Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 15 (2007) 6298–6310
Bioorganic & Medicinal Chemistry

QSAR modeling of matrix metalloproteinase inhibition by *N*-hydroxy-α-phenylsulfonylacetamide derivatives

Michael Fernández[a] and Julio Caballero[b,*]

[a]*Molecular Modeling Group, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba*
[b]*Centro de Bioinformática y Simulación Molecular, Universidad de Talca, 2 Norte 685, Casilla 721, Talca, Chile*

Received 30 March 2007; revised 24 May 2007; accepted 6 June 2007
Available online 10 June 2007

Abstract—The main molecular features which determine the selectivity of a set of 80 *N*-hydroxy-α-phenylsulfonylacetamide derivatives (HPSAs) in the inhibition of three matrix metalloproteinases (MMP-1, MMP-9, and MMP-13) have been identified by using linear and nonlinear predictive models. The molecular information has been encoded in 2D autocorrelation descriptors, obtained from different weighting schemes. The linear models were built by multiple linear regression (MLR) combined with genetic algorithm (GA), and a robust QSAR mapping paradigm. The Bayesian-regularized genetic neural network (BRGNN) was employed for nonlinear modeling. In such approaches each model could have its own set of input variables. All models were predictive according to internal and external validation experiments; but the best results correspond to nonlinear ones. The 2D autocorrelation space brings different descriptors for each MMP inhibition, and suggests the atomic properties relevant for the inhibitors to interact with each MMP active site. On the basis of the current results, the reported models have the potential to discover new potent and selective inhibitors and bring useful molecular information about the ligand specificity for MMP S'_1 and S'_2 subsites.

1. Introduction

Matrix metalloproteinases (MMPs) constitute a family of zinc endopeptidases which are collectively able to degrade all components of the extracellular matrix such as collagens, proteoglycans, fibronectin, laminin, elastin, and many nonmatrix proteins.[1] They are involved in connective tissue remodeling and are implicated in some processes such as ovulation, embryonic growth, angiogenesis, differentiation, and healing.[2] Any disturbance of the generally well-balanced equilibrium between the MMPs and their physiological inhibitors: the tissue inhibitors of MMPs can give rise to pathological situations such as rheumatoid and osteoarthritis, atherosclerosis, tumor development, tumor metastasis and pulmonary emphysema.[3] In this context, MMP inhibitors have caught the interest as an important class of drugs for the development of innovative chemotherapeutics in several fields where effective treatments are lacking.[4]

Despite MMPs share certain biochemical properties, each has distinct substrate specificity and, up to date, several mammalian enzymes have been identified ranging from well-characterized enzymes such as collagenase, stremolysin, gelatinase, and membrane type MMPs. At the same time, it has been identified that different MMPs contribute to different stages of disease processes; therefore, the design of selective MMP inhibitors should limit potential side effects. A broad-spectrum of peptidic or nonpeptidic structures bearing a zinc-binding ligand (e.g., carboxylic or hydroxamic acids) have been recognized as MMP inhibitors.[5–9] The selectivity has been tried 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. The use of computer-aided design methods can more closely extract the structural features and binding characteristics of the MMP active sites and thereby minimize MMP inhibitor specificity-related side effects. Molecular Dynamics and docking-type techniques have helped to explore the structural differences of MMPs and their interactions with MMP inhibitors.[10,11] In addition, quantitative

Keywords: MMP inhibitors; Bayesian-regularized genetic neural networks; QSAR analysis; 2D autocorrelation space.
* Corresponding author. Tel.: +56 71 201 662; fax: +56 71 201 561; e-mail addresses: jcaballero@utalca.cl; jmcr77@yahoo.com

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.06.014

structure–activity relationship (QSAR) studies have been successfully applied for modeling activities of MMP inhibitors.[12–18]

In a recent work, we carried out QSAR modeling for relating inhibitor structural features with biological activities for a set of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives against several MMP family members.[19] 2D autocorrelation pool was used for encoding structural information and the relevant information that relates the topological features of these compounds with their inhibitory activities against the studied MMP family members was extracted by linear and nonlinear genetic algorithm (GA) feature selection. In the current work, we applied the 2D autocorrelation methodology to a set of 80 *N*-hydroxy-α-phenylsulfonylacetamide derivatives (HPSAs) (the chemical structures are shown in Table 1) as inhibitors of MMP-1, MMP-9, and MMP-13. We established the structure–activity relationships with both multiple linear regression (MLR) and Bayesian-regularized genetic neural network (BRGNN) approaches.

2. Results and discussion

The studied data set includes selective compounds which are potent inhibitors of MMP-9 and MMP-13, and moderate inhibitors of MMP-1. Inhibitors of MMP-9 are potentially valuable for arresting tumor metastasis, while inhibitors of MMP-13 can offer protection from the cartilage degradation associated with osteoarthritis. Meanwhile, the inhibition of MMP-1 is a possible source of the musculoskeletal side effects that have been seen in clinical trials of broad-spectrum MMP inhibitors.[20] Correlation matrix (Table 2) shows that inhibitory activities of HPSAs employed in this study against MMP-9 and MMP-13 are related to each other. However, the activities against MMP-1 are completely unrelated with the rest.

Six models are reported in this work. In total, 27 descriptors from the whole 2D autocorrelation pool were employed. The colinearity of the variables should be as low as possible for guarantying the absence of redundant information.[21] The correlation of each one of these descriptors in these equations with each other was calculated. There are only five correlated pairs ($R^2 > 0.7$) from 351 pairs. Furthermore, there are six pairs with correlations between 0.5 and 0.6, 98 pairs with correlations between 0.1 and 0.5, and 241 pairs with correlations between 0 and 0.1. The five colinear pairs are: MATS6v-GATS6v ($R^2 = 0.946$), GATS6v-GATS6p ($R^2 = 0.782$), GATS7v-GATS7p ($R^2 = 0.766$), MATS6v-GATS6p ($R^2 = 0.738$), and GATS4v-GATS4p ($R^2 = 0.706$).

2.1. Multiple linear regression approach

Linear correlations were developed by means of MLR models for inhibitory activities of HPSAs against three MMPs with acceptable statistical significances and predictive power (Eqs. 1–3).

MLR-MMP-1:

$$\log(10^6/\mathrm{IC}_{50}) = -17.557 \times \mathrm{MATS4m} - 5.396 \times \mathrm{MATS3v} + 17.908 \times \mathrm{MATS6v} - 4.396 \times \mathrm{MATS5e} - 4.375 \times \mathrm{MATS6e} + 10.359 \times \mathrm{GATS6v} - 5.118 \times \mathrm{GATS7v} + 15.274 \quad (1)$$

$$N_{\mathrm{training}} = 63 \quad R^2 = 0.736 \quad S = 0.312 \quad p < 10^{-5}$$
$$R^2_{\mathrm{CV}} = 0.559 \quad S_{\mathrm{CV}} = 0.403$$
$$N_{\mathrm{test}} = 10 \quad R^2_{\mathrm{EP}} = 0.664 \quad S_{\mathrm{EP}} = 0.282$$

MLR-MMP-9:

$$\log(10^6/\mathrm{IC}_{50}) = -0.005 \times \mathrm{ATS6m} + 0.018 \times \mathrm{ATS3e} + 8.881 \times \mathrm{MATS2e} - 7.718 \times \mathrm{MATS4e} - 4.655 \times \mathrm{GATS1v} + 14.788 \times \mathrm{GATS1e} + 2.379 \times \mathrm{GATS6p} - 4.571 \quad (2)$$

$$N_{\mathrm{training}} = 66 \quad R^2 = 0.731 \quad S = 0.416 \quad p < 10^{-5}$$
$$R^2_{\mathrm{CV}} = 0.605 \quad S_{\mathrm{CV}} = 0.504$$
$$N_{\mathrm{test}} = 12 \quad R^2_{\mathrm{EP}} = 0.713 \quad S_{\mathrm{EP}} = 0.415$$

MLR-MMP-13:

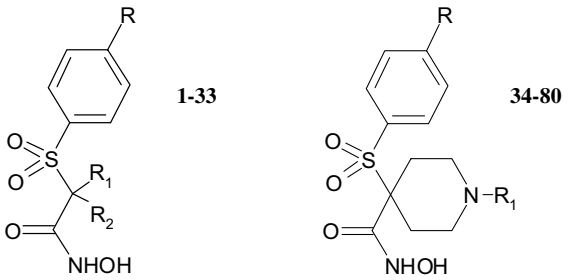
$$\log(10^6/\mathrm{IC}_{50}) = 0.017 \times \mathrm{ATS3m} + 11.363 \times \mathrm{MATS6v} - 1.118 \times \mathrm{MATS6e} - 1.826 \times \mathrm{GATS1v} + 11.911 \times \mathrm{GATS6v} + 6.097 \times \mathrm{GATS1e} + 1.297 \times \mathrm{GATS4e} - 11.139 \quad (3)$$

$$N_{\mathrm{training}} = 68 \quad R^2 = 0.692 \quad S = 0.376 \quad p < 10^{-5}$$
$$R^2_{\mathrm{CV}} = 0.598 \quad S_{\mathrm{CV}} = 0.430$$
$$N_{\mathrm{test}} = 12 \quad R^2_{\mathrm{EP}} = 0.727 \quad S_{\mathrm{EP}} = 0.429$$

In Eqs. 1–3, N_{training} and N_{test} are the number of compounds included in the training and test sets, respectively, R^2 is the square of correlation coefficients, S is the standard deviation of the regressions, p is the significance of the variables in the models, R^2_{CV} and S_{CV} are the correlation coefficients and standard deviations of the leave-one-out (LOO) cross-validation, respectively. R^2_{EP} and S_{EP} are the correlation coefficients and standard deviations of test set regressions, respectively.

The MLR training and test set predictions ($\log(10^6/\mathrm{IC}_{50})$) for the HPSAs against MMP-1. MMP-9, and MMP-13 appear in Table 3. In turn, plots of training and test set predictions versus experimental

Table 1. Structural features of *N*-hydroxy-α-phenylsulfonylacetamide derivatives (HPSAs)



Compound[a]	R	R_1	R_2
1	OMe	Bn	H
2	OMe	Bn	Me
3	OMe	2-Naphthylmethyl	H
4	OMe	2-Naphthylmethyl	Me
5	OMe	4-Biphenylmethyl	Me
6	OMe	Isoprenyl	Me
7	OMe	Isoprenyl	Isoprenyl
8	OMe	Allyl	Allyl
9	OMe	3-Phenylallyl	Me
10	OMe	*n*-Pr	*n*-Pr
11	OMe	IsoPr	H
12	OMe	*n*-Bu	H
13	OMe	Cyclohexylmethyl	Me
14	OMe	*n*-Dodecyl	H
15	OMe	Propargyl	Propargyl
16	OMe	3-Picolyl	Me
17	OMe	3-Picolyl	Isoprenyl
18	OMe	3-Picolyl	IsoBu
19	OMe	3-Picolyl	Isopentyl
20	OMe	3-Picolyl	*n*-Bu
21	OMe	3-Picolyl	*n*-Octyl
22	OMe	3-Picolyl	Propargyl
23	OMe	4-[2-(1-Piperidinyl)ethoxy]benzyl	Me
24	OMe	4-[2-(1-Azepanyl)ethoxy]benzyl	Me
25	OMe	4-[2-(Diisopropylamino)ethoxy]benzyl	Me
26	OMe	4-[2-(Diethylamino)ethoxy]benzyl	Me
27	OMe	4-{3-[4-(3Cl-phenyl)-1-piperazinyl]propoxy}benzyl	Me
28	OMe	4-[2-(4-Morpholinyl)ethoxy]benzyl	Me
29	OEt	4-[2-(Diethylamino)ethoxy]benzyl	Me
30	O-*n*-Bu	4-[2-(1-Piperidinyl)ethoxy]benzyl	Me
31	2-Furyl	4-[2-(Diethylamino)ethoxy]benzyl	Me
32	Br	4-[2-(Diethylamino)ethoxy]benzyl	Me
33	Me	Isoprenyl	Isoprenyl
34	OMe	Bn	—
35	OMe	3-Methoxy benzyl	—
36	OMe	3,4-Dichloro benzyl	—
37	OMe	4-Me benzyl	—
38	OMe	2-Naphthylmethyl	—
39	OMe	4-Biphenylmethyl	—
40	OMe	Isoprenyl	—
41	OMe	4-Br benzyl	—
42	OMe	3-Ph propyl	—
43	OMe	*t*-Bu	—
44	OMe	*n*-Bu	—
45	OMe	Cyclo octyl	—
46	OMe	Et	—
47	OMe	IsoPr	—
48	OMe	Me	—
49	O-*n*-Bu	Bn	—
50	OMe	4-F benzyl	—
51	O-*n*-Bu	4-F benzyl	—
52	OMe	4-Methoxy benzyl	—
53	OMe	4-Methoxy phenyl ethyl	—

Table 1 (*continued*)

Compound[a]	R	R_1	R_2
54	OMe	2-Ph ethyl	—
55	O-*n*-Bu	4-Methoxy benzyl	—
56	OMe	3-Phenoxy propyl	—
57	O-*n*-Bu	3-Phenoxy propyl	—
58	OMe	2-Phenoxy ethyl	—
59	O-*n*-Bu	2-Phenoxy ethyl	—
60	OMe	4-[2-(1-Piperidinyl)ethoxy]benzyl	—
61	O-*n*-Bu	4-[2-(1-Piperidinyl)ethoxy]benzyl	—
62	O-*n*-Bu	3-[2-(4-Morpholinyl)ethoxy]benzyl	—
63	O-*n*-Bu	Me	—
64	O-*n*-Bu	Et	—
65	O-*n*-Bu	*n*-Bu	—
66	O-Benzyl	Bn	—
67	O-4Cl-Phenyl	Me	—
68	O-4Cl-Phenyl	Et	—
69	O-4Cl-Phenyl	*n*-Bu	—
70	O-4Cl-Phenyl	Bn	—
71	O-4Cl-Phenyl	H	—
72	O-Isopentyl	Bn	—
73	2-Ethylbutoxy	Bn	—
74	O-*n*-Bu	3-Methoxy benzyl	—
75	OMe	4-(2-Thienyl)benzyl	—
76	OMe	4-(2-Pyridinyl)benzyl	—
77	O-*n*-Bu	3,4-Dichloro benzyl	—
78	O-4Cl-Benzyl	4-Me benzyl	—
79	2-Furanyl	Bn	—
80	O-4Cl-Phenyl	4-Methoxy benzyl	—

[a] Compounds **1**–**33** are from Ref. 8 and **34**–**80** are from Ref. 9.

$\log(10^6/IC_{50})$ values for the MLR models are shown in Figure 1. In general, MLR models were able to explain data variance and were quite stable to the inclusion–exclusion of compounds as measured by LOO correlation coefficients ($Q^2 > 0.5$). In addition, the MLR models were able to describe the test set variances with $R^2_{EP} > 0.6$ and small S_{EP} values.

The relationship between inhibitory activities has been reflected by some similarities between the linear models:

- There is supremacy of atomic van der Waals volume and Sanderson electronegativity weighted terms in MMP-1 and MMP-13 models.
- The colinear pair MATS6v-GATS6v influences the inhibition of MMP-1 and MMP-13. In turn, GATS6p, which is colinear with the both above-mentioned descriptors, influences the inhibition of MMP-9. The signs of the coefficients of these descriptors are positive in all MLR models.
- GATS1v has a negative influence in the inhibition of MMP-9 and MMP-13.
- GATS1e has a positive influence in the inhibition of MMP-9 and MMP-13.

Table 2. Correlation matrix for the *N*-hydroxy-α-phenylsulfonylacetamide activities against MMPs

	MMP-1	MMP-9	MMP-13
MMP-1	1		
MMP-9	0.020	1	
MMP-13	0.000	0.682	1

The common features expressed in the 2D autocorrelation space represent the similarities between target enzymes (MMPs) relevant to interactions with HPSA inhibitors. By contrast, the peculiarities must contain information about the relevant molecular characteristics for each selective MMP inhibition. In this sense, the more showy dissimilarities are summed up:

- Model of activity against MMP-1 has the major contribution of atomic van der Waals volume weighted terms.
- Model of activity against MMP-9 has the major contribution of atomic Sanderson electronegativity weighted terms.
- Model of activity against MMP-1 does not contain short lag descriptors ($l = 1$ or 2). In contrast, MLR-MMP-9 and MLR-MMP-13 models contain three and two short lag descriptors, respectively.

2.2. 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[22] which surpassed the limits of the linear solutions when modeling of inhibitory activities was performed.[19,22,23] 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

Table 3. Experimental and predicted $\log(10^6/IC_{50})$ by linear and nonlinear models

Compound	MMP-1			MMP-9			MMP-13		
	Exp	Predictions		Exp	Predictions		Exp	Predictions	
		Lin	Nonlin		Lin	Nonlin		Lin	Nonlin
1	3.50	3.12	3.48	3.97	3.78[a]	3.90[a]	4.00	4.26	4.14
2	4.00	3.82	4.07	4.96	4.70	4.69	4.96	4.85	5.04
3	3.23	3.17	3.31	3.71	3.95	4.37	4.85	4.54	4.59
4	3.86	4.01	3.83	5.10	4.82[a]	4.64[a]	5.05	5.11	4.87
5	3.80	3.76	3.83	4.64	5.03	4.75	5.10	5.21	5.07
6	3.62	3.76	3.62	4.96	5.13	4.34	4.85	5.24	5.15
7	4.60	3.66	4.29	6.30	5.22	5.68	6.40	5.74	5.89
8	3.68	3.30[a]	3.40[a]	4.46	3.87	4.07	4.41	4.41	4.33
9	3.52	3.87[a]	3.70[a]	4.80	4.54	4.41	4.92	5.07	5.02
10	—	—	—	3.35	3.43	3.52	3.85	3.84	3.99
11	3.19	3.28	2.99	—	—	—	3.73	4.07[a]	3.30[a]
12	—	—	—	3.89	3.40	3.78	4.19	3.93[a]	3.51[a]
13	3.49	3.16[a]	3.08[a]	5.05	4.07	4.57	4.62	4.18	4.60
14	—	—	—	2.53	3.06	3.31	3.50	3.84	3.49
15	3.52	3.61	3.33	3.85	4.60	4.35	4.92	5.21	4.80
16	3.59	3.49	3.21	4.42	4.61	4.82	4.66	4.61	4.67
17	3.81	3.40	3.90	5.05	4.90	4.76	5.52	5.02	4.97
18	3.00	3.19	3.39	4.20	3.72	3.75	4.89	4.19[a]	4.04[a]
19	3.24	3.14	3.16	3.92	4.69	4.36	4.05	4.89	4.73
20	2.94	2.91	3.10	—	—	—	3.90	4.72[a]	4.65[a]
21	3.28	2.71	3.04	3.76	4.22	4.19	4.37	4.63	4.33
22	3.17	3.48	3.52	4.08	4.55[a]	4.65[a]	4.49	4.93	4.68
23	3.62	3.52	3.47	5.05	4.95[a]	5.22[a]	6.00	5.11	5.21
24	3.27	3.36	3.47	4.72	5.11	4.89	4.92	5.07	5.33
25	3.37	3.96	3.51	4.82	4.39	4.70	4.72	4.25	4.45
26	3.50	3.39	3.76	4.89	4.62	4.80	4.82	4.93	4.87
27	3.23	3.46	3.04	5.15	5.61[a]	5.62[a]	5.40	5.72	5.54
28	3.38	3.34	3.41	4.68	4.72	4.96	4.51	5.26	4.78
29	3.20	3.21	3.42	4.96	4.89	4.91	4.80	5.01	5.12
30	3.12	3.07[a]	3.07[a]	5.52	5.50	5.46	5.70	5.46[a]	5.47[a]
31	4.85	4.42	4.53	5.40	5.38	5.36	5.70	5.42	5.43
32	4.29	3.98	4.34	4.24	4.63	4.16	4.96	4.89	5.04
33	3.58	4.40	3.70	4.29	4.97	4.64	5.22	5.61	5.62
34	3.31	3.15	3.13	5.00	4.59	4.76	5.70	5.33	5.51
35	3.28	3.42[a]	3.08[a]	5.05	4.94	4.94	5.70	5.68	5.38
36	3.35	3.09[a]	3.14[a]	5.22	5.02	5.18	5.70	5.73	5.53
37	3.31	3.30	3.33	4.77	4.67	4.83	5.70	5.34	5.66
38	3.43	3.09	3.32	5.30	4.94	5.03	5.70	5.43	5.46
39	2.88	2.93	3.36	4.89	5.10	5.12	5.52	5.59[a]	5.38[a]
40	2.86	2.87	3.02	4.41	4.21	4.35	5.15	5.36	5.12
41	3.22	2.99	3.06	5.00	5.09	5.07	5.70	5.42	5.74
42	2.71	3.02	2.68	4.89	4.38	4.44	4.96	4.91	5.29
43	—	—	—	3.19	3.27	3.53	3.83	4.13	4.13
44	—	—	—	3.34	4.04	3.96	4.42	5.00	4.70
45	2.58	2.55	2.28	3.86	4.16	3.54	4.54	4.62	4.99
46	2.43	2.51	2.07	3.44	3.85	3.27	4.48	4.70	4.44
47	2.35	2.48	2.25	3.43	3.58[a]	3.28[a]	4.47	4.43	4.32
48	2.29	2.83	2.58	3.32	4.06	3.55	4.36	4.91[a]	4.43[a]
49	2.62	2.54	2.57	5.40	5.03	5.19	6.00	5.59[a]	5.71[a]
50	3.18	3.18	3.00	4.80	4.40	4.71	5.70	5.16	5.48
51	2.33	2.57	2.63	4.72	4.81	5.50	5.30	5.37	5.73
52	3.19	3.07	3.04	4.92	4.89	4.49	5.70	5.59	5.46
53	3.18	2.69[a]	3.15[a]	4.47	4.77	4.02	5.70	5.42	5.45
54	2.88	2.65	3.00	4.35	4.41	4.32	5.05	5.12[a]	5.36[a]
55	2.58	2.54	2.67	5.52	5.39	5.36	6.00	5.85[a]	5.89[a]
56	2.92	3.37	2.86	4.36	4.96[a]	4.74[a]	5.40	5.36	5.40
57	2.42	2.84	2.41	5.30	5.44	5.33	6.00	5.63	6.00
58	3.20	3.15	3.04	4.59	4.84	4.85	5.52	5.33	5.42
59	2.54	2.62	2.65	5.30	5.32	5.40	5.70	5.61	5.86
60	3.41	3.11	2.91	5.52	5.29	5.31	5.52	5.62	5.74
61	2.71	2.81	2.70	5.70	5.89	5.59	6.00	5.96	5.69
62	2.66	2.77	2.66	5.70	5.70	5.89	5.70	6.13	5.90

Table 3 (*continued*)

Compound	MMP-1			MMP-9			MMP-13		
	Exp	Predictions		Exp	Predictions		Exp	Predictions	
		Lin	Nonlin		Lin	Nonlin		Lin	Nonlin
63	2.47	2.31[a]	2.45[a]	5.00	4.46[a]	4.65[a]	5.70	5.42	5.53
64	2.15	2.12	2.28	4.37	4.31	4.36	5.70	5.19	5.46
65	—	—	—	4.55	4.53	4.83	5.70	5.47	5.64
66	—	—	—	5.52	5.26	5.38	4.80	5.73	5.20
67	2.85	2.93	2.94	5.70	5.54	5.74	5.70	5.57	6.12
68	2.76	2.85	2.92	6.00	5.36[a]	5.53[a]	6.00	5.49	5.94
69	2.97	2.70[a]	2.89[a]	6.00	5.44	5.73	6.00	5.55	5.96
70	3.10	3.04[a]	2.95[a]	6.00	5.80	5.63	6.00	5.87	5.62
71	2.90	3.08	2.97	5.70	6.05	5.68	5.70	5.79	5.88
72	2.59	2.72	2.82	5.40	4.88	5.20	5.70	5.55	5.60
73	2.50	2.57	2.73	4.85	5.17	5.31	5.30	5.70	5.32
74	2.82	2.85	2.76	5.52	5.41[a]	5.51[a]	6.00	5.93[a]	5.95[a]
75	3.27	3.10	3.19	4.96	4.92	5.11	5.52	5.61	5.38
76	3.37	3.09	3.19	5.22	5.33	5.34	5.70	5.87	5.65
77	2.44	2.27	2.76	4.70	5.35[a]	4.53[a]	5.30	5.75[a]	5.25[a]
78	2.32	3.10	2.94	5.22	5.86	5.71	5.22	5.81	5.78
79	4.40	4.11	3.98	5.52	5.37	5.29	6.00	5.79	5.59
80	3.31	3.05	2.92	6.00	6.07[a]	5.97[a]	6.00	6.11	5.88

[a] Test set predictions, the rest are training predictions.

application, ANN architectures were varied testing different quantities of neurons in hidden layers.

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.[24] For this reason, a variety of fitness functions which are proportional to the residual error of the validation set,[24–27] or even the cross-validation set from the neural network simulations,[24,28] 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 the validation set does not contain some outliers, by fortuitous manner, 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 4 and references therein).

The descriptors and statistics of BRGNN models are depicted in Table 4 and plots of predicted versus experimental $\log(10^6/IC_{50})$ values are shown in Figure 1. The best models include seven variables and contain two nodes in the hidden layer. The number of optimum parameters yielded by the Bayesian regularization was 16 or 17 in all cases. BRGNN statistics reveal that neural network approaches surpass the results achieved by MLR in regard to fitness and predictive capacity ($R^2_{CV} > 0.6$, $R^2_{EP} > 0.7$ in all cases). Inhibitory activities ($\log(10^6/IC_{50})$) of the HPSAs against the MMPs predicted by nonlinear models appear in Table 3.

On the basis of the great reliability that they show statistically, expressed by means of internal and external validation experiments, we consider the nonlinear solutions as best-suited for analyzing structural properties relevant to inhibitor-MMP interactions. The inspection of BRGNN models reveals quite a few coincidences: BRGNN-MMP-9 and BRGNN-MMP-13 models share the ATS6m descriptor, while BRGNN-MMP-1 and BRGNN-MMP-9 models share the information provided by the colinear pair MATS7v-GATS7v, however BRGNN-MMP-1 and BRGNN-MMP-9 do not share any descriptor.

In order to gain a deeper insight into the relative effects of each 2D autocorrelation descriptor in our model, a recently reported weight-based input ranking scheme was carried out. Black-box nature of three layer ANNs has been 'deciphered' in a recent report of Guha et al.[29] Their method allows understanding how an input descriptor is correlated to the predicted output by the network and consists of two parts: first, the nonlinear transform for a given neuron is linearized. Afterward, the magnitude in which a given neuron affects the downstream output is determined. Next, a ranking scheme for neurons in the hidden layer is developed. The ranking scheme is carried out by determining the square contribution values (SCV) for each hidden neuron (see Ref. 29 for details). This method for ANN model interpretation is similar in manner to the partial least squares interpretation method for linear models described by Stanton.[30]

The results of ANN deciphering studies are displayed in Table 5. The reported effective weight matrixes show that the second hidden neurons have the major contributions to all the models. The SCV value is 5.5-fold higher for second hidden neuron with respect to the first hidden neuron in the BRGNN-MMP-1 model; while SCV value is quite bigger for second neurons in BRGNN-MMP-9 and BRGNN-MMP-13 models.

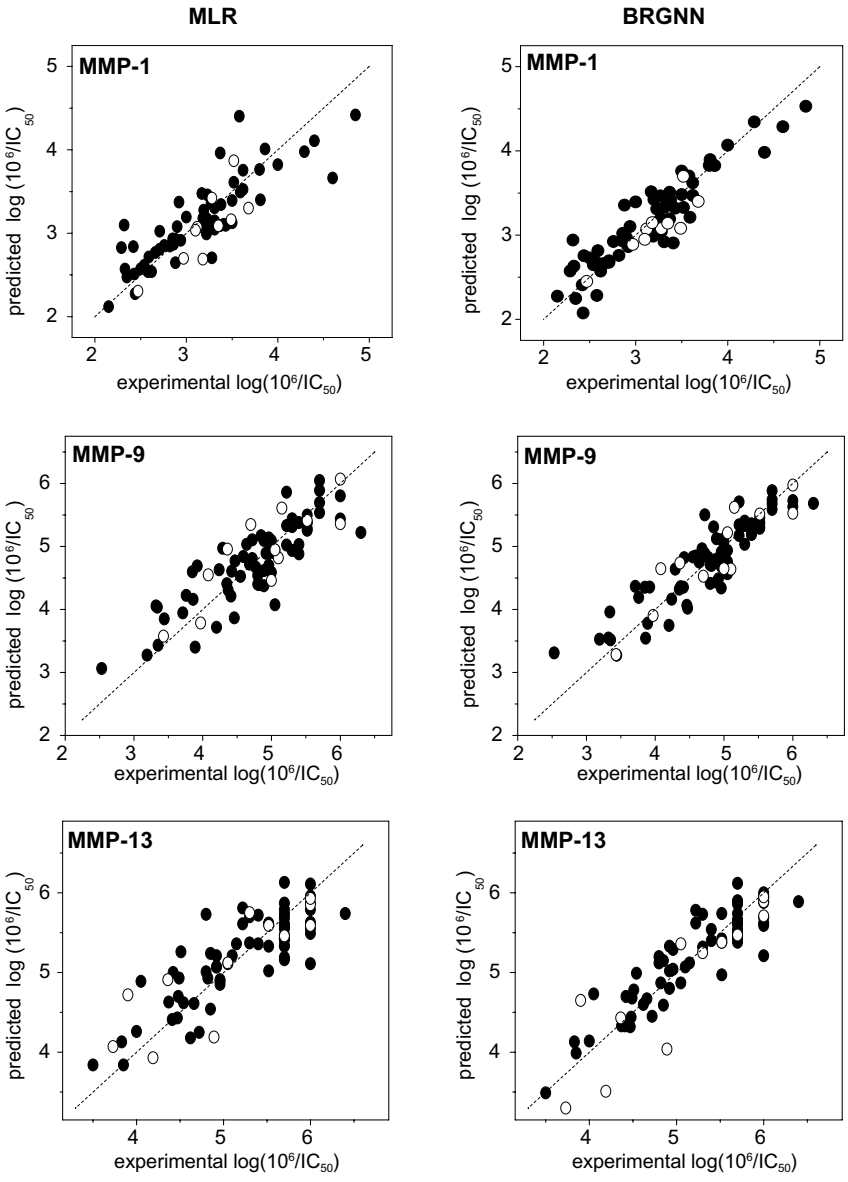


Figure 1. Plot of predicted versus experimental $\log(10^6/IC_{50})$ values for MMP inhibition by *N*-hydroxy-α-phenylsulfonylacetamide derivatives using linear (left) and nonlinear (right) models. (●) Training set predictions; (○) test set predictions.

Table 4. Descriptors and statistics for BRGNN models[a]

Model	Descriptors	Training set					LOO cross-validation		Test set		
		n	Num. par.	Opt. par.	R^2	*S*	R^2_{CV}	S_{CV}	*n*	R^2_{EP}	S_{EP}
BRGNN-MMP1	ATS3e, MATS3m, MATS3e, MATS5e, MATS6e, GATS1v, GATS7p	63	19	17	0.844	0.224	0.601	0.377	10	0.781	0.362
BRGNN-MMP9	ATS6m, MATS2m, MATS5v, MATS1e, GATS4v, GATS5e, GATS4p	66	19	16	0.813	0.327	0.692	0.421	12	0.814	0.332
BRGNN-MMP13	ATS3m, ATS6m, MATS1v, GATS7v, GATS3e, GATS4e, GATS6p	68	19	16	0.816	0.273	0.647	0.384	12	0.785	0.429

[a] 7-2-1Architecture 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

According to such characteristics, we can derive the approximate effect of the descriptors by the analysis of the sign of the weights in second neurons. In the model derived using MMP-1 inhibitory activities, MATS3e has the highest impact equal to 2.071 with a positive influence, however, there is a considerable influence of

Table 5. Effective weight matrix for the optimum BRGNN models

MMP-1			MMP-9			MMP-13		
Network inputs	Hidden neurons		Network inputs	Hidden neurons		Network inputs	Hidden neurons	
	2	1		2	1		2	1
ATS3e	−0.552	1.330	ATS6m	**1.372**	−1.098	ATS3m	**1.567**	−0.420
MATS3m	0.405	−1.656	MATS2m	0.931	−0.987	ATS6m	0.260	−0.788
MATS3e	**2.071**	−2.264	MATS5v	−0.470	1.077	MATS1v	0.865	−0.676
MATS5e	**−1.385**	1.540	MATS1e	0.162	−0.691	GATS7v	0.550	−0.477
MATS6e	−0.998	0.346	GATS4v	−0.532	−0.099	GATS3e	−0.095	0.402
GATS1v	0.985	−2.069	GATS5e	0.783	−1.090	GATS4e	0.591	−0.302
GATS7p	0.534	−0.534	GATS4p	0.223	0.208	GATS6p	**1.392**	−0.169
SCV	0.848	0.152	SCV	1.000	5×10^{-6}	SCV	0.999	0.001

Most relevant descriptors appear in bold letter.
[a]The columns are ordered by the SCVs for the hidden neurons, shown in the last row.

MATS5e and MATS6e descriptors with a negative influence (similar to MLR-MMP-1 model). Pursuant to these results, the atomic Sanderson electronegativities are the most important properties for MMP-1 inhibition; the difference of signs (between weights in second neuron and weights in first and second neurons) suggesting a complex nonlinear effect. In the model derived using MMP-9 inhibitory activities, the atomic mass weighted terms (ATS6m and MATS2m) showed the higher weights with positive influences. The same analysis in the model derived using MMP-13 inhibitory activities showed higher weights for ATS3m and GATS6p with positive influences.

2.3. Qualitative comparison of linear and nonlinear models

According to the analysis of the weighted atomic properties inside the 2D autocorrelation space, the differences between structure–activity fitting models can be interpreted as the differences between the inhibitor relevant molecular information for having a certain inhibitory activity against several MMPs. Our QSAR study is based on previous structure–activity relationship study in which authors modified functional groups at the P'_1 and P'_2 sites (R, R_1, and R_2 in Fig. 2) of the inhibitors as functional probes for S'_1 and S'_2 subsites of MMPs.[8,9] MMP-1 has a characteristic Arg214 in its S'_1 subsite (Fig. 2) (numberings correspond to human MMPs as in SWISS-PROT[31]). The long side chain of the Arg214 extends to the bottom of the S'_1 subsite and forms a rather shallow pocket. In MMP-9/MMP-13, this Arg is replaced by Leu397/Leu218 residue, which causes a deep pocket. Modification of the P'_2 substituent could provide such selectivity due to the hydrophilic nature of the S'_2 site of MMP-1 containing Asn180 and Ser239, as compared to the hydrophobic nature of MMP-9/MMP-13, which contain Leu187/Leu184 and Met422/Ile243, respectively.[32,33] Authors searched for selectivity in accordance with the differences between these pockets. We consider the differences between 2D autocorrelation spaces extracted by linear and nonlinear GA as the relevant features for the inhibitor-pocket interactions; that is, the different spaces are caused by the differences between MMP S'_1 and S'_2 pockets. The interpretation of these differences on the basis of the driving forces for inhibitor inclusion in enzyme active sites is feasible.

The success of HPSAs as MMP inhibitors lies on hydroxamic acid moiety chelating Zn^{2+} ion and sulfonamide group-related hydrogen bonds. The occupation of the pockets allows modulating the selectivity by steric, hydrophobic, and electronic differences among MMP active sites. The importance of electrostatic interactions and hydrophobicity as the key features which drive the inhibitor-MMP affinities was assessed by Gupta et al.[12–16] by means of QSAR studies where linear dependencies between molecular properties and inhibitory activities were determined using hydrophobicity related descriptors ($^1\chi^v$ or $\log P$) and electrotopological state (E-state) indices. These effects could be itemized by QSAR analysis of a set of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives as inhibitors of several MMPs employing 2D autocorrelation approach.[19]

The use of 2D autocorrelation space proved advantageous in previous works.[19,23,34,35] It can be readily derived directly from the molecular structures without

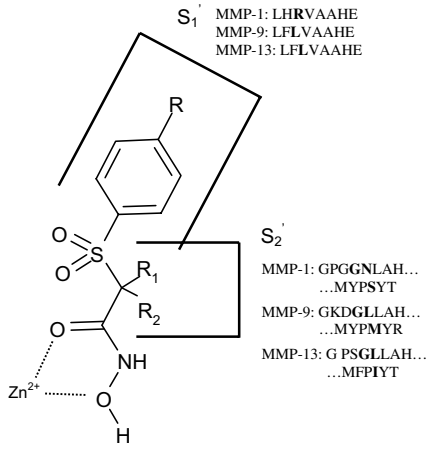


Figure 2. Position of *N*-hydroxy-α-phenylsulfonylacetamide derivatives inside MMP active site. Comparison of amino acid sequences of MMPs. Bold letters indicate the amino acid of S'_1 and S'_2 pockets that contribute to the ligand specificity.

any experimental effort and their computation 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 (Fig. 3). Like other descriptors giving an all-embracing representation of the molecule, 2D autocorrelation descriptors are difficult to interpret. 2D autocorrelation descriptors cannot offer the specific positions of the atoms since they encode global and dimension-limited information, but the inclusion of atomic property weighting schemes brings greater applicability. As a result, these descriptors address the topology of the structure or parts thereof in association with selected physicochemical properties. By contrast, reductionistic approaches (chemical interpretation in terms of local properties, functional groups, additive schemes based on molecular fragments or on atomic types, etc.) are interpretable, but encompass rather limited information.[36]

In order to interpret our results, we evaluated the relevance of the physicochemical properties in each linear and nonlinear model. For this, we chose to estimate the relative contribution of each descriptor in the MMP inhibitory activity. The descriptor under study was removed from the model and mean of the absolute deviation values Δmi between the observed and estimated value for all compounds was calculated. Finally, the contribution Ci[37] of descriptor i is given by:

$$Ci = \frac{100 \times \Delta mi}{\sum \Delta mi} \tag{4}$$

The contributions (Ci) of descriptors weighted by the same physicochemical property were added up, in this way the contribution Cp_k was obtained ($p_k = m, v, e,$ and p). The contribution profiles of the atomic properties are given in Figure 4.

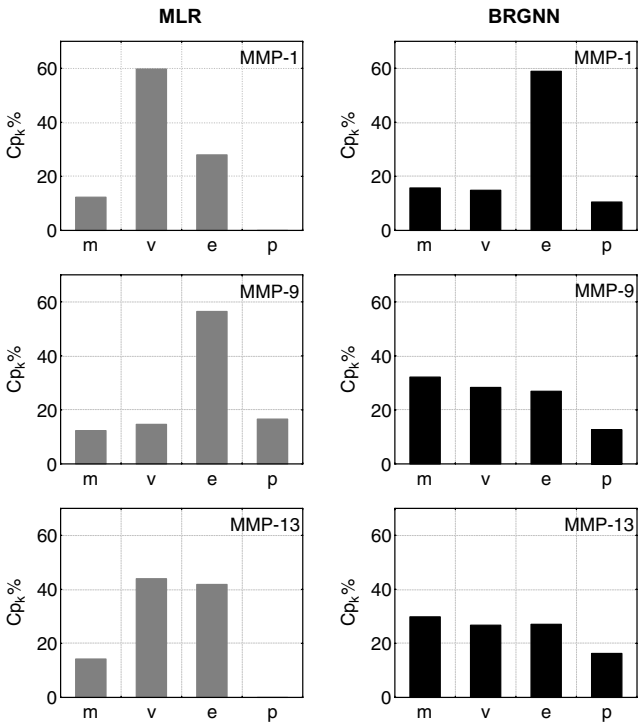


Figure 4. Contribution profile of atomic properties (Cp_k) for linear (left) and nonlinear (right) models.

The linear approach leads to the following results:

- Inhibition of MMP-1 is greatly influenced by atomic van der Waals volume terms ($Cv = 60\%$).
- Inhibition of MMP-9 is greatly influenced by atomic Sanderson electronegativity terms ($Ce = 56\%$).
- Inhibition of MMP-13 has the same contributions of atomic van der Waals volume and Sanderson electronegativity terms ($Cv = 44\%$ and $Ce = 42\%$).

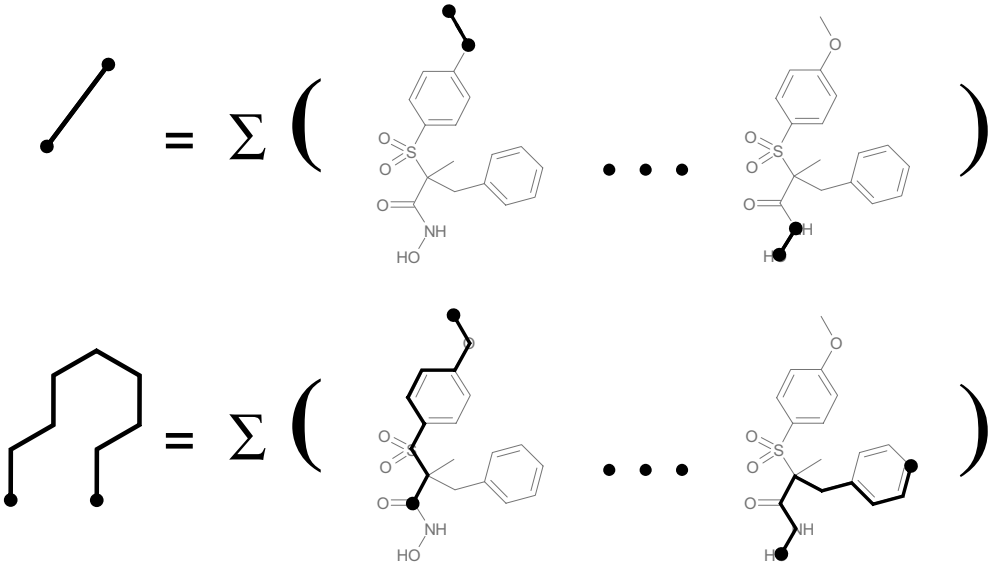


Figure 3. Representation of 2D autocorrelation terms at topological distances 1 and 8 in *N*-hydroxy-α-phenylsulfonylacetamide derivatives.

– In general, the influences of atomic mass and polarizability terms are poor ($Cm < 15\%$ and $Cp < 20\%$). Indeed, inhibition of MMP-1 and MMP-13 was not influenced by atomic polarizability terms.

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

– The model describing the inhibitory activity of MMP-1 is mainly influenced by atomic Sanderson electronegativity terms ($Ce = 59\%$), while the remaining atomic properties have a poor contribution.
– MMP-9 and MMP-13 contribution profiles are very similar. There are the same contributions of atomic mass, van der Waals volume, and Sanderson electronegativity terms ($Cp_k \approx 30\%$).
– Atomic polarizability terms have a poor contribution in all MMPs ($Cp \approx 15\%$).

From the analysis of contribution profiles we can extract the main features relevant for the design of selective inhibitors. The linear approach identified the atomic Sanderson electronegativity as the most relevant property for inhibition of MMP-9; meanwhile, the model describing MMP-13 inhibitory activity has a balanced contribution of atomic Sanderson electronegativities and van der Waals volumes. In general, it is recognized that nonpolar substituent at the P'_1 position favors MMP inhibition due to van der Waals interactions provided by the nonpolar atoms; however, the replacement of oxygen atom in R position of HPSAs affects the inhibitory potency against MMPs. Replacement of the methoxy group of **7** by a methyl (**33**) decreases inhibitory activities against all MMPs, but the bigger downfall was encountered for MMP-9 inhibition. In addition, replacement of methoxy group of **26** by bromine atom (**32**) decreased only MMP-9 inhibition. Such effects suggest that electronic surroundings in S'_1 subsite of MMP-9 affect the inhibitory activity in a bigger rate, which was reflected in MLR-MMP-9 model. Furthermore, the linear contribution profiles marked the significant contribution of atomic van der Waals volume weighted terms to MMP-1 inhibition. In this sense, it is noticed that changes in atomic volumes of substituents at the P'_1 position influence the MMP-1 inhibition: when methoxy group at R was replaced with *n*-butoxy group, inhibitory potency of MMP-13 and MMP-9 is retained while MMP-1 inhibition is significantly lowered (for instance, **26** and **29**, **23** and **30**, **34** and **49**, **50** and **51**, etc). Aside from this, when changes in groups interacting with S'_2 subsite are achieved, neither electronic changes (for instance, **2** and **16**, **34** and **35**) nor steric changes (for instance, **2** and **4**, **34** and **37**, **38** or **41**) allow establishing differences between each MMP inhibition.

On the other hand, nonlinear model describing MMP-1 inhibition pointed out the relevance of atomic Sanderson electronegativity terms, in a manner similar to that obtained for *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives.[19] Such effect can be attributed to the additional electrostatic interactions that provoke Arg214 and Asn180/Ser239 in S'_1 and S'_2 subsites of MMP-1 (Fig. 2). This is less stressed in MMP-9 and MMP-13, where hydrophobic residues replace the above-mentioned. This is further evidenced that the 2D autocorrelation space provided similar information concerning the structural features responsible for MMP-9 and MMP-13 inhibitions. The similarity between MMP-9 and MMP-13 active sites (Fig. 2), and the correlation between inhibitory activities of HPSAs against both enzymes (Table 2), is reflected by contribution profiles of nonlinear models. The best statistics and the great adequacy of nonlinear models bring reliability for predicting purposes; however, the similitude between BRGNN-MMP-9 and BRGNN-MMP-13 models does not allow finding qualitative valuations about the design of selective inhibitors. Nevertheless, the differences should be comprehended by the analysis of the different mathematical spaces defined by the 2D autocorrelation space for both models.

3. Conclusions

Predictive QSAR models were derived for *N*-hydroxy-α-phenylsulfonylacetamide derivatives, which should be useful for assisting the design of selective compounds. Such models correlate well structural features with inhibitory activities against several MMPs and bring valuable information about the relevant characteristics of inhibitors. 2D autocorrelation space was employed, obtained from different weighting schemes, viewed as an adaptive descriptor space, containing topological information able to capture structural complexity. The 2D autocorrelation descriptors appeared to capture sufficient structural detail to yield very useful results in modeling biological properties.

Linear and nonlinear models were developed by MLR combined with GA and BRGNN procedures. Nonlinear models bring more reliable statistics according to validation experiments. Different models were developed for describing inhibition of MMP-1, MMP-9, and MMP-13, and they established the relevance of electronic interactions for MMP-1 inhibition, in accordance with the increase of hydrophilic residues in its active site. Linear models were developed which, although not as statistically sound as the nonlinear models, underscore crucial requirements for selective MMPs, such as the relevance of electronic interactions in S'_1 subsite of MMP-9 and steric interactions in S'_1 subsite of MMP-1. 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.

4. Computational methods

4.1. Datasets: source and prior preparation

Inhibitions of MMP-1, MMP-9, and MMP-13 (IC_{50}) for 80 HPSAs were taken from the literature.[8,9] 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 3. 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 human MMP. Prior to molecular descriptor calculations, 3D structures of the studied compounds were geometrically optimized using the semiempirical quantum-chemical method PM3[38] implemented in the MOPAC 6.0 computer software.[39]

The data set was divided in training and test sets for each MMP inhibitory activity. Fifteen percentage of compounds out of the total ones were chosen randomly as a test set and were used for external validation for the MLR and BRGNN models. The compounds in the test sets were reserved to validate potential models. For the development of MLR and BRGNN models, the training sets included all the remaining compounds.

4.2. 2D autocorrelation pool

Three spatial autocorrelation vectors were employed for modeling the inhibitory activities: Broto–Moreau's autocorrelation coefficients (ATS) (Eq. 5),[40] Moran's indices (MATS) (Eq. 6),[41] and Geary's coefficients (GATS) (Eq. 7).[42]

$$\mathrm{ATS}(p_k, l) = \sum_i \delta_{ij} p_{ki} p_{kj} \tag{5}$$

$$\mathrm{MATS}(p_k, l) = \frac{N}{2L} \frac{\sum_{ij} \delta_{ij}(p_{ki} - \bar{p}_k)(p_{kj} - \bar{p}_k)}{\sum_i (p_{ki} - \bar{p}_k)} \tag{6}$$

$$\mathrm{GATS}(p_k, l) = \frac{(N-1)}{4L} \frac{\sum_{ij} \delta_{ij}(p_{ki} - \bar{p}_k)(p_{kj} - \bar{p}_k)}{\sum_i (p_{ki} - \bar{p}_k)} \tag{7}$$

where $\mathrm{ATS}(p_k, l)$, $\mathrm{MATS}(p_k, l)$, and $\mathrm{GATS}(p_k, l)$ are Broto–Moreau's autocorrelation coefficient, Moran's index, and Geary's coefficient at spatial lag l, respectively; p_{ki} and p_{kj} are the values of property k of atom i and j, respectively; $\bar{p}_k$ is the average value of property k, L is the number of nonzero values in the sum, N is the number of atoms in the molecule, and $\delta(\mathrm{l}, \mathrm{d}_{ij})$ is a Dirac-delta function defined as

$$\delta(l, d_{ij}) = \begin{cases} 1 & \text{if } d_{ij} = l \\ 0 & \text{if } d_{ij} \neq l \end{cases} \tag{8}$$

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

Spatial autocorrelation measures the level of interdependence between properties, and the nature and strength of that interdependence. 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 software[43] was used for calculating weighted Broto–Moreau, Moran, and Geary 2D autocorrelation vectors. Four different weighting schemes have been used: atomic masses (m), atomic van der Waals volumes (v), atomic Sanderson electronegativities (e), and atomic polarizabilities (p). Autocorrelation vectors were calculated at spatial lags l ranging from 1 up to 8 (Fig. 3). The autocorrelation descriptors are denoted by the scheme: type of descriptor-spatial lag-weighting property; for instance, GATS6v is the Geary autocorrelation of lag 6 weighted by atomic van der Waals volumes.

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. 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 48.

4.3. Modeling procedure

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,[44] 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. The quality of each model was proven by the square multiple correlation coefficient (R^2) and the standard deviation (S). The models with R^2-value above 0.8 were selected and they were tested in cross-validation experiments.

4.4. Linear GA search

Linear GA search was carried out exploring MLR models. The mean square error of data fitting was tried as the individual fitness function. An initial population of 50 individuals is randomly extracted from the data matrix in the first generation. The succeeding generations were generated by crossover and single-point mutation operators, while the two best scoring individuals were automatically retained as members for the next round of evolution. The GA search ends when 90% of the generations showed the same target fitness score. Linear GA was programmed within the MATLAB environment using the genetic algorithm toolbox.[45] The best models were selected according to R value ($R > 0.8$) and the results of cross-validation experiments (higher R^2_{CV}).

4.5. Bayesian-regularized genetic neural networks (BRGNN)

Bayesian-regularized genetic neural network (BRGNN) is a framework that combines Bayesian-regularized artificial neural networks (BRANNs) with GA feature selection.[22,23] Our BRGNN approach is a version of the So and Karplus GA feature selection method[24] incorporating Bayesian regularization.

Bayesian networks are optimal devices for solving learning problems. They diminish the inherent complexity of artificial neural networks (ANNs), being governed by Occam's Razor, when complex models are automatically self-penalizing under Bayes' rule. The Bayesian approach to ANN modeling considers all possible values of network parameters weighted by the probability of each set of weights. The BRANN method was designed by Mackay[46,47] for overcoming the deficiencies of ANNs. Bayesian approach yields a posterior distribution of network parameters $P(w|D,H)$ from a prior probability distribution $P(w|H)$ according to updates provided by the training set D using the BRANN model H. Predictions are expressed in terms of expectations with respect to this posterior distribution. Bayesian methods can simultaneously optimize the regularization constants in ANNs, a process that is very laborious using cross-validation. Instead of trying to find the global minimum, the Bayesian approach finds the (locally) most probable parameters (see in more detail in Ref. 23).

Bayesian approach produces predictors that are robust and well matched to the data. These properties become BRANNs in accurate predictors for QSAR analysis.[48,49] They give models which are relatively independent of ANN architecture, above a minimum architecture, since the Bayesian regularization method estimates the number of effective parameters. The concerns about overfitting and overtraining are also eliminated by this method so that the production of a definitive and reproducible model is attained. The joining of BRANN and GA feature selection (BRGNN) increases the possibilities of BRANNs for modeling as we indicated in previous works.[19,22,23,34] This method is relatively fast and considers the whole data set in training process. For other hybrids of ANN and GA the use of the mean square error as fitness function could lead to undesirable well fitted but poor generalized networks as algorithm solutions. In this connection, BRGNN avoids such results by two aspects: (1) keeping network architectures as simple as possible inside the GA framework and (2) implementing Bayesian regulation in the network training function.

Fully connected, three-layer BRANNs with back-propagation training were implemented in the MATLAB environment.[45] 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 $\log(10^5/IC_{50})$ values, respectively; both were normalized prior to network training. BRANN training was carried out according to the Levenberg–Marquardt optimization.[50] 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}.

The GA implemented in this paper keeps the same characteristics of the previously reported in earlier work.[22] 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 two 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. Crossover operator and single-point mutations were used in the evolution process until a 90% of the generations showed the same target fitness score. Our GA was programmed within the MATLAB environment using the genetic algorithm and neural networks toolboxes.[45] The predictors are BRANNs with a simple architecture (two or three neurons in a sole hidden layer). We tried the mean square error of data fitting for BRANN models, as the case may be, as the individual fitness function. The best models were selected according to R value ($R > 0.8$) and the results of cross-validation experiments (higher R^2_{CV}).

4.6. Analysis of the quality of the models

The quality of the fit of the training set of a specific model was measured by its R^2. However, a most important measure is the prediction quality. An internal LOO cross-validation process was carried out by estimating R^2 of LOO cross-validation (R^2_{CV}) and standard deviation (S_{CV}). A data point was removed (left-out) from the training set, and the model was 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 R^2_{CV}. In addition, the predictive power of the model was also measured by an external validation process that consists in predicting the activity of unknown compounds forming the test set. In this case, R^2 and S of test set fitting are calculated as criterion of the quality of the external predictions (R^2_{EP}, S_{EP}). Such criteria have been formulated as the requirements for a QSAR model to have highly predictive power.[51]

References and notes

1. Nagase, H.; Woessner, J. F., Jr. *J. Biol. Chem.* **1999**, *274*, 21491.
2. Shapiro, S. D. *Curr. Opin. Cell Biol.* **1998**, *10*, 602.
3. Lafleur, M.; Underwood, J. L.; Rappolee, D. A.; Werb, Z. *J. Exp. Med.* **1996**, *184*, 2311.
4. Whittaker, M. C.; Floyd, D.; Brown, P.; Gearing, A. J. H. *Chem. Rev.* **1999**, *99*, 2735.
5. Beckett, R. P.; Davidson, A. H.; Drummond, A. H.; Whittaker, M. *Drug Discov. 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. Aranapakam, V.; Grosu, G. T.; Davis, J. M.; Hu, B.; Ellingboe, J.; Baker, J. L.; Skotnicki, J. S.; Zask, A.; DiJoseph, J. F.; Sung, A.; Sharr, M. A.; Killar, L. M.; Walter, T.; Jin, G.; Cowling, R. *J. Med. Chem.* **2003**, *46*, 2361.
9. Aranapakam, V.; Davis, J. M.; Grosu, G. T.; Baker, J.; Ellingboe, J.; Zask, A.; Levin, J. I.; Sandanayaka, V. P.; Du, M.; Skotnicki, J. S.; DiJoseph, J. F.; Sung, A.; Sharr, M. A.; Killar, L. M.; Walter, T.; Jin, G.; Cowling, R.; Tillett, J.; Zhao, W.; McDevitt, J.; Xu, Z. B. *J. Med. Chem.* **2003**, *46*, 2376.
10. Hou, T.; Zhang, W.; Xu, X. *J. Comput. Aided Mol. Des.* **2002**, *16*, 27.
11. Hanessian, S.; Moitessier, N.; Therrien, E. *J. Comput. Aided Mol. Des.* **2001**, *15*, 873.
12. Kumar, D.; Gupta, S. P. *Bioorg. Med. Chem.* **2003**, *11*, 421.
13. Gupta, S. P.; Kumar, D.; Kumaran, S. A. *Bioorg. Med. Chem.* **2003**, *11*, 1975.
14. Gupta, S. P.; Kumaran, S. *Bioorg. Med. Chem.* **2003**, *11*, 3065.
15. Gupta, S. P.; Maheswaran, V.; Pande, V.; Kumar, D. *J. Enzyme Inhib. Med. Chem.* **2003**, *18*, 7.
16. Gupta, S. P.; Kumaran, S. *Bioorg. Med. Chem.* **2005**, *13*, 5454.
17. Amin, E.; Welsh, W. *J. Med. Chem.* **2001**, *44*, 3849.
18. Amin, E.; Welsh, W. *J. Chem. Inf. Model.* **2006**, *46*, 1775.
19. Fernández, M.; Caballero, J.; Tundidor-Camba, A. *Bioorg. Med. Chem.* **2006**, *14*, 4137.
20. Levin, J. I.; Du, M. T.; DiJoseph, J. F.; Killar, L. M.; Sung, A.; Walter, T.; Sharr, M. A.; Roth, C. E.; Moy, F. J.; Powers, R.; Jin, G.; Cowling, R.; Skotnicki, J. S. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 235.
21. Cronin, M. T. D.; Schultz, T. W. *J. Mol. Struct. (Theochem)* **2003**, *622*, 39.
22. Caballero, J.; Fernández, M. *J. Mol. Model.* **2006**, *12*, 168.
23. Fernández, M.; Caballero, J. *J. Mol. Graph. Model.* **2006**, *25*, 410.
24. So, S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.
25. Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1328.
26. Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Nesari, N. *J. Chem. Inf. Model.* **2005**, *45*, 190.
27. Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Taghavi, F. *J. Comput. Chem.* **2004**, *25*, 1495.
28. Mattioni, B. E.; Jurs, P. C. *J. Mol. Graph. Model.* **2003**, *21*, 391.
29. Guha, R.; Stanton, D. T.; Jurs, P. C. *J. Chem. Inf. Model.* **2005**, *45*, 1109.
30. Stanton, D. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1423.
31. Bairoch, A.; Apweiler, R. *Nucleic Acids Res.* **2000**, *28*, 45.
32. 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**, *6*, 217.
33. Welch, A. R.; Holman, C. M.; Huber, M.; Brenner, M. C.; Browner, M. F.; Van Wart, H. E. *Biochemistry* **1996**, *35*, 10103.
34. Caballero, J.; Garriga, M.; Fernández, M. *Bioorg. Med. Chem.* **2006**, *14*, 3330.
35. Fernández, M.; Tundidor-Camba, A.; Caballero, J. *Mol. Simul.* **2005**, *31*, 575.
36. Todeschini, R.; Gramatica, P. *Perspect. Drug Discov. Des.* **1998**, *9/10/11*, 355.
37. Cherqaoui, D.; Esseffar, M.; Villemin, D.; Cence, J. M.; Chastrette, M.; Zakarya, D. *New J. Chem.* **1998**, *22*, 839.
38. Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 210.
39. MOPAC version 6.0. U.S. Air Force academy: Colorado Springs CO.
40. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359.
41. Moran, P. A. P. *Biometrika* **1950**, *37*, 17.
42. Geary, R. F. *Incorporated Statistician* **1954**, *5*, 115.
43. Todeschini, R.; Consonni, V.; Pavan, M. DRAGON, version 2.1. Talete SRL: Milan, Italy.
44. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1.
45. MATLAB, version 7.0. The Mathworks Inc.: Natick, MA, http://www.mathworks.com.
46. Mackay, D. J. C. *Neural Comput.* **1992**, *4*, 415.
47. Mackay, D. J. C. *Neural Comput.* **1992**, *4*, 448.
48. Burden, F. R.; Winkler, D. A. *J. Med. Chem.* **1999**, *42*, 3183.
49. Winkler, D. A.; Burden, F. R. *Biosilico* **2004**, *2*, 104.
50. Foresee, F. D.; Hagan, M. T. Gauss-Newton Approximation to Bayesian learning. In *Proceedings of the 1997 International Joint Conference on Neural Networks*, IEEE: Houston, 1997; pp 1930.
51. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269.