

Proteochemometric Modeling of the Inhibition Complexes of Matrix Metalloproteinases with *N*-Hydroxy-2-[(Phenylsulfonyl)Amino]Acetamide Derivatives Using Topological Autocorrelation Interaction Matrix and Model Ensemble Averaging

Michael Fernández^{1,2,*}, Leyden Fernández¹, Julio Caballero^{1,3}, José Ignacio Abreu^{1,4} and Grethel Reyes¹

¹Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, Matanzas 44740, Cuba

²Department of Bioscience and Bioinformatics, Kyushu Institute of Technology (KIT), 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

³Centro de Bioinformática y Simulación Molecular, Universidad de Talca, 2 Norte 685, Casilla 721, Talca, Chile

⁴Artificial Intelligence Lab, Faculty of Informatics, University of Matanzas, Matanzas 44740, Cuba

*Corresponding author: Michael Fernández, michael_llamosa@yahoo.com; re0701m@bio.kyutech.ac.jp

A target-ligand QSAR approach using autocorrelation formalism was developed for modeling the inhibitory potency (pIC_{50}) toward matrix metalloproteinases (MMP-1, MMP-2, MMP-3, MMP-9, and MMP-13) of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives. Target and ligand structural information was encoded in the Topological Autocorrelation Interaction matrix calculated from 2D topological representation of inhibitors and protein sequences. The relevant Topological Autocorrelation Interaction descriptors were selected by genetic algorithm-based multilinear regression analysis and Bayesian-regularized genetic neural network approaches. A model ensemble strategy was employed for achieving robust and reliable linear and non-linear predictors having nine topological autocorrelation interaction descriptors with square correlation coefficients of ensemble test-set fitting (R^2_{test}) about 0.80 and 0.87, respectively. Electrostatic and hydrophobicity/hydrophilicity properties were the most relevant on the optimum models. In addition, the distribution of the inhibition complexes on a self-organized map depicted target dependence rather than an inhibitor similarity pattern.

Key words: Bayesian-regularized genetic neural networks, genetic algorithm, MMP inhibitors, QSAR analysis

Received 22 December 2007, revised 14 May 2008 and accepted for publication 15 May 2008

Matrix metalloproteinases (MMPs), a family of zinc endopeptidases, collectively degrade all components of the extracellular matrix such as collagens, proteoglycans, fibronectin, laminin, elastin, and many non-matrix 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). Because any disturbance of the generally well-balanced equilibrium between the MMPs and their physiological inhibitors can provoke pathological situations such as rheumatoid and osteoarthritis, atherosclerosis, tumor development, tumor metastasis, and pulmonary emphysema, 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 (3). Despite MMPs share certain biochemical properties, they vary in substrate specificity. Several mammalian enzymes ranging from well-characterized enzymes, such as collagenase, stromelysin, gelatinase, and membrane-type MMPs, have been identified as MMPs. In addition, 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 non-peptidic structures bearing a zinc-binding ligand (e.g., carboxylic or hydroxamic acids) has been recognized as MMP inhibitors (4–8). 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 (9,10). However, quantitative structure–activity relationship (QSAR) studies have been successfully applied for

modeling activities of MMP inhibitors (11). In recent works of our group, 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 non-linear genetic algorithm (GA) feature selection (12,13).

In both previous studies, classical ligand-based QSARs were developed for deriving individual models for each MMP family member. In the current paper, we also applied the 2D autocorrelation methodology to a set of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives (HPSAAs) (the chemical structures are shown in Table 1) that show inhibitory activities against several MMP family members (MMP-1, MMP-2, MMP-3, MMP-9, and MMP-13), but a target–ligand QSAR approach named proteochemometrics (PCM) (14) was employed for modeling the inhibition of MMPs. The structural infor-

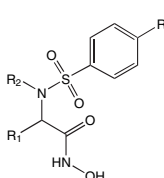
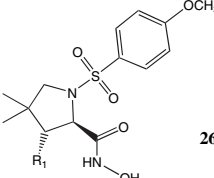
mation of the target MMPs was encoded in Amino Acids Sequence Autocorrelation (AASA) vectors, a structure encoding scheme of protein sequence previously reported by us in proteometrics studies (15–18). Afterwards, the target–ligand Topological Autocorrelation Interaction (TAI) matrix was computed as the matrix product of the MMPs AASA vectors and inhibitor 2D autocorrelation descriptors. Target–ligand QSAR studies were performed by both GA-based multilinear regression analysis (GA-MRA) and Bayesian-regularized genetic neural network (BRGNN) approaches.

Computational Methods

Data sets: source and prior preparation

In order to study the inhibition of MMP family, a series of 32 HPSAAs having well-distributed inhibitory activity (IC₅₀) over five

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
HPSAA ^a	<i>R</i> ₁	<i>R</i> 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	—	

^aHPSAA 1–6, 22, and 23 are from (6); 26–32 are from (7) and 7–21, 24, and 25 are from (8).

^bRacemic pair.

MMPs – MMP-1, MMP-2, MMP-3, MMP-9, and MMP-13 – was taken from the literature (6–8). For modeling, IC_{50} activities were converted in logarithmic activities $pIC_{50} = -\ln(IC_{50})$. IC_{50} , a measurement of drug effectiveness, is the functional strength of the inhibitor. The chemical structures and inhibitory potencies (pIC_{50}) are shown in Tables 1 and S1, respectively. 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 (19) implemented in the MOPAC 6.0^a computer software.

Amino acid sequences of the five human MMP family members (primary accession number in parentheses) – MMP-1 (P03956), MMP-2 (P08253), MMP-3 (P08254), MMP-9 (P14780), and MMP-13 (P45452) – were obtained from the Swiss-Prot/TrEMBL database (20).

Proteochemometric modeling

Proteochemometrics, proposed by Wikberg (14), originates from chemometrics, the mathematical methods used to analyze chemical data. PCM models describe the interactions between a series of macromolecules (such as proteins) and a series of ligands. These models are useful for predicting the affinities of new proteins for their ligands if the new molecules fall within the description space of the protein-ligand pairs of the training data set. Similarly, one PCM model can predict the affinity of new ligands toward a group of related targets. A PCM experiment is typically described by three descriptor blocks: the ligand descriptor (DL), protein descriptor (DP), and ligand-protein cross-term (DLP) blocks. A vector of variables, called the ligand descriptors (DL), characterizes each ligand L . Similarly, each protein P has its protein descriptors (DP). Depending on the problem faced, one or more descriptor blocks can be discarded. In our study, a DLP block called TAI matrix was calculated as the matrix product of the inhibitor 2D autocorrelation vectors and the AASA vectors of the MMPs. The TAI matrix was used for deriving linear and non-linear models of the inhibition of five MMPs (MMP-1, MMP-2, MMP-3, MMP-9, and MMP-13) by GA-MRA and BRGNN function mapping techniques.

2D spatial autocorrelation vectors

The binding of a ligand to a target depends on the shape of the ligand and on a variety of effects such as the molecular electrostatic potential, polarizability, hydrophobicity and lipophobicity. Therefore, in a QSAR study, the strategy for encoding molecular information must in some way, either explicitly or implicitly, account for these physicochemical effects. Furthermore, usually data sets include molecules of different sizes with different numbers of atoms, and so the structural encoding schemes must allow comparing such molecules. Thus, we were faced with the problem of having to compare molecules with different numbers of atoms. Information of variable length can be transformed into fixed-length information by autocorrelation (21).

Autocorrelation vectors have several useful properties. First, a substantial reduction in data can be achieved by limiting the topological distance, l . Second, the autocorrelation coefficients are independent of the original atom numberings, and so they are canonical. Third, the length of the correlation vector is independent of the size of the molecule (21).

For the autocorrelation vectors, H-depleted molecular structure is represented as a graph, and physico-chemical properties of atoms (i.e., atomic van der Waals volumes, atomic Sanderson electronegativities, and atomic polarizabilities) as real values assigned to the graph vertices (Figure 1).

These descriptors can be obtained by summing up the products of certain properties of two atoms, located at given topological distances or spatial lag in the molecular graph. Broto-Moreau's autocorrelation vectors were employed for encoding the topological structure of the MMP inhibitors.

Broto-Moreau's autocorrelation coefficient (22) is:

$$ATS/p_k = \sum_i \delta_{ij} p_{ki} p_{kj} \quad (1)$$

where ATS/p_k is Broto-Moreau's autocorrelation coefficient at spatial lag l ; 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 non-zero elements in the sum; and $\delta(l, d_{ij})$ is a Dirac-delta function defined as

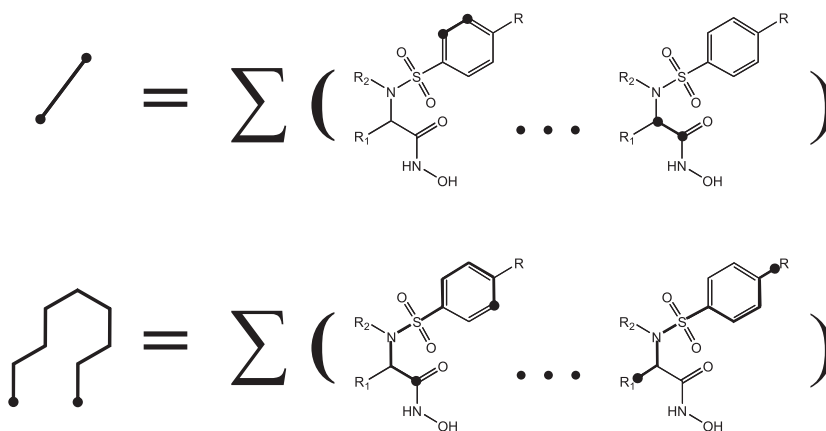


Figure 1: Representation of 2D autocorrelation terms at topological distances 1 and 8 in generic *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivative.

$$\delta(l, d_{ij}) = \begin{cases} 1 & \text{if } d_{ij} = l \\ 0 & \text{if } d_{ij} \neq l \end{cases} \quad (2)$$

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

Dragon computer software^b was used for calculating the 2D autocorrelation vectors at spatial lags ranging from 1 to 8 and weighted by three atomic properties: atomic van der Waals volumes, atomic Sanderson electronegativities, and atomic polarizabilities; thus, a total of 24 (8×3) 2D autocorrelation vectors were computed.

Amino Acid Sequence Autocorrelation vectors

Protein-ligand interactions depend on a variety of intramolecular interactions, such as hydrophobic, electrostatic, van der Waals, and hydrogen bond, which are ruled by the amino acid sequence and the chemical characteristics of the ligand. Therefore, in structure–property/activity relationship studies, the strategy for encoding protein structural information must in some way, either explicitly or implicitly, account for these interactions. The autocorrelation vector formalism can be easily extended to amino acid sequences considering protein primary structure as a linear graph with nodes formed by amino acid residues. We recently introduced the AASA vectors for modeling the functional variations upon mutation of the ghrelin receptor (16) and the conformational stability of human lysozyme (15) and gene V protein mutants (17). The calculated autocorrelation vectors encode structural information concerning the whole protein. Particularly, AASA vectors of lag l are calculated as follows:

$$AASA/p_k = \frac{1}{L} \sum_i \delta_{ij} p_{ki} p_{kj} \quad (3)$$

where $AASA/p_k$ is the AASA at spatial lag l weighted by the p_k property; L is the number of elements in the sum; p_{ki} and p_{kj} are the values of property k of amino acids i and j in the sequence, respectively, and $\delta(l, d_{ij})$ is the Dirac–delta function in eqn 2.

For example, if we consider the decapeptide ASTCGFHCS, AASA vectors at spatial lag 1 and 5 are calculated as follows:

$$AASA1p_k = \frac{1}{9} (p_{kA} \cdot p_{kS} + p_{kS} \cdot p_{kT} + p_{kT} \cdot p_{kC} + p_{kC} \cdot p_{kG} + p_{kG} \cdot p_{kF} + p_{kF} \cdot p_{kH} + p_{kH} \cdot p_{kC} + p_{kC} \cdot p_{kS} + p_{kS} \cdot p_{kD}) \quad (4)$$

$$AASA5p_k = \frac{1}{5} (p_{kA} \cdot p_{kF} + p_{kS} \cdot p_{kH} + p_{kT} \cdot p_{kC} + p_{kC} \cdot p_{kS} + p_{kG} \cdot p_{kD}) \quad (5)$$

In a protein, autocorrelation analysis will test whether the value of a property at one residue is independent of the values of the property at neighboring residues. If dependence exists, the property is said to exhibit spatial autocorrelation. AASA vectors represent the degree of similarity between amino acid sequences.

As weights for sequence residues we used seven physicochemical and conformational amino acid/residue properties (Table S2 in

Supplementary Material) selected from the AAindex database (23). In our work, spatial lag, l , was ranging from 1 to 5. Computational code for AASA vector calculation was written in MATLAB environment^c. A data matrix of 35 AASA vectors, 7 properties \times 5 different lags, were generated with the autocorrelation vectors calculated for each MMP family member considering whole protein sequence.

Topological Autocorrelation Interaction matrix

The TAI matrix was calculated as the matrix product of the AASA vectors of the five MMPs and the inhibitor 2D autocorrelation vectors, resulting in 480 TAI descriptors, 35 AASA vectors \times 24 2D autocorrelation vectors. TAI descriptors are calculated as follows:

$$TAI/p_k^1 p_o^2 = AASA/p_k^1 \times ATSl/p_o^2 \quad (6)$$

where $TAI/p_k^1 p_o^2$ is the TAI at spatial lag l_1 in the protein sequence weighted by the amino acid and/or residue property p_k^1 and at spatial lag l_2 in the ligand topological structure weighted by atomic property p_o^2 ; $AASA/p_k^1$ is the AASA at spatial lag l_1 in the protein sequence weighted by the amino acid and/or residue property p_k^1 ; and $ATSl/p_o^2$ is Broto–Moreau's autocorrelation coefficient at spatial lag l_2 weighted by the atomic property p_o^2 .

Descriptors that stayed constant or almost constant were eliminated, and pairs of variables with an absolute value of correlation coefficient (R) greater than 0.95 were classified as intercorrelated, and only one of these was included for building the model. Finally, 60 TAI descriptors were obtained. Afterwards, optimum predictive models were built with reduced subsets of variables by means of GA-MRA and BRGNN algorithm.

Genetic algorithm search

Linear GA search was carried out exploring MRA models. The mean square error (MSE) of data fitting was tried as the individual fitness function. An initial population of 100 individuals was randomly extracted from the data matrix in the first generation. The succeeding generations were generated by crossover and single-point mutation operators, while the 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. The best models were selected according to R value ($R > 0.8$) and the results of leave-one-out (LOO) and three-fold-out (TFO) crossvalidation experiments (higher values of R_{LOO}^2 and R_{TFO}^2).

Bayesian-regularized genetic neural network

Bayesian-regularized genetic neural network is a framework that combines Bayesian-regularized artificial neural network (BRANN) (24,25) and GA feature selection (26,27). Our BRGNN approach was a version of the So and Karplus GA feature selection method (28) incorporating Bayesian regularization. Bayesian networks are optimal devices for solving learning problems. They diminish the inher-

ent complexity of artificial neural networks (ANNs), being governed by Occam's razor, when complex models are automatically self-penalized 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 (24,25) 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 crossvalidation.

Instead of trying to find the global minimum, the Bayesian approach finds the (locally) most probable parameters [see (24,25) for more details]. Bayesian approach produces predictors that are robust and well matched to the data. These properties become BRANNs in accurate predictors for QSAR analysis (29,30). They give models that are relatively independent of ANN architecture, above a minimum architecture, because the Bayesian regularization method estimates the number of effective parameters. The concern about overfitting and overtraining is 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 indicated by us in the previous works (12,13,15–18,26,27,31–36). 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 MSE as fitness function could lead to undesirable well-fitted but poorly generalized networks as algorithm solutions. In this connection, BRGNN avoids such results by two aspects: (i) keeping network architecture as simple as possible inside the GA framework and (ii) implementing Bayesian regulation in the network training function.

Fully connected, three-layer BRANNs with back-propagation training were implemented in the MATLAB environment^c. 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 pIC_{50} values, respectively; both were normalized prior to network training. BRANN training was carried out according to Levenberg–Marquardt optimization (37). The initial value for λ was 0.005 with decrease and increase factors of 0.1 and 10, respectively. The training was stopped when λ became larger than 10 (10).

The GA implemented in this paper keeps the same characteristics of the ones previously reported in earlier works (26,27). Initially, a set of chromosomes were randomly generated. The population fitness was then calculated, and the members were rank-ordered according to fitness. The 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 90% of the generations showed the same target fitness score. The predictors are BRANNs with a simple architecture (two or three neurons in a sole hidden layer). We tried the MSE of data fitting

for BRANN models, as the case may be, as the individual fitness function. Finally, from crossvalidation experiments over the subpopulation of well-fitted models, it can derive the best generalizable network with the highest predictive power. The best models were selected according to R value ($R > 0.8$) and the results of LOO and TFO crossvalidation experiments (higher R^2_{LOO} and R^2_{TFO}). BRGNN toolbox for MATLAB^d was programmed within the MATLAB environment using the Genetic Algorithm and Direct Search^e and Neural Networks^f toolboxes.

Ensemble averaging

Model ensemble (ME) is a learning paradigm where many predictors are jointly used to solve a problem (38). On the basis of this judgment, a collection of a finite number of predictors is trained for the same task and the outputs can be combined to form one unified prediction. As a result, the generalization ability of the system can be significantly improved by reducing overfitting (39). Recently, Baumann (40) demonstrated that ensemble averaging significantly improves prediction accuracy by assembling the predictions of several models that are obtained in parallel with bootstrapped training sets and provide a more realistic meaning of the predictive capacity of any regression model (MRA, partial least squares, ANNs).

Model diversity can be introduced by manipulating the input features (feature selection), randomizing the training procedure (overfitting, underfitting, training with different topologies and/or training parameters, etc.), manipulating the response value (adding noise), or manipulating the training set (41). Because linear models with fixed number of variables have a unique topology and BRANN predictors have demonstrated to be highly stable to network topology variations (29,30), the latter method, i.e., manipulating the training set, was used for introducing diversity in MRA and BRGNN ensembles.

Data-diverse MEs were previously used by us for BRGNN model validation (15,17,18,27). For generating the constituent predictors, we partitioned the whole data into several training and test sets. The assembled predictors aggregate their outputs to produce a single prediction. In this way, instead of predicting a sole randomly selected external set; we predicted the result of averaging several ones. In this way, each target–inhibitor complex was predicted several times forming training and test sets, and an average of values was calculated. The ensemble predictive power was measured accounting R^2_{test} and $\text{RMSE}_{\text{test}}$ values of the averaged test set of MRA and BRGNN ensembles having an optimum number of members.

Self-organizing maps

In order to settle structural similarities among the inhibitory potency of HPSAAs toward five MMPs, a self-organizing map (SOM) was built. Kohonen (42) introduced a neural network model that generates SOMs. In such maps, molecules with similar descriptor vectors are projected into the same or closely adjacent neurons (43). These networks have been widely used for addressing structural similarities among chemical data sets (44). SOMs were implemented in MATLAB environment; neurons were initially located at a grid topology. The ordering phase was developed in 1000 steps with 0.9 learning rate until tuning neighborhood distance (1.0) was

achieved. The tuning phase learning rate was 0.02. Training was performed for a period of 2000 epochs in an unsupervised manner.

Results and Discussion

The topology and the nature of the residues in the active sites of MMPs are highly conserved among the different MMP family members. However, our QSAR study was based on previous structure–activity relationship study in which authors modified functional groups at the $P'1$, $P1$, and $P'2$ sites (R , $R1$, and $R2$ in Figure 2) of the inhibitors in Table 1 as functional probes for $S'1$, $S1$, and $S'2$ subsites of MMP family members (6–8), searching for selectivity in accordance with the differences between these pockets. Correlation matrix (Table 2) shows that inhibitory activities of HPSAAs employed in this study against five MMP family members correlated to each other to some extent. Despite the inhibitory potency (pIC_{50}) against MMP-1 being poor for these compounds, the activities against collagenases (MMP-1 and MMP-13) and gelatinases (MMP-2 and MMP-9) correlate to a large extent with R^2 higher than 0.7. The target–ligand interaction information gathered in the TAI matrix accounts for the specificity of MMPs for the inhibitors. TAI matrix was explored by both linear and non-linear techniques in order to achieve general predictive models for the inhibition of MMPs; in this way, specificity-encoding models were built by varying the dimensions of the target–ligand interaction subspaces from 3 to 12 inside the GA frameworks.

Optimum linear and non-linear MMP/HPSAA complexes interaction subspaces

Genetic algorithm-based multilinear regression analysis

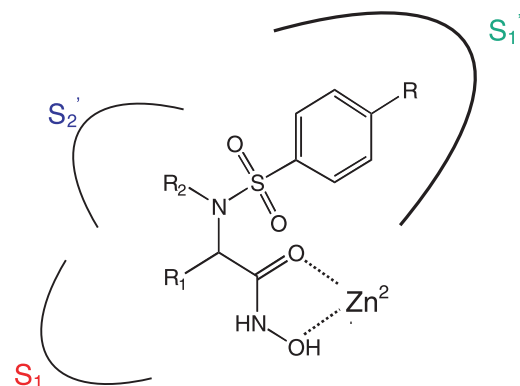
In the first approach, a MRA model for the inhibitory potency of the studied target–ligand complexes was achieved by means of GA search routine. The model selection was subjected to the principle of parsimony (45). Then, we chose a function with high statistical signification but having fewer descriptors as possible. Two MMP/HPSAA complexes (MMP-2/HPSAA-16 and MMP-9/HPSAA-21) were removed as outliers because of their high standard deviation during the model selection process. The best linear QSAR model obtained is given below together with the statistical parameters of the regression.

$$\begin{aligned} \text{pIC}_{50} = & 1.5373 \times \text{TAI}5f3v + 9.7097 \times \text{TAI}2\text{Ht}8p + 2.9598 \\ & \times \text{TAI}5R_x6e + 18.4776 \times \text{TAI}2pK'2v - 9.2497 \times \text{TAI}3pK'3e \\ & + 5.3590 \times \text{TAI}1pK'5v - 0.01469 \times \text{TAI}5V^{\circ}2v \\ & + 7.2279 \times \text{TAI}5s4v - 4.0983 \times \text{TAI}1s6p - 5.6160 \end{aligned} \quad (7)$$

$$N = 147; R^2 = 0.824; S = 0.476; F = 71.314; p < 10^{-5}$$

$$R^2_{\text{LOO}} = 0.799; S_{\text{LOO}} = 0.491; R^2_{\text{TFO}} = 0.794; S_{\text{TFO}} = 0.500$$

where pIC_{50} is the studied property; N is the number of compounds included in the model; R^2 is the square correlation coefficient; S is the standard deviation of the regression; F is the Fisher ratio; p is



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

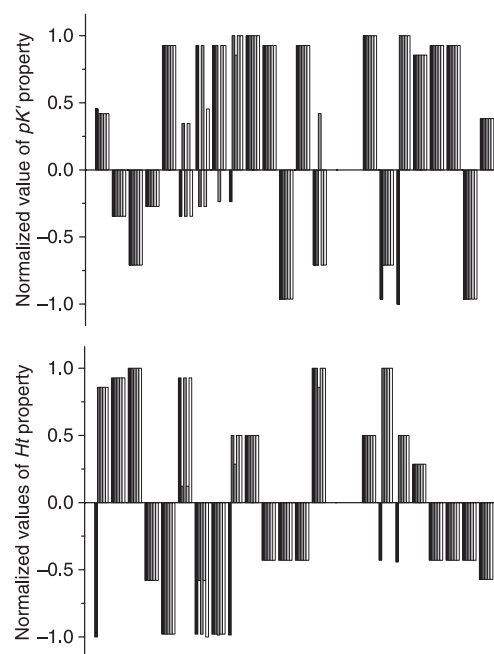


Figure 2: Position of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives inside MMP active site and distributions of the most relevant amino acid/residue properties pK' and Ht along the MMP pockets. Comparison of amino acid sequences of MMPs. Colored letters indicate the amino acids of $S'1$ (green), $S1$ (red) and $S'2$ (blue) pockets that contribute to ligand specificity.

Table 2: 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

the significance of the variables in the model; R^2_{LOO} and R^2_{TFO} are the square correlation coefficients of the LOO and TFO crossvalidations, respectively; and S_{LOO} and S_{TFO} are the standard deviations of the LOO and TFO crossvalidations, respectively. This nine-*TAI* descriptor model (*TAI*-MRA) explains nearly 80% of inhibitory potency in crossvalidation test as having very good predictive power.

The variables in *TAI*-MRA model in eqn 7 mean the following: *TAI5f3v* is the *TAI* of lag 5 weighted by flexibility in the target sequence and lag 3 weighted by van der Waals volume in the ligand; *TAI2Ht8p* is the *TAI* of lag 2 weighted by thermodynamic transfer hydrophobicity in the target sequence and lag 8 weighted by polarizability in the ligand; *TAI5R_x6e* is the *TAI* of lag 5 weighted by solvent-accessible reduction ratio in the target sequence and lag 6 weighted by electronegativity in the ligand; *TAI2pK'2v* is the *TAI* of lag 2 weighted by equilibrium constant with reference to the ionization property of COOH group in the target sequence and lag 2 weighted by van der Waals volume in the ligand; *TAI3pK'3e* is the *TAI* of lag 3 weighted by equilibrium constant with reference to the ionization property of COOH group in the target sequence and lag 3 weighted by electronegativity in the ligand; *TAI1pK'5v* is the *TAI* of lag 1 weighted by equilibrium constant with reference to the ionization property of COOH group in the target sequence and lag 5 weighted by van der Waals volume in the ligand; *TAI5V°2v* is the *TAI* of lag 5 weighted by partial specific volume in the target sequence and lag 2 weighted by van der Waals volume in the ligand; *TAI5s4v* is the *TAI* of lag 5 weighted by shape (position of branch point in a side chain) in the target sequence and lag 4 weighted by van der Waals volume in the ligand; and *TAI1s6p* is the *TAI* of lag 1 weighted by shape (position of branch point in a side chain) in the target sequence and lag 6 weighted by polarizability in the ligand. As can be observed in Table 3, there is no significant intercorrelation among selected descriptors, and so different information is brought to the model by each *TAI* descriptor.

According to eqn 7, MMPs–HPSAA affinity linearly depends on the interactions of six out of the seven properties related to the target protein – flexibility *f*, thermodynamic transfer hydrophobicity *Ht*, solvent-accessible reduction ratio *R_x*, equilibrium constant with reference to the ionization property of COOH group *pK'*, partial specific volume *V*, and shape (position of branch point in a side chain) *s* – with the three atomic properties related to the ligands – atomic

van der Waals volumes, atomic Sanderson electronegativities, and atomic polarizabilities.

Bayesian-regularized genetic neural networks

Despite the adequate results found by linear GA, we carried out an additional non-linear search for exploring other possibilities. Recently, we proposed the BRGNN approach (26,27), which surpassed the limits of the linear solutions when modeling inhibitory activities (12,13,15–18,26,27,31–36). 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 non-linear GA search is developed. In our current application, ANN architectures were varied testing different quantities of neurons in hidden layers. Similar to the linear attempt, we searched optimum models varying the dimension of the training subspace from 3 to 12.

Topological Autocorrelation Interaction descriptors and statistics of the optimum BRGNN predictor appear in Table 4. As can be observed, optimum predictor was found having three hidden nodes. The same two MMP/HPSAA complexes (MMP-2/HPSAA-16 and MMP-9/HPSAA-21) were removed as outliers because of their high standard deviation during the non-linear model selection process. The optimum non-linear model *TAI*-BRGNN 2 describes about 92%

Table 4: Statistics of the optimum *TAI*-BRGNN predictors for the inhibition of MMPs by HPSAAs. Optimum neural network predictor appears in bold letter

<i>TAI</i> descriptors: <i>TAI5f8p</i> , <i>TAI5s3e</i> , <i>TAI2pK'1v</i> , <i>TAI2s3v</i> , <i>TAI2pK'8e</i> , <i>TAI2pK'2v</i> , <i>TAI1Ht1v</i> , <i>TAI1Ht5v</i> , <i>TAI4s2p</i>							
Model	hidd. nod.	R^2	<i>S</i>	R^2_{LOO}	S_{LOO}	R^2_{TFO}	S_{TFO}
1	2	0.856	0.416	0.773	0.523	0.785	0.512
2	3	0.923	0.304	0.878	0.383	0.864	0.405
3	4	0.932	0.285	0.860	0.412	0.840	0.443
4	5	0.936	0.278	0.833	0.453	0.822	0.466
5	6	0.947	0.252	0.803	0.488	0.706	0.616

hidd. nod. represents the number of hidden nodes; R^2 , R^2_{LOO} , and R^2_{TFO} are square correlation coefficients of data set fitting, LOO, and TFO crossvalidations, respectively; *S*, S_{LOO} , and S_{TFO} are standard deviations of data set fitting, LOO, and TFO crossvalidations, respectively.

Table 3: Correlation matrix of the inputs of the optimum linear predictor *TAI*-MRA

	<i>TAI5f3v</i>	<i>TAI2Ht8p</i>	<i>TAI5R_x6e</i>	<i>TAI2pK'2v</i>	<i>TAI3pK'3e</i>	<i>TAI1pK'5v</i>	<i>TAI5V°2v</i>	<i>TAI5s4v</i>	<i>TAI1s6p</i>
<i>TAI5f3v</i>	1.000	0.064	0.459	0.018	0.005	0.014	0.545	0.145	0.184
<i>TAI2Ht8p</i>		1.000	0.043	0.324	0.075	0.029	0.369	0.000	0.223
<i>TAI5R_x6e</i>			1.000	0.013	0.013	0.184	0.225	0.000	0.110
<i>TAI2pK'2v</i>				1.000	0.023	0.101	0.486	0.024	0.064
<i>TAI3pK'3e</i>					1.000	0.019	0.070	0.030	0.022
<i>TAI1pK'5v</i>						1.000	0.035	0.052	0.081
<i>TAI5V°2v</i>							1.000	0.099	0.334
<i>TAI5s4v</i>								1.000	0.482
<i>TAI1s6p</i>									1.000

variance of data fitting and about 85% variance of the data in LOO and TFO crossvalidation processes. The internal predictive power of the non-linear model is more than 5% points higher in comparison to model *TAI-MRA* in eqn 7. The superior behavior of the non-linear models describing the inhibitory potency of HPSAA toward the five MMPs studied suggests that the *TAI* descriptors built a non-linear vectorial space that better resembles the target–ligand interactions in comparison to the linear one.

With respect to the possibility of change correlations following the method used by So and Karplus (34), we performed a randomization test. Randomized values were given to the dependent variable (pIC_{50}), and networks were trained using this randomized target and the real set of independent variables (optimum *TAI* descriptors). By repeating this processes 500 times, no correlation was found between R^2 values for training and test sets, similar to the results of So and Karplus (28).

The variables in optimum non-linear model *TAI-BRNN 2* in Table 4 mean the following: *TAI5f8p* is the *TAI* of lag 5 weighted by flexibility in the target sequence and lag 8 weighted by polarizability volume in the ligand; *TAI5s3e* is the *TAI* of lag 5 weighted by shape (position of branch point in a side chain) in the target sequence and lag 3 weighted by electronegativity in the ligand; *TAI2pK'1v* is the *TAI* of lag 2 weighted by equilibrium constant with reference to the ionization property of COOH group in the target sequence and lag 1 weighted by van der Waals volume in the ligand; *TAI2s3v* is the *TAI* of lag 2 weighted by shape (position of branch point in a side chain) in the target sequence and lag 3 weighted by van der Waals volume in the ligand; *TAI2pK'8e* is the *TAI* of lag 2 weighted by equilibrium constant with reference to the ionization property of COOH group in the target sequence and lag 8 weighted by electronegativity in the ligand; *TAI2pK'2v* is the *TAI* of lag 2 weighted by equilibrium constant with reference to the ionization property of COOH group in the target sequence and lag 2 weighted by van der Waals volume in the ligand; *TAI1Ht1v* is the *TAI* of lag 1 weighted by thermodynamic transfer hydrophobicity in the target sequence and lag 1 weighted by van der Waals volume in the ligand; *TAI1Ht5v* is the *TAI* of lag 1 weighted by thermodynamic transfer hydrophobicity in the target sequence and lag 5 weighted by van der Waals volume in the ligand and *TAI4s2p* is the *TAI* of lag 4 weighted by shape (position of branch point in a side chain) in the target sequence and lag 2 weighted by polarizability in the ligand. As can be observed in Table 5, only three intercorrelations appear significant ($R^2 > 0.7$): *TAI5s3e* versus *TAI2s3v*,

TAI2pK'1v versus *TAI2pK'2v*, and *TAI2s3v* versus *TAI4s2p*. Despite this level intercorrelation, the adequate fitting of the data set and the crossvalidations obtained by such descriptor subset reflect that relevant structural information is brought to the model by each *TAI* descriptor.

According to model *TAI-BRNN 2* and differently to model *TAI-MRA*, MMP-HPSAA affinity non-linearly depends on the interactions of four out of the seven properties related to the target protein – flexibility *f*, thermodynamic transfer hydrophobicity *Ht*, equilibrium constant with reference to the ionization property of COOH group pK' , and shape (position of branch point in a side chain) *s* – with the three atomic properties related to the ligands – atomic van der Waals volumes, atomic Sanderson electronegativities, and atomic polarizabilities. This fact suggests that the linear model gathered linear contributions from many properties for resembling a relevant and simple linear target–ligand interaction space. The six protein-related properties in eqn 7 are related to four main features: hydrophobicity–hydrophilicity balance (properties *Ht*, R_{H}), electrostatic state (property pK'), shape/size distribution (properties V^0 , *s*), and freedom degrees or enthalpy level (property *f*) along the target sequence. On the contrary, the non-linear model establishes a more relievable predictor by gathering information from fewer protein-related properties (four instead of six) but depicting a more reliable and convoluted target–ligand interaction space. It is noteworthy that the non-linear models also account for the four main features related to the above, but only one property per feature is needed for encompassing the proper information from the target sequence: hydrophobicity–hydrophilicity balance (property *Ht*), electrostatic state (property pK'), shape/size distribution (property *s*), and freedom degrees or enthalpy level (property *f*). However, both optimum linear and non-linear models exhibited contributions of all three atomic properties used, suggesting that, from the ligand's point of view, this is necessary to account for electronic, hydrophobicity–hydrophilicity, and size distributions in order to adequately resemble the contribution of MMP inhibitors to the target–ligand interaction space.

Data-diverse linear and non-linear ensembles

In order to build robust models, we assembled linear equations and networks instead of using single linear equation or network to calculate the inhibitory potency of HPSAA toward MMPs. Ensembles of the eqn 7 and neural network *TAI-BRNN 2* were built by varying training and test data. This approach consists of

	<i>TAI5f8p</i>	<i>TAI5s3e</i>	<i>TAI2pK'1v</i>	<i>TAI2s3v</i>	<i>TAI2pK'8e</i>	<i>TAI2pK'2v</i>	<i>TAI1Ht1v</i>	<i>TAI1Ht5v</i>	<i>TAI4s2p</i>
<i>TAI5f8p</i>	1	0.119	0.019	0.151	0.063	0.015	0.233	0.171	0.08
<i>TAI5s3e</i>		1	0.004	0.771	0.024	0.002	0.291	0.17	0.679
<i>TAI2pK'1v</i>			1	0.004	0.035	0.788	0.199	0.141	0.12
<i>TAI2s3v</i>				1	0.004	0.031	0.148	0.135	0.759
<i>TAI2pK'8e</i>					1	0.016	0.017	0.09	0.005
<i>TAI2pK'2v</i>						1	0.174	0.036	0.118
<i>TAI1Ht1v</i>							1	0.618	0.1
<i>TAI1Ht5v</i>								1	0.064
<i>TAI4s2p</i>									1

Table 5: Correlation matrix of the inputs of the optimum linear predictor *TAI-BRNN 2*. High intercorrelations ($R^2 > 0.7$) appear in bold letter

training several predictors with different randomly partitioned training sets (67% of the data) and predicting the inhibitory potency of the rest of the target–inhibitor complex (33% of the data) in test sets. In this regard, the outputs of the trained models were combined to form one unified prediction. In this sense, we reported calculated pIC_{50} values for each target–inhibitor complex averaged over the test sets (Table S1 in Supplementary Material). The optimum number of elements in the ensemble predictor was selected by studying the behavior of ensemble root mean square error (RMSE) of test sets ($\text{RMSE}_{\text{test}}$) versus the number of predictors in the ensemble. Concerning this, Figure 3 depicts the plots of $\text{RMSE}_{\text{test}}$ values for linear and non-linear ensembles with number

of members varying from 2 to 100. Beside the lower $\text{RMSE}_{\text{test}}$ values of the ensembles of BRGNNs in comparison to the ensembles of linear equations, the mean of this statistical quantity tends to decrease with the increment of assembled networks, and the $\text{RMSE}_{\text{test}}$ values have lower dispersions (Figure 3B). However, the mean $\text{RMSE}_{\text{test}}$ values for the linear ensembles in Figure 3A slightly decrease with the increment of its members, but the dispersions of the $\text{RMSE}_{\text{test}}$ values significantly decrease. These differences in the behaviors of linear and non-linear ensembles are related to the higher intrinsic diversity of non-linear models in comparison to the linear ones. In this regard, model diversity has been reported as a property that accounts for the decrease of the ensemble error in comparison to the errors of single predictors (40) represented as start dots in Figure 3.

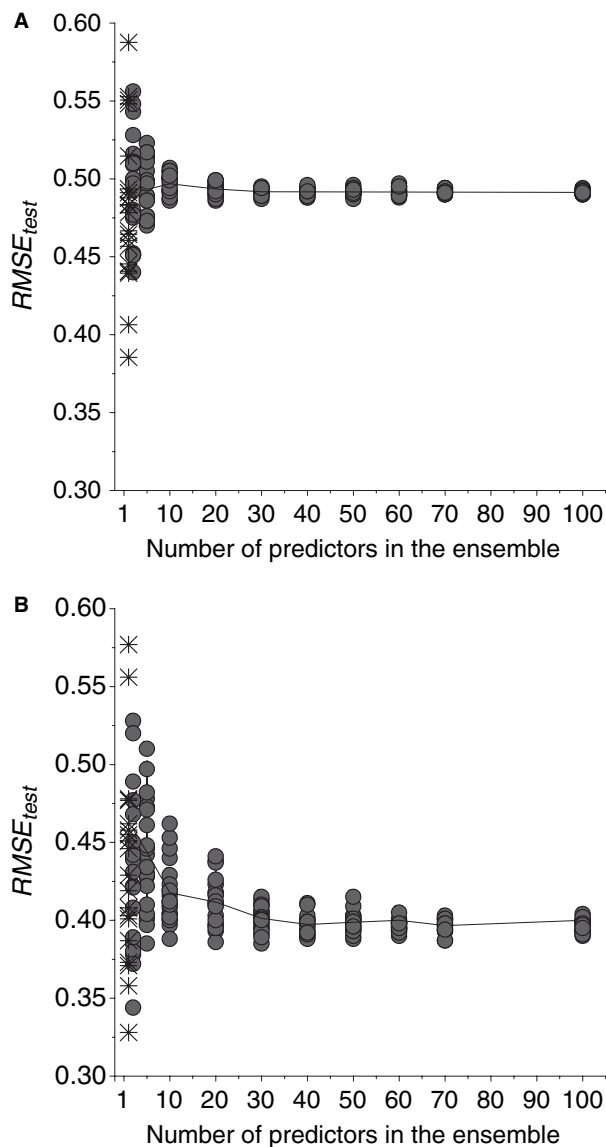


Figure 3: Plots of RMSE of average test set values ($\text{RMSE}_{\text{test}}$) of the inhibitory potency (pIC_{50}) of HPSAAs toward MMP family members according to 20 linear (A) and 30 non-linear (B) ensembles versus number of individual predictors in each ensemble. Start dots represent single models.

Differently to the network conglomerates, the linear ensembles do not show a significant precision increase in comparison to individual linear models. However, they do exhibit a very stable behavior of the prediction errors when averaging more than 20 linear equations. Considering this, we selected the optimum linear ensemble having 20 linear equations. Otherwise, as can be observed in Figure 3B, $\text{RMSE}_{\text{test}}$ values of ensembles having more than 30 networks remained stable. Because of the higher variability of the neural networks in comparison to linear equations, 30 BRGNN (10 more predictors) are needed for achieving a reliable BRGNN ensemble.

Figure 4 depicts the plots of calculated versus experimental pIC_{50} for target–ligand complexes calculated as an average over test sets according to the conglomerates of models *TAI*-MRA and *TAI*-BRGNN 2. The accuracy for test data fitting was about 80% and 87% for the linear and non-linear ensembles, respectively. Test set average values for the inhibitory potency of MMP/HPSAA complex according to model *TAI*-BRGNN 2 appear in Table S1 in Supplementary Material. Similar to the internal validation process, *TAI* approach better fits in a non-linear way the inhibitory potency of HPSAAs toward MMPs by means of the interaction of sequence information of the target encoded in AASA vectors and ligand information encoded in 2D autocorrelation descriptors. The MMP inhibitory pattern that the optimum nine-*TAI* descriptor resembled was successfully learned by the ensemble of BRGNNs during supervised training.

Optimum *TAI*-BRGNN 2 model's interpretation and comparison with previous QSAR studies

In order to gain a deeper insight into the relative effects of each *TAI* descriptor in the model *TAI*-BRGNN 2, a weight-based input ranking scheme was carried out. Black-box nature of three-layer ANNs has been 'deciphered' by Guha *et al.* (46). Their method determines the square contribution values (SCV) for each hidden neuron [see (46) for details]. This approach for ANN model interpretation is similar in manner to the partial least squares interpretation method for linear models described by Stanton (47).

The results of the model interpretation analysis appear in Table 6. Among the three hidden nodes in the predictor *TAI*-BRGNN 2, the most ranked is node 3 having a SCV value about 0.79,

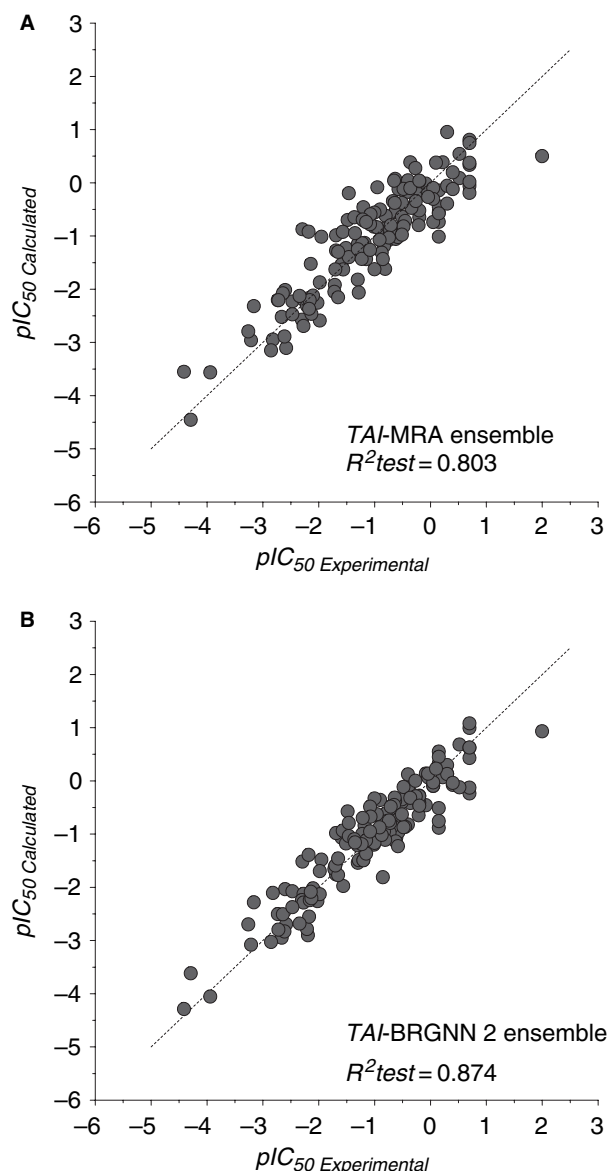


Figure 4: Plots of average calculated versus experimental inhibitory potency (pIC_{50}) of HPSAAs toward MMP family members for test sets according to 20-member linear ensemble (A) and 30-member non-linear ensemble (B).

which is 5.4-fold higher than second-ranked hidden node 1 and about 13-fold higher than hidden node 2. According to the Guha's analysis (46), the most ranked node has a major impact on the overall output of the neural network. Consequently, the most weighted inputs in such node represent the most relevant descriptors for the regression problem under study. Specifically in Table 6, TAI descriptors having weights >1.51 on the most ranked nodes 3 and 1 are the most relevant inputs. As can be observed, such descriptors are *TAI2pK^{1v}*, *TAI2pK^{8e}*, *TAI1Ht1v*, and *TAI1Ht5v*, which represent interactions of equilibrium constant with reference to the ionization property of COOH group and thermodynamic transfer hydrophobicity in the target sequence

Table 6: Effective weight matrix for the TAI-BRGNN 2 model for the inhibitory potency of HPSAAs toward MMPs.^a TAI descriptors with the highest impacts appear in bold letter

Inputs	Hidden nodes		
	3	1	2
<i>TAI5f8p</i>	0.493	-1.232	0.164
<i>TAI5s3e</i>	-0.954	0.739	-0.145
<i>TAI2pK^{1v}</i>	-1.580	-0.537	1.581
<i>TAI2s3v</i>	0.180	0.352	-0.616
<i>TAI2pK^{8e}</i>	2.481	-2.062	0.010
<i>TAI2pK^{2v}</i>	-0.331	-0.376	0.666
<i>TAI1Ht1v</i>	-2.340	1.130	1.405
<i>TAI1Ht5v</i>	3.193	-1.722	-1.290
<i>TAI4s2p</i>	0.424	-0.657	0.729
SCV	0.793	0.147	0.060

^aThe columns are ordered by the SCV for the hidden neurons shown in the last row.

with van der Waals volume and electronegativity in the ligand structure.

The data set of HPSAAs used in our study comes from the previous structure–activity relationship studies in which functional groups at the P_1' , P_1 , and P_2' sites (R , R_1 , and R_2 in Figure 2) of the inhibitors were modified as functional probes for S_1' , S_1 , and S_2' subsites of MMP family members (6–8). Selectivity was searched according to the differences between these pockets in five MMPs (MMP-1, MMP-2, MMP-3, MMP-9, and MMP-13). Amino acid/residue-related properties with the highest impact in the optimum model can be interpreted as the most relevant MMP pocket features ruling the interaction with HPSAAs. In turn, the 2D autocorrelation vectors extracted can be viewed as the relevant features for the inhibitor to match the MMP S_1' , S_1 , and S_2' pockets.

The inhibition of MMPs by HPSAAs lies on 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 that varies in depth for different MMPs (48). MMP-1 has a characteristic Arg in its S_1' subsite (Figure 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 (Figure 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 (Figure 2), which is replaced by Gly-Asn in MMP-1 and by Asn-Val in MMP-3 (49–51). Figure 2 also depicts the distribution of the most relevant amino acid/residue properties (pK' and Ht) along the sequence of MMP family members MMP-1, MMP-2, MMP-3, MMP-9, and MMP-13. As was expected, the properties vary at MMP S_1' , S_1 , and S_2' pockets, accounting for affinity differences among the family members. These differences, which are mainly related to the steric, hydrophobic, and electronic availability of these side chains, reflect on inhibitor binding and enzy-

matic activity. In this regard, it is noteworthy that the most relevant amino acid/residue properties for modeling the MMPs inhibition according to the optimum model *TAI*-BRGNN 2, equilibrium constant with reference to the ionization property of COOH group, and thermodynamic transfer hydrophobicity, specifically account for the variability in electrostatic state and hydrophobicity–hydrophilicity balance among the MMP sequences. However, it should be mentioned that optimum model *TAI*-BRGNN 2 also exhibits contributions from size/shape and enthalpy-related properties.

A broad review on QSAR studies on MMPs has been recently reported by Verma and Hansch (11) describing more than 90 models. The prediction results on the inhibition of various compound series against MMP-1, MMP-2, MMP-3, MMP-7, MMP-8, MMP-9, MMP-12, MMP-13, and MMP-14 revealed that the most important molecular properties were hydrophobicity and molar refractivity, which were the main determinants of the activity. On the contrary, recent 2D autocorrelation linear and non-linear QSAR studies for the inhibition of three MMPs (MMP-1, MMP-9, and MMP-13) by set of 80 HPSAAs were reported by us (13). Different models were developed describing the inhibition of MMP-1, MMP-9, and MMP-13, and the relevance of electronic interactions for MMP-1 inhibition was in accordance with the increase of hydrophilic residues in its active site. Electronic interactions in *S'*1 subsite of MMP-9 and steric interactions in *S'*1 subsite of MMP-1 were found to be crucial requirements for selective MMPs.

Similarly, a varied pool of 2D autocorrelation descriptors was also used by our group in order to search the relevant structural information for individual modeling of the same data set of HPSAAs toward five MMP family members used here (12). Five different models were developed for each MMP by both linear and non-linear modeling techniques. The linear approach identified different relevant inhibitor-related properties in each case. Atomic masses have high contributions to the inhibition of all MMPs, except for MMP-3 for which atomic Sanderson electronegativities are the key features. Inhibition of MMP-1, MMP-3, and MMP-9 was greatly influenced by atomic Sanderson electronegativities. Inhibition of MMP-9 showed to be similarly influenced by atomic masses and Sanderson electronegativities. In general, the linear influence of atomic van der Waals volumes and polarizabilities was poor (12). Nevertheless, the non-linear approach brought more reliable conclusions in accordance with the crossvalidation tests. The high non-linear contribution in the inhibition of all MMPs comes from atomic van der Waals volumes, which was the most relevant feature for the inhibition of MMP-2, MMP-3, and MMP-9. Atomic masses were the key features to the inhibition of MMP-1 whilst the inhibition of MMP-13 was mainly ruled by atomic polarizabilities. On the contrary, atomic Sanderson electronegativities were relevant for the inhibition of MMP-1 and MMP-3. However, it should be mentioned that the optimum non-linear models in (12) exhibited lower LTO crossvalidation accuracies (MMP-1 72%, MMP-2 73%, MMP-3 69%, MMP-9 71%, and MMP-13 73%) in comparison to the test set accuracies per each MMP of the *TAI*-BRGNN 2 ensemble (MMP-1 79%, MMP-2 78%, MMP-3 79%, MMP-9 74%, and MMP-13 88%).

It is interesting to note that, from the ligand's point of view, the most relevant atomic properties according to our optimum model

TAI-BRGNN 2 are atomic Sanderson electronegativities and van der Waals volumes. Despite the target free nature of previous 2D autocorrelation models on MMP inhibition in (12,13), our present results are in concordance with them. From the ligand's point of view, beside some individual contributions of other properties, electronic and hydrophobicity distributions on the molecular structure are the main features governing MMP family inhibition according to the non-linear models.

Self-organizing maps of the MMP/HPSAA complexes

Finally, we aimed to settle some similarity among MMP/HPSAA complexes by building a SOM of the inhibitory potency using the optimum subset of nine-*TAI* descriptor in model *TAI*-BRGNN 2. Figure 5 depicts 14×14 SOM of the pIC_{50} values for the studied MMP/HPSAA complexes. Eighty neurons were occupied of a total of 196 neurons, yielding about 41% of occupancy in the map. As can be observed, target–ligand complexes with similar inhibitory potency range were located at neighboring neurons in the map. Less potent inhibitors were placed at the right top region of the map, whilst most potent inhibitors were allocated at the left top region of the map. Middle potent inhibitors occupied the center, the left bottom, and right bottom regions of the map.

By analyzing the pIC_{50} topological map in Figure 5, instead of some structural similarities among inhibitors, taking into account their allocation at adjacent neurons with similar level of inhibitory potency, it can be observed as a differential allocation of target–ligand complexes for each MMP. This fact suggests that the SOM could help elucidate the affinity preference of an inhibitor to a certain MMP family member according to its allocation when it is projected on the map. In this sense, MMP-1, MMP-2, MMP-3, and MMP-9 target–ligand complex appear very well differentiated in the map. On the contrary, MMP-13 complexes appear mixed with other MMP family members. This result agrees with the fact that

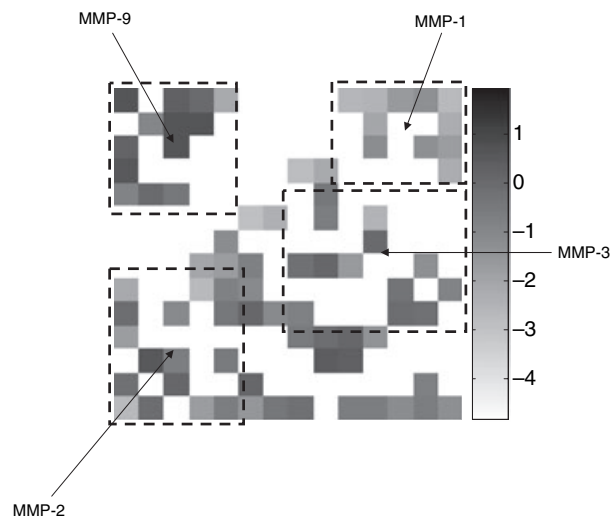


Figure 5: Kohonen self-organizing map of the inhibitory potency (pIC_{50}) of HPSAAs toward MMP family members. Inhibitory potency legend is placed at the right hand of the map.

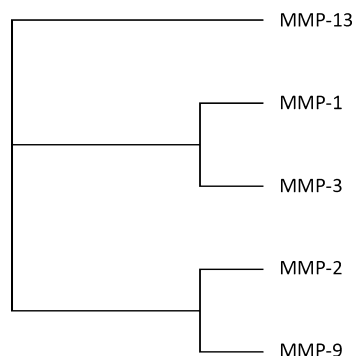


Figure 6: Multiple alignment of MMP family members MMP1, MMP2, MMP3, MMP9 and MMP13 obtained by BioEdit (52) software using ClustalW algorithm (53).

MMP-13 inhibition has higher pair correlations among all the MMP family members according to Table 2. The HPSAAs studied have lower affinity toward MMP-1 but do exhibit high affinity toward MMP-9; the target–ligand complexes of such enzymes are allocated at very well differentiated zones in the map with low and high affinity levels, respectively. However, the studied inhibitors show average inhibitory potency toward the rest of the MMP family members, and their distributions on the map depict scrambled patterns.

The distribution of the inhibition complexes on the SOM depicts a target arrangement rather than an inhibitor arrangement. When comparing the distribution of the five MMPs on the map with the tree view of the multiple alignments of the same proteins in Figure 6 obtained by BIOEDIT (52) software using CLUSTALW algorithm (53), we found that some similarities exist between both grouping approaches. Two clusters appear in Figure 6: one cluster with MMP-2 and MMP-9 and a second cluster with MMP-1 and MMP-3, whilst MMP-13 appears alone in a different leaf of the tree. This MMP distribution pattern is quite similar to the distribution on the SOM of the inhibition complexes; MMP-1 and MMP-5 are allocated at neighboring regions on the left side of the map, while on the right side appear MMP-2 and MMP-9 inhibition complexes. In turn, inhibition complexes of MMP-13 appear disperse on the map. The obtained distribution on the map could also be associated with similarities in the inhibition mechanisms for these MMP family members, not only involving the catalytic domain. In this sense, determination of the crystal structure of the complex of new inhibitor with MMP-13 revealed a novel binding mode characterized by the absence of interaction with the catalytic zinc. A structural alignment of this crystal structure with the catalytic domains of 11 other MMPs revealed some new critical interactions for the selectivity of this class of inhibitors (54). In fact, MMP-13 inhibition could involve mechanisms with and without binding to the catalytic zinc, such a variable target–ligand interaction pattern depicted in the SOM as a scramble distribution.

Conclusions

Classical linear and non-linear QSAR studies are mainly ligand-based, but recently PCMs have been introduced for considering

target structural information in the QSAR models. In this sense, we applied a topological approximation to the PCM modeling of the inhibition of five MMP family members. *TAI* matrix represents the target–ligand *TAI* space of enzyme–inhibitor complexes. Linear and non-linear variable searching strategies achieved nine-descriptor models describing about 80% and 85% of data variance in crossvalidation experiments. In addition, a non-linear ensemble of the optimum BRGNN fitted the test sets with an accuracy of 87%, overcoming the linear ensemble that described only 80% of test set variance. The optimum linear model reflected the influence of six amino acid/residue properties accounting for four main features: hydrophobicity–hydrophilicity balance (properties H_t , R_z), electrostatic state (property pK'), shape/size distribution (properties V^0 , s), and freedom degrees or enthalpy level (f) along the target sequence. Similarly, the optimum non-linear model accounted for such features but having influence only of four amino acids/residues: hydrophobicity–hydrophilicity balance (property H_t), electrostatic state (property pK'), shape/size distribution (property s), and freedom degrees or enthalpy level (property f). On the contrary, from the ligand's point of view, all the three weighting properties – atomic van der Waals volumes, atomic Sanderson electronegativities, and atomic polarizabilities – appeared in both optimum linear and non-linear predictors, but the sensitive analysis of the non-linear model demonstrated that atomic van der Waals volumes and atomic Sanderson electronegativities have the highest impact on the inhibitory potency. Likewise, this analysis showed that, despite some influences of residue size/shape and enthalpy, the most relevant protein-related properties were hydrophobicity–hydrophilicity balance and electrostatic state. This fact suggests that variability in electrostatic state and hydrophobicity–hydrophilicity balance among the MMP S_1' , S_1 , and S_2' pockets rules the affinity of MMP/HPSAA complexes.

Acknowledgments

The authors would like to acknowledge Prof. Akinori Sarai for providing useful information at the time of preparation of the manuscript. Financial support of this research was provided by Cuban Ministerio de Ciencia, Tecnología y Medio Ambiente (CITMA) through a grant to M. Fernandez (Grant No. 20104102).

References

1. Birkedal-Hansen H. (1995) Proteolytic remodeling of the extracellular matrix. *Curr Opin Cell Biol*;7:728–735.
2. Baker A.H., Edwards D.R., Murphy G. (2002) Metalloproteinase inhibitors: biological actions and therapeutic opportunities. *J Cell Sci*;115:3719–3727.
3. Leung D., Abbenante G., Fairlie D.P. (2000) Protease inhibitors: current status and future prospects. *J Med Chem*;43:305–341.
4. Beckett R.P., Davidson A.H., Drummond A.H., Whittaker M. (1996) Recent advances in matrix metalloproteinase inhibition. *Drug Discov Today*;1:16–26.
5. MacPherson L.J., Bayburt E.K., Capparelli M.P., Carroll B.J., Goldstein R., Justice M.R., Zhu L. *et al.* (1997) Discovery of CGS 27023A, a non-peptide, potent, and orally active stromelysin

- inhibitor that blocks cartilage degradation in rabbits. *J Med Chem*;40:2525–2532.
6. Hanessian S., Bouzbouz S., Boudon A., Tucker G.C., Peyroulan D. (1999) Picking the S₁, S₁' and S₂' pockets of matrix metalloproteinases. A niche for potent acyclic sulfonamide inhibitors. *Bioorg Med Chem Lett*;9:1691–1696.
7. Hanessian S., Moitessier N., Gauchet C., Viau M. (2001) N-Aryl sulfonyl homocysteine hydroxamate inhibitors of matrix metalloproteinases: further probing of the S₁, S₁', and S₂' pockets. *J Med Chem*;44:3066–3074.
8. Hanessian S., MacKay D.B., Moitessier N. (2001) Design and synthesis of matrix metalloproteinase inhibitors guided by molecular modeling. Picking the S₁ pocket using conformationally constrained inhibitors. *J Med Chem*;44:3074–3082.
9. Hanessian S., Moitessier N., Therrien E. (2001) A comparative docking study and the design of potentially selective MMP inhibitors. *J Comput Aided Mol Des*;15:873–881.
10. Kumar D., Gupta S.P. (2003) A quantitative structure-activity relationship study on some matrix metalloproteinase and collagenase inhibitors. *Bioorg Med Chem*;11:421–426.
11. Verma R.P., Hansch C. (2007) Matrix metalloproteinases (MMPs): chemical–biological functions and (Q)SARs. *Bioorg Med Chem*;15:2223–2268.
12. Fernández M., Caballero J., Tundidor-Camba A. (2006) Linear and nonlinear QSAR study of N-hydroxy-2-[(phenylsulfonyl)amino] acetamide derivatives as matrix metalloproteinase inhibitors. *Bioorg Med Chem*;14:4137–4150.
13. Fernández M., Caballero J. (2007) QSAR modeling of matrix metalloproteinase inhibition by N-hydroxy- α -phenylsulfonylacetamide derivatives. *Bioorg Med Chem*;15:6298–6310.
14. Wikberg S.J.E., Lapinsch M., Prusis P. (2004) Proteochemometrics: a tool for modeling the molecular interaction space. In: Kubinyi H., Müller G., editors. *Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective*. Weinheim: Wiley-VCH; p. 289–309.
15. Caballero J., Fernández L., Abreu J.I., Fernández M. (2006) Amino acid sequence autocorrelation vectors and ensembles of Bayesian-regularized genetic neural networks for prediction of conformational stability of human lysozyme mutants. *J Chem Inf Model*;46:1255–1268.
16. Caballero J., Fernández L., Gariga M., Abreu J.I., Collina S., Fernández M. (2007) Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines. *J Mol Graph Model*;26:166–178.
17. Fernández L., Caballero J., Abreu J.I., Fernández M. (2007) Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: gene V protein mutants. *Proteins*;67:834–852.
18. Fernández M., Abreu J.I., Caballero J., Gariga M., Fernández L. (2007) Comparative modeling of the conformational stability of chymotrypsin inhibitor 2 protein mutants using amino acid sequence autocorrelation (AASA) and amino acid 3D autocorrelation (AA3DA) vectors and ensembles of Bayesian-regularized genetic neural networks. *Mol Simul*;33:1045–1056.
19. Stewart J.J.P. (1989) Optimization of parameters for semi-empirical methods. *J Comput Chem*;10:210–220.
20. Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Natale D.A., O'Donovan C., Redaschi N., Yeh L.S. (2005) The universal protein resource (UniProt). *Nucleic Acids Res*;33:154–159.
21. Bauknecht H., Zell A., Bayer H., Levi P., Wagener M., Sadowski J., Gasteiger J. (1996) Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J Chem Inf Comput Sci*;36:1205–1213.
22. Moreau G., Broto P. (1980) Autocorrelation of a topological structure: a new molecular descriptor. *Nouv J Chim*;4:359–360.
23. Kawashima S., Kanehisa M. (2000) AAindex: amino acid index database. *Nucleic Acids Res*;28:374.
24. Mackay D.J.C. (1992) Bayesian interpolation. *Neural Comput*;4:415–447.
25. Mackay D.J.C. (1992) A practical Bayesian framework for back-prop networks. *Neural Comput*;4:448–472.
26. Caballero J., Fernández M. (2006) Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regulated neural networks. *J Mol Model*;12:168–181.
27. Caballero J., Tundidor-Camba A., Fernández M. (2007) Modeling of the inhibition constant (K_i) of some cruzain ketone-based inhibitors using 2D spatial autocorrelation vectors and data-diverse ensembles of Bayesian-regularized genetic neural networks. *QSAR Comb Sci*;26:27–40.
28. So S., Karplus M. (1996) Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural network. *J Med Chem*;39:1521–1530.
29. Burden F.R., Winkler D.A. (1999) Robust QSAR models using Bayesian regularized neural networks. *J Med Chem*;42:3183–3187.
30. Winkler D.A., Burden F.R. (2004) Bayesian neural nets for modeling in drug discovery. *Biosilico*;2:104–111.
31. Fernández M., Tundidor-Camba A., Caballero J. (2005) Modeling of cyclin-dependent kinase inhibition by 1H-pyrazolo [3,4-d] pyrimidine derivatives using artificial neural networks ensembles. *J Chem Inf Model*;45:1884–1895.
32. González M.P., Caballero J., Tundidor-Camba A., Helguera A.M., Fernández M. (2006) Modeling of farnesyltransferase inhibition by some thiol and non-thiol peptidomimetic inhibitors using genetic neural networks and RDF approaches. *Bioorg Med Chem*;14:200–213.
33. Fernández M., Caballero J. (2006) Modeling of activity of cyclic urea HIV-1 protease inhibitors using regularized-artificial neural networks. *Bioorg Med Chem*;14:280–294.
34. Fernández M., Caballero J. (2006) Bayesian-regularized genetic neural networks applied to the modeling of non-peptide antagonists for the human luteinizing hormone-releasing hormone receptor. *J Mol Graph Model*;25:410–422.
35. Fernández M., Carreiras M.C., Marco J.L., Caballero J. (2006) Modeling of acetylcholinesterase inhibition by tacrine analogues using Bayesian-regularized genetic neural networks and ensemble averaging. *J Enzyme Inhib Med Chem*;21:647–661.
36. Fernández M., Caballero J. (2007) QSAR models for predicting the activity of non-peptide luteinizing hormone-releasing

- hormone (LHRH) antagonists derived from erythromycin A using quantum chemical properties. *J Mol Model*;13:465–476.
37. Foresee F.D., Hagan M.T. (1997) Gauss-Newton approximation to Bayesian learning. In: *Proceedings of the 1997 International Joint Conference on Neural Networks*, Houston: IEEE, p. 1930.
38. Tetko I., Livingstone D.J., Luik A.I. (1995) Neural network studies. 1. Comparison of overfitting and overtraining. *J Chem Inf Comput Sci*;35:826–833.
39. Hansen L.K., Salamon P. (1990) Neural network ensembles. *IEEE Trans Pattern Anal Mach Intell*;12:993–1001.
40. Baumann K. (2005) Chance correlation in variable subset regression: influence of the objective function, the selection mechanism, and ensemble averaging. *QSAR Comb Sci*;24:1033–1046.
41. Agrafiotis D.K., Cedeño W., Lobanov V.S. (2002) On the use of neural network ensembles in QSAR and QSPR. *J Chem Inf Comput Sci*;42:903–911.
42. Kohonen T. (1982) Self-organized formation of topologically correct feature maps. *Biol Cybern*;43:59–69.
43. Gasteiger J., Zupan J. (1993) Neural networks in chemistry. *Angew Chem Int Ed Engl*;32:503–527.
44. Zupan J., Gasteiger J. (1999) *Neural Networks in Chemistry and Drug Design*. Weinheim: Wiley-VCH.
45. Hawkins D.M. (2004) The problem of overfitting. *J Chem Inf Comput Sci*;44:1–12.
46. Guha R., Stanton D.T., Jurs P.C. (2005) Interpreting computational neural network QSAR models: a detailed interpretation of the weights and biases. *J Chem Inf Model*;45:1109–1121.
47. Stanton D.T. (2003) On the physical interpretation of QSAR models. *J Chem Inf Comput Sci*;43:1423–1433.
48. Welch A.R., Holman C.M., Huber M., Brenner M.C., Browner M.F., Van Wart H.E. (1996) Understanding the P1' specificity of the matrix metalloproteinases: effect of S1' pocket mutations in matrilysin and stromelysin-1. *Biochemistry*;35:10103–10109.
49. Chen L., Rydel T.J., Gu F., Dunaway C.M., Pikul S., Dunham K.M., Barnett B.L. (1999) Crystal structure of the stromelysin catalytic domain at 2.0 Å resolution: inhibitor-induced conformational changes. *J Mol Biol*;293:545–557.
50. 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. (1999) Crystal structures of MMP-1 and -13 reveal the structural basis for selectivity of collagenase inhibitors. *Nat Struct Biol*;3:217–221.
51. 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. (2002) Crystal structure of human MMP9 in complex with a reverse hydroxamate inhibitor. *J Mol Biol*;319:173–181.
52. Hall T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows **95/98/NT**. *Nucleic Acids Symp Ser*;41:95–98.
53. Thompson J.D., Higgins D.G., Gibson T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*;22:4673–4680.
54. Pirard B. (2007) Insight into the structural determinants for selective inhibition of matrix metalloproteinases. *Drug Discov Today*;12:640–646.

Notes

^aMOPAC version 6.0. Frank J. Seiler Research Laboratory, U.S. Air Force Academy, Colorado Springs, CO, 1993.

^bDRAGON Software version 3.0, Milano Chemometrics, 2003.

^cMATLAB version 7.0, The MathWorks, Inc., MA, 2004, <http://www.mathworks.com>.

^dBRGNN toolbox for MATLAB version 1.0, Molecular Modeling Group, University of Matanzas, 2007.

^eThe MathWorks Inc. Genetic algorithm and direct search toolbox user's guide for use with MATLAB, The MathWorks Inc., MA, 2004.

^fThe MathWorks Inc. Neural network toolbox user's guide for use with MATLAB, The MathWorks Inc., MA, 2004.

Supplementary Material

The following supplementary material is available for this article:

Table S1. Experimental, calculated, and residual pIC₅₀ for MMP/HPSAA inhibition complexes according to linear and non-linear ensembles of models TAI-MRA and TAI-BRGNN 2, respectively.

Table S2. Numerical values of the seven selected physicochemical, energetic, and conformational properties of the 20 amino acids/residues (23).

This material is available as part of the online article from:

<http://www.blackwell-synergy.com/doi/abs/10.1111/j.1747-0285.2008.00675.x>

(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.