 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. Following the Occam's Razor,[20] we selected just the variables that contain the information that is necessary for the modeling but nothing more. In this sense, linear and nonlinear GA searches have been carried out in order to build the linear and nonlinear models.

Linear GA search was carried out exploring MLR models. In turn, neural network feature selection procedures that extract nonlinear information from the data set were employed for data dimensionality reduction before network training.[46] Linear and nonlinear models were generated between the activities log ($10^6/IC_{50}$) and the respective selected molecular descriptors. The quality of each model was proven by the square multiple correlation coefficient (R^2), the standard deviation (S), and their predictive abilities.

5.3. Bayesian-regularized Genetic Neural Networks (BRGNNs)

Nonlinear GA implemented in this paper is a version of the So and Karplus report[29] incorporating Bayesian regularization which was previously reported by our group[23] and was programmed within the Matlab environment using genetic algorithm and neural network toolboxes.[47] An individual in the population is represented by a string of integers which reflects the numbering of the columns in the data matrix.

Each individual encodes the same number of descriptors; the descriptors were randomly chosen from a common data matrix, and in a way such that (1) no two individuals can have exactly the same set of descriptors and (2) all descriptors in a given individual must be different. The fitness of each individual in this generation is determined by the fitness score computed by a BRANN and scaled using a scaling function. A top scaling fitness function scaled a top fraction of the individuals in a population equally; these individuals have the same probability to be reproduced, while the rest are assigned the value 0.

We employed two alternatives for selecting the fitness scores (Fig. 4). In alternative A, we tried the MSE of data fitting for BRANN models as the individual fitness function. MSE reflects the quality of the fit of the training data. Models with R-values higher than 0.80 were selected and they were tested in cross-validation experiments. This method can provide good solutions because Bayesian regularization reduces the likelihood of overfitted solutions.[23–26] In alternative B, the models were optimized for their predictive capacity by LOO cross-validation.

Initially, a set of 50 chromosomes were randomly generated. The population fitness was then calculated and the members were rank ordered according to fitness. The 2 best scoring models were automatically retained as members for the next round of evolution. More progeny models were then created for the next generation by preferentially mating parent models with higher scores. The crossover probability was set equal to 0.6 and the mutation rate was 0.3. The evolution proceeds until the best scoring model remains constant for at least 10 generations.

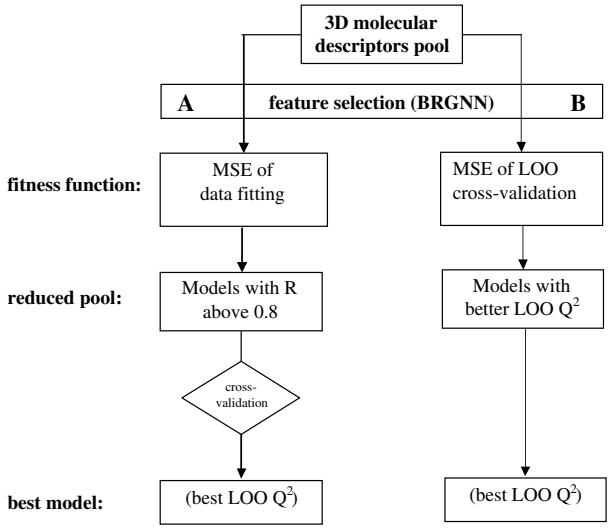


Figure 4. Schematic diagram describing both alternatives for nonlinear GA search.

We carried out several GA searches including 4–7 descriptors in the input layer. BRANNs were trained with a simple architecture (two or three neurons in a sole hidden layer). In these nets, the transfer functions of input and output layers were linear, and the hidden layer had neurons with a hyperbolic tangent transfer function. Inputs and targets took the values from independent variables selected by the GA and $\log(10^6/IC_{50})$ values, respectively; both were normalized prior to network training. BRANN training was carried out according to the Levenberg–Marquardt optimization.[48] The initial value for μ was 0.005 with decrease and increase factors of 0.1 and 10, respectively. The training was stopped when μ became larger than 10^{10}.

5.4. Model validation

The reliability of the GA selected models (with high R values) was indicated by cross-validation experiments quantified with predictive Q^2 for these models including from 2 to 5 neurons. For 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.

In addition to the traditional LOO cross-validation, leave-three-out (L3O) cross-validations were also performed, where in each experiment the objects were left out randomly. In this case, the results were reported as the averaged Q^2 of 10 replies.

References and notes

1. Birkedal-Hansen, H. *Curr. Opin. Cell Biol.* **1995**, *7*, 728.
2. Baker, A. H.; Edwards, D. R.; Murphy, G. *J. Cell Sci.* **2002**, *115*, 3719.
3. Lafleur, M.; Underwood, J. L.; Rappolee, D. A.; Werb, Z. *J. Exp. Med.* **1996**, *184*, 2311.
4. Leung, D.; Abbenante, G.; Fairlie, D. P. *J. Med. Chem.* **2000**, *43*, 305.
5. Beckett, R. P.; Davidson, A. H.; Drummond, A. H.; Whittaker, M. *Drug Discovery Today* **1996**, *1*, 16.
6. MacPherson, L. J.; Bayburt, E. K.; Capparelli, M. P.; Carroll, B. J.; Goldstein, R.; Justice, M. R.; Zhu, L.; Hu, S.-i.; Melton, R. A.; Fryer, L.; Goldberg, R. L.; Doughty, J. R.; Spirito, S.; Blancuzzi, V.; Wilson, D.; O'Byrne, E. M.; Ganu, V.; Parker, D. T. *J. Med. Chem.* **1997**, *40*, 2525.
7. Hanessian, S.; Bouzbouz, S.; Boudon, A.; Tucker, G. C.; Peyroulan, D. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 1691.
8. Hanessian, S.; Moitessier, N.; Gauchet, C.; Viau, M. *J. Med. Chem.* **2001**, *44*, 3066.
9. Hanessian, S.; MacKay, D. B.; Moitessier, N. *J. Med. Chem.* **2001**, *44*, 3074.
10. Hanessian, S.; Moitessier, N.; Therrien, E. *J. Comput. Aided Mol. Des.* **2001**, *15*, 873.
11. Kumar, D.; Gupta, S. P. *Bioorg. Med. Chem.* **2003**, *11*, 421.
12. Gupta, S. P.; Kumar, D.; Kumaran, S. *Bioorg. Med. Chem.* **2003**, *11*, 1975.
13. Gupta, S. P.; Kumaran, S. *Bioorg. Med. Chem.* **2003**, *11*, 3065.
14. Gupta, S. P.; Maheswaran, V.; Pande, V.; Kumar, D. *J. Enzyme Inhib. Med. Chem.* **2003**, *18*, 7.
15. Gupta, S. P.; Kumaran, S. *Bioorg. Med. Chem.* **2005**, *13*, 5454.
16. Holland, H. *Adaption in Natural and Artificial Systems*; The University of Michigan: Ann Arbor, MI, 1975.
17. Cartwright, H. M. *Applications of artificial intelligence in chemistry*; Oxford University: Oxford, 1993.
18. Zupan, J.; Gasteiger, J. *Anal. Chim. Acta* **1991**, *248*, 1.
19. Mackay, D. J. C. *Neural Comput.* **1992**, *4*, 415.
20. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1.
21. Burden, F. R.; Winkler, D. A. *J. Med. Chem.* **1999**, *42*, 3183.
22. Winkler, D. A.; Burden, F. R. *Biosilico* **2004**, *2*, 104.
23. Caballero, J.; Fernández, M. *J. Mol. Model.* **2006**, *12*, 168.
24. González, M. P.; Caballero, J.; Tundidor-Camba, A.; Helguera, A. M.; Fernández, M. *Bioorg. Med. Chem.* **2006**, *14*, 200.
25. Fernández, M.; Caballero, J. *Bioorg. Med. Chem.* **2006**, *14*, 280.
26. Fernández, M.; Tundidor-Camba, A.; Caballero, J. *J. Chem. Inf. Model.* **2005**, *45*, 1884.
27. Randić, M. *New J. Chem.* **1991**, *15*, 517.
28. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. *J. Am. Chem. Soc.* **1997**, *119*, 10509.
29. So, S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.
30. Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1328.
31. Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Nesari, N. *J. Chem. Inf. Model.* **2005**, *45*, 190.
32. Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Taghavi, F. *J. Comput. Chem.* **2004**, *25*, 1495.
33. Mattioni, B. E.; Jurs, P. C. *J. Mol. Graphics Modell.* **2003**, *21*, 391.
34. Welch, A. R.; Holman, C. M.; Huber, M.; Brenner, M. C.; Browner, M. F.; Van Wart, H. E. *Biochemistry* **1996**, *35*, 10103.
35. Chen, L.; Rydel, T. J.; Gu, F.; Dunaway, M.; Pikul, S.; McDow Dunham, K.; Barnett, B. L. *J. Mol. Biol.* **1999**, *293*, 545.
36. Lovejoy, B.; Welch, A. R.; Carr, S.; Luong, C.; Broka, C.; Hendricks, R. T.; Campbell, J. A.; Walker, K. A. M.; Martin, R.; Van Wart, H.; Browner, M. F. *Nat. Struct. Biol.* **1999**, *3*, 217.
37. Rowsell, S.; Hawtin, P.; Minshull, C. A.; Jepson, H.; Brockbank, S. M. V.; Barratt, D. G.; Slater, A. M.; McPheat, W. L.; Waterson, D.; Henney, A. M.; Pauptit, R. A. *J. Mol. Biol.* **2002**, *319*, 173.
38. Fernández, M.; Tundidor-Camba, A.; Caballero, J. *Mol. Simulat.* **2005**, *31*, 575.
39. Cherqaoui, D.; Esseffar, M.; Villemin, D.; Cence, J. M.; Chastrette, M.; Zakarya, D. *New J. Chem.* **1998**, *22*, 839.
40. Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 210.
41. MOPAC version 6.0. Frank J. Seiler Research Laboratory, U. S. Air Force academy Colorado Springs, CO, 1993.
42. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359.
43. Moran, P. A. P. *Biometrika* **1950**, *37*, 17.
44. Geary, R. F. *Incorporated Statistician* **1954**, *5*, 115.
45. DRAGON Software version 3.0, Milano Chemometrics, 2003.
46. Yasri, A.; Hartsough, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218.
47. MATLAB version 7.0. The MathWorks, Inc. 2004. WEB: www.mathworks.com.
48. Foresee, F. D.; Hagan M. T. Gauss–Newton approximation to Bayesian learning. In *Proceedings of the 1997 International Joint Conference on Neural Networks*, 1997; pp 1930–1935.