Bioorganic & Medicinal Chemistry 16 (2008) 7470–7476

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

QSAR prediction of inhibition of aldose reductase for flavonoids

Andrew G. Mercader [a,*], Pablo R. Duchowicz [a], Francisco M. Fernández [a], Eduardo A. Castro [a], Daniel O. Bennardi [b], Juan C. Autino [b], Gustavo P. Romanelli [b,c]

[a] INIFTA (UNLP, CCT La Plata-CONICET), División Química Teórica, Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina
[b] Cátedra de Química Orgánica, Facultad de Ciencias Agrarias y Forestales, UNLP, Calles 60 y 119, B1904AAN La Plata, Argentina
[c] Centro de Investigación y Desarrollo en Ciencias Aplicadas "Dr J.J. Ronco" (CINDECA), Facultad de Ciencias Exactas, Universidad Nacional de La Plata—CONICET, Calle 47 No. 257, B1900AJK La Plata, Argentina

ARTICLE INFO

Article history:
Received 6 March 2008
Revised 2 June 2008
Accepted 4 June 2008
Available online 10 June 2008

Keywords:
QSAR
Flavone derivative
Aldose reductase inhibition
Cataract prevention
Replacement method
Enhanced replacement method
Genetic algorithm
Dragon molecular descriptors

ABSTRACT

We performed a predictive analysis based on quantitative structure–activity relationships (QSAR) of an important property of flavonoids, which is the inhibition (IC_{50}) of aldose reductase (AR). The importance of AR inhibition is that it prevents cataract formation in diabetic patients. The best linear model constructed from 55 molecular structures incorporated six molecular descriptors, selected from more than a thousand geometrical, topological, quantum-mechanical, and electronic types of descriptors. As a practical application, we used the obtained QSAR model to predict the AR inhibitory effect of newly synthesized flavonoids that present 2-, 7-substitutions in the benzopyrane backbone.

1. Introduction

Diabetic patients normally suffer from complications such as cataract, peripheral neuropathy, and vascular disease particularly of the retina, kidney, and heart. Increased activity of sorbitol pathway of glucose metabolism has been implicated in the pathogenesis of these complications.[1] The sorbitol pathway contains two enzymes aldose reductase (AR) (EC 1.1.1.21) and sorbitol dehydrogenase (EC 1.1.1.14).

AR normally reduces glucose to sorbitol using nicotinamide-adeninedinucleotide phosphate (NADPH) as a cofactor; at the same time another sorbitol dehydrogenase oxidizes sorbitol to fructose. However, in diabetis conditions, glucose level in this pathway is increased and sorbitol is produced faster than being oxidized to fructose.[2]

The accumulation of sorbitol in lens, nerve, or retina results in hyperosmotic effect, which leads to lens swelling and subsequent cataract formation as well as the pathologic changes in other tissues.[3] The inhibition of AR is a possible prevention or treatment of these effects.[4]

Several flavonoids and flavonoid derivatives have been reported to have inhibitory activity against AR enzyme.[5–7]

AR being a relevant area of research, selectivity of AR against other reductases is not a minor topic, however, aldehyde reductase inhibition IC_{50} values are not available for the set of flavones used in this work. Nevertheless, scarce data were located in literature, for example, isoaffinetin (5,7,3′,4′,5′-pentahydroxyflavone-6-C-glucoside) is a potent inhibitor of AR (rat lens, porcine lens, and recombinant human) with no inhibition against aldehyde reductase.[8]

Flavonoids (phenyl-benzopyranes) are low molecular weight plant products, that are abundant, relatively simple to synthesize and present several interesting biological activity profiles in enzymatic systems, consequently their study is greatly interesting in many research fields.

Clearly, it is of great interest to be able to predict the IC_{50} of compounds that have no experimental values yet, as well as attempting to determine the structural parameters that the AR inhibition depends on. A generally accepted remedy for overcoming the lack of experimental data in complex chemical phenomena is the analysis based on quantitative structure–activity relationships (QSAR).[9]

A recent QSAR study on a data set of inhibitory activities against AR enzyme of 75 flavonoids was reported using multilinear regression analysis with classical and quantum chemical descriptors.[10] This model lacked statistical significance showing low correlation

* Corresponding author. Tel.: +54 221 425 7430/54 221 425 7291; fax: +54 221 425 4642.
E-mail addresses: amercader@inifta.unlp.edu.ar, andrewmercader@yahoo.com (A.G. Mercader).

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.06.004

coefficients and no predictive ability. A second study used the same data set for multilinear regression analysis selecting the models by a genetic algorithm, followed by artificial networks to further improve the linear models and the predictive power of the correlations in a small extent.[11]

In the present study, we investigate a QSAR model for the inhibition of AR enzyme that could serve as a guide for the rational design of further potent and selective inhibitors having the flavone (Fig. 1) or cromone backbone (Fig. 2); the latter being basically a flavone without the phenyl group in position 2.

A great number of structural molecular descriptors including definitions of all classes were searched using the replacement method (RM)[12–15] and further refined using the recently proposed enhanced replacement method (ERM)[16] for the optimal variable subset selection. We compare our results with those provided by the widely applied genetic algorithm (GA)[17] that provides suitable benchmark data. Our main interest is to apply the new QSAR model to estimate the activity of a group of newly synthesized flavonoids that present 2-, 7-substitutions in the benzopyrane backbone,[18] since they do not have experimentally measured inhibitory effects on AR at the present time. Up to now, few attempts have been carried out to synthesize flavonoids with substitutions of that type. There are few biological characterizations for this sort of newly synthesized molecules, and in this way we expect to provide more knowledge on the above-mentioned phenomena.

2. Methods

2.1. Data set

In the present study, we choose a training set of 56 flavonoid derivatives for which their activities are reported in the literature by Štefanič-Petek et al.[10] We first try to use all 75 molecules from that paper, but a further revision of the references containing the experimental data to clarify some doubts on the structures[6,19] revealed some important errors in the representation of some of them. For this reason the number of flavones was reduced to just 56 reliable structures. More precisely, that reduction was due to the fact that some molecules either did not exhibit the desired flavone backbone or because their exact structure was not found in the references.

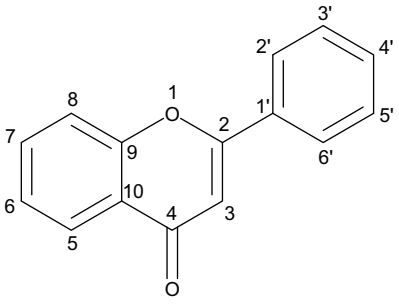

Figure 1. Molecular structure of flavone.

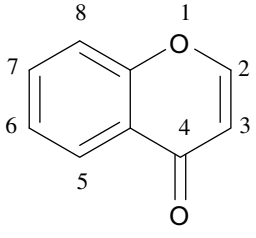

Figure 2. Molecular structure of cromone.

The experimentally inhibitory effects on AR enzyme of the selected molecules were measured spectrophotometrically; the reaction was initiated by addition of the flavonoid derivatives and the rate of NADPH oxidation was determined by monitoring the decrease in absorbance at 340 nm. AR was obtained from the lenses of the eyes of rats of the Wistar strain weighting 200–250 g[6,19] and purified according to the method of Inagaki et al.[20] IC_{50} refers to the micromolar concentration of the compound required for 50% inhibition of the enzyme and was determined by the method of Kador et al.[21]

In addition to it, a set of four flavones with the desired backbone[22] was chosen to test the prediction ability of the new model. In this case all the experimental conditions were almost identical to those of the training set, with the difference that the AR was obtained from the lenses of the eyes of Sprague–Dawley rats weighting 250–280 g.[22] It is expected that such difference will not affect significantly the measurement.

Table 1 summarizes the molecular structures, numbering of the substituents, and experimental $-\log IC_{50}$ of the above-mentioned flavonoid derivatives.

2.2. Molecular descriptors

The structures of the compounds are first pre-optimized with the molecular mechanics force field (MM+) procedure included in the Hyperchem 6.03 package,[23] and the resulting geometries are further refined by means of the semiempirical method PM3 (Parametric Method-3) using the Polak-Ribiere algorithm and a gradient norm limit of 0.01 kcal $Å^{-1}$. We computed the molecular descriptors using the software Dragon 5.0,[24] including parameters of all types such as constitutional, topological, geometrical, charge, GETAWAY (Geometry, Topology, and Atoms-Weighted AssemblY), WHIM (weighted holistic invariant molecular descriptors), 3D-MoRSE (3D-molecular representation of structure based on electron diffraction), molecular walk counts, BCUT descriptors, 2D-autocorrelations, aromaticity indices, Randic molecular profiles, radial distribution functions, functional groups, atom-centered fragments, and empirical properties.[25] We enlarged that pool by the addition of 18 constitutional and 4 quantum-chemical descriptors (molecular dipole moments, total energies, homo–lumo energies) not provided by the program Dragon. The resulting total pool thus consists of $D = 1233$ descriptors.

2.3. Model search

It is our purpose to search the set **D**, containing D descriptors, for an optimal subset **d** of $d \ll D$ ones with minimum standard deviation S

$$S = \frac{1}{(N-d-1)} \sum_{i=1}^{N} \mathrm{res}_i^2 \qquad (1)$$

by means of the multivariable linear regression (MLR) technique. In this equation, N is the number of molecules in the training set, and res_i is the residual for molecule i, the difference between the experimental property (**p**) and predicted property ($\mathbf{p}_{\mathbf{pred}}$). More precisely, we want to obtain the global minimum of $S(\mathbf{d})$, where **d** is a point in a space of $D!/[d!(D-d)!]$ ones. A full search (FS) of optimal variables is impractical because it requires $D!/[d!(D-d)!]$ linear regressions. Some time ago, we proposed the replacement method (RM)[12–15] and more recently the enhanced replacement method (ERM),[16] both approaches produce linear regression QSPR–QSAR models that are quite close to the FS ones with much less

Table 1
Experimental and predicted (Eq. 4) $-\log IC_{50}$

No.	Substituents	$-\log IC_{50}$ Exp.	$-\log IC_{50}$ Pred.
Training set			
1	5,7,3′,4′-OH; 3,6-OCH_3	7.553	7.374
2	3′,4′-OH; 5,6,7,8-OCH_3	7.490	6.922
3	6,3′,4′-OH; 5,7,8-OCH_3	7.456	7.170
4	5,7,3′,4′-OH; 6-OCH_3; 8-CH_2Ph	7.470	7.654
5	5,3′,4′-OH; 6,7,8-OCH_3	7.410	7.014
6	3′,4′-OH; 5,7,8-OCH_3	7.350	6.568
7	5,6,7,3′,4′-OH; 3-OCH_3	7.240	7.518
8	5,6,3′,4′-OH; 7,8-OCH_3	7.190	7.333
9	7,3′,4′-OH; 5,8-OCH_3	7.130	7.078
10	5,3′,4′-OH; 7,8-OCH_3	7.110	7.117
11	3′,4′-OH; 5,6,7-OCH_3	7.040	7.187
12	5,6,7,3′,4′-OH; 8-OCH_3	6.920	6.585
13	6,3′,4′-OH; 5,7-OCH_3	6.850	6.509
14	4′-OH; 5,6,7,8-OCH_3	6.796	6.558
15	8,3′,4′-OH; 5,7-OCH_3	6.790	6.508
16	3′,4′-OH; 3,5,7,8-OCH_3	6.770	6.689
17	5,6,7,3′,4′-OH	6.690	6.598
18	5,3′,4′-OH; 6,7-OCH_3	6.770	6.995
19	5,8,3′,4′-OH; 7-OCH_3	6.640	7.167
20	5,7,3′,4′-OH; 3,8-OCH_3	6.620	6.697
21	6,4′-OH; 5,7,8-OCH_3	6.600	6.631
22	3′,4′-OH; 3,5,6,7-OCH_3	6.570	6.853
23	5,7,3′,4′-OH; 8-OCH_3	6.550	6.667
24	7,3′,4′-OH; 3,5,8-OCH_3	6.550	6.367
25	8-OCH_3; 5,6,7,3′,4′-$OCOCH_3$	6.520	6.336
26	5,6,3′,4′-OH; 7-OCH_3	6.520	6.467
27	6,3′,4′-OH; 3,5,7-OCH_3	6.520	6.830
28	5,3′,4′-OH; 3,6,7-OCH_3	6.458	6.267
29	5,7,4′-OH; 6,8-OCH_3	6.390	6.402
30	5,4′-OH; 6,7,8-OCH_3	6.270	6.394
31	5,6,3′,4′-OH; 3,7-OCH_3	6.090	6.668
32	5,6,4′-OH; 7,8-OCH_3	6.070	6.679
33	5,6,7,4′-OH; 8-OCH_3	5.920	5.782
34	5,6,7,4′-OH; 8,3′-OCH_3	5.920	5.207
35	5,4′-OH; 6,7-OCH_3	5.850	5.475
36	5,7,3′,4′-OH; 3-O-Rh	5.933	5.966
37	5,7,4′-OH; 6,8,3′-OCH_3	5.350	5.276
38	6,4′-OH; 5,7,8,3′-OCH_3	5.200	5.118
39	5,4′-OH; 6,7,3′-OCH_3	5.170	5.284
40	5,7-OH; 6,8,4′-OCH_3	5.140	4.824
41	5,6,7-OH; 8-OCH_3	5.090	4.964
42	5,6-OH; 7,8-OCH_3	5.076	5.155
43	3′,4′-OH; 5,6,7-OCH_3; 3-$COCH_3$	5.050	4.581
44	5,3′-OH; 6,7-OCH_3; 4′-O-Glc	5.086	4.689
45	5-OH; 6,7,3′-OCH_3; 4′-O-Glc	4.880	4.900
46	5-OH; 6,7-OCH_3; 4′-O-Glc	4.790	4.477
47	5,7-OH; 6,8,3′-OCH_3; 4′-O-Glc	4.740	4.521
48	4′-OH; 5,6,7,8,3′-OCH_3	4.730	5.406
49	5,4′-OH; 6,8,3′-OCH_3; 7-O-Glc	4.680	5.185
50	5,7-OH; 6,8,3′,4′-OCH_3	4.530	4.783
51	5,4′-OH; 6,7,8,3′-OCH_3	4.340	5.323
52	5,6,4′-OH; 7,8,3′-OCH_3	3.960	4.748
53	6-OH; 5,7,8-OCH3	3.540	–
54	5,5′-OH; 7,2′,4′-OCH_3	3.500	3.591
55	7-OH; 5-OCH_3	3.000	2.977
56	5,4′-OH; 7,2′,5′-OCH_3	3.000	3.291
Test set			
57	7-OH; 2′-OH	5.780	6.206
58	7-OH; 2′ 4′-OH	5.640	5.254
59	6-OH; 4′-OH	5.280	10.33
60	7-OH; 2′,4′-OH	6.456	5.592

Note: Substituents indication is based on a flavone backbone (Fig. 1).

computational work. These alternative techniques approach the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors $\mathbf{d}=\{X_1,X_2,\ldots,X_d\}$. The RM gives models with better statistical parameters than the forward stepwise regression procedure[26] and similar ones to the more elaborated genetic algorithms,[17] and the ERM leads to even better statistical parameters.[16]

A genetic algorithm is a search technique based on natural evolution, where variables play the role of genes (in this case a set of descriptors) in an individual of the species. An initial group of random individuals (population) evolve according to a fitness function (in this case the standard deviation) that determines the survival of the individuals. The algorithm searches for those individuals that lead to better values of the fitness function through a selection, mutation, and crossover genetic operation. The selection operators guarantee the propagation of individuals with better fitness in future populations. The GAs explore the solution space combining genes from two individuals (parents) using the crossover operator to form two new individuals (children) and also by randomly mutating individuals using the mutation operator. The GAs offer a combination of hill-climbing ability (natural selection) and a stochastic method (crossover and mutation) and explore many solutions in parallel processing information in a very efficient manner. The practical application of GAs requires the tuning of some parameters such as population size, generation gap, crossover rate, and mutation rate. These parameters typically interact among themselves nonlinearly and cannot be optimized one at a time. There is considerable discussion about parameter settings and approaches to parameter adaptation in the evolutionary computation literature; however, there does not seem to be conclusive results on which may be the best.[27]

The Kubinyi function (FIT) is a statistical parameter that closely relates to the Fisher ratio (F), but avoids the main disadvantage of the latter that is too sensitive to changes in small d values, and poorly sensitive to changes in large d values. The FIT($\mathbf{d}$) criterion has a low sensitivity to changes in small d values and a substantially increasing sensitivity for large d values. The greater the FIT value the better the linear equation; it is given by

$$\mathrm{FIT}=\frac{R(d)^2(N-d-1)}{(N+d^2)(1-R(d)^2)} \tag{2}$$

where $R(d)$ is the correlation coefficient for a model with d descriptors. In this paper, we determine the optimal number of molecular descriptors (d_{opt}) in the linear regression equation from the plot of FIT versus d. Assuming that the Kubinyi function exhibits a maximum at d_{max}, we choose d_{opt} in the following way:

(a) if $d_{max} < 7$, then $d_{opt} = d_{max}$.
(b) if $d_{max} > 7$, we define $d_1 = \left[\frac{d_{max}}{2}\right] + 1$, where $[x]$ denotes the integer part of x. Then if the slope of FIT at d_1 is greater than at $d_1 + 1$, then $d_{opt} = d_1$, otherwise, $d_{opt} = d_1 + 1$.

We believe that the value of d_{opt} obtained in this way reflects a 'breaking point' beyond which the FIT improvement can be considered negligible.

We resort to the less time-consuming RM to determine d_{opt}, and finally apply the new ERM[16] to find the best model for d_{opt} descriptors.

As a theoretical validation of all the models, we choose the well-known leave-one-out (loo) and the leave-more-out cross-validation procedures (*l-n%-o*),[28] where *n%* represents the number of molecules removed from the training set. We generated 5,000,000 cases of random data removal for *l-n%-o*, where *n%* = 30% (16 flavonoids).

3. Results and discussion

We first established different predictive relationships to link the molecular structure of flavonoids with their inhibitory activities by means of linear regression models with 1–10 parameters (d) that were searched from the pool of 1233 (D) descriptors. The application of the RM to the training set of 56 flavone derivatives sug-

gested that molecule **53** was an outlier in most resulting models. More precisely, it was the molecule with highest error in 110 out of 143 models tested, and in the best six-parameter model this molecule had an absolute error equal to 2.97S. Surprisingly, the structure of this molecule does not present significant differences with the rest of the molecules belonging to the training set. Since prediction from any QSAR model cannot be intrinsically better than the experimental data employed to develop the model, and the quality of the input data will greatly influence the performance of the QSAR model,[29] molecule **53** was taken out of the training set. The RM and this resulting 55-molecule training set were then used to calculate the best QSAR models with $1 \leqslant d \leqslant 13$. Figure 3 shows that the maximum of the FIT function appears at $d_{max} = 12$, and according to the criterion outlined in Section 2.3 we conclude that the optimal model should have $d_{opt} = 6$ descriptors. More precisely, the best QSAR model provided by the RM is

$$\begin{aligned} -\log IC_{50} = {} & 4.8501(\pm 1.7) + 12.773(\pm 1.4)\text{BELp4} \\ & - 4.950(\pm 0.7)\text{GGI8} - 12.191(\pm 0.9)\text{MATS4e} \\ & + 0.905(\pm 0.2)\text{Mor22e} - 16.422(\pm 2.1)\text{E1p} \\ & - 16.844(\pm 1.7)\text{R4v} \end{aligned} \quad (3)$$

$$N = 55, \quad R = 0.9362, \quad S = 0.4364, \quad \text{FIT} = 3.744, \quad p < 10^{-4}$$
$$R_{loo} = 0.914, \quad S_{loo} = 0.507, \quad R_{l\text{-}30\%\text{-}o} = 0.763, \quad S_{l\text{-}30\%\text{-}o} = 0.891$$
$$\text{RMSE}_{\text{Test Set}} = 3.9059$$

Here, the absolute errors of the regression coefficients are given in parentheses, p is the significance of the model, and $\text{RMSE}_{\text{Test Set}}$ stands for root mean squared errors of the test set. In our calculations we employ the computer system Matlab 5.0.[30]

We then used ERM[16] to search for an improved model with $d_{opt} = 6$ descriptors which in this case is given by

$$\begin{aligned} -\log IC_{50} = {} & -85.5375(\pm 10) - 10.882(\pm 1.4)\text{E1u} \\ & - 15.398(\pm 0.9)\text{MATS4e} + 55.920(\pm 5.35)\text{BELm2} \\ & + 7.7606(\pm 0.9)\text{HATS6e} - 2.6755(\pm 0.5)\text{DISPe} \\ & - 18.253(\pm 1.1)\text{R4p} \end{aligned} \quad (4)$$

$$N = 55, \quad R = 0.9523, \quad S = 0.3789, \quad \text{FIT} = 5.14, \quad p < 10^{-5}$$
$$R_{loo} = 0.934, \quad S_{loo} = 0.447, \quad R_{l\text{-}30\%\text{-}o} = 0.803, \quad S_{l\text{-}30\%\text{-}o} = 0.886$$
$$\text{RMSE}_{\text{Test Set}} = 2.9127$$

As a benchmark, we let a GA to select an optimal model with $d_{opt} = 6$ descriptors. To this end we optimized the GA parameters for this particular problem by means of several trials and thus arrived at the following convenient settings: number of individuals, 250; generation gap, 0.9; single point crossover probability, 0.6; mutation probability, $0.7/d$. We decided to stop the evolution process when one individual occupied more than 90% of the population or when the number of generations reached 2500.

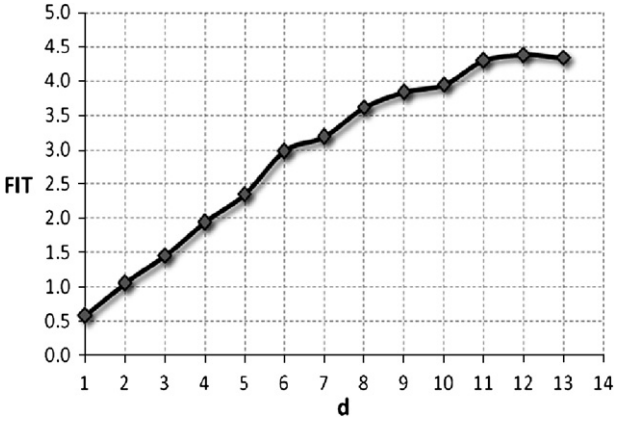

Figure 3. FIT parameter as a function of the number of descriptors for the training set.

From the optimal GA model

$$\begin{aligned} -\log IC_{50} = {} & 12.2967(\pm 1.5) - 0.1898(\pm 0.02)\text{TIC0} \\ & - 15.6392(\pm 1)\text{MATS4e} + 1.7611(\pm 0.4)\text{H7e} \\ & - 10.0425(\pm 1.6)\text{E1u} + 16.6513(\pm 1.9)\text{BELe4} \\ & - 14..0207(\pm 1.5)\text{R3v} \end{aligned} \quad (5)$$

$$N = 55, \quad R = 0.9374, \quad S = 0.4325, \quad \text{FIT} = 3.822, \quad p < 10^{-4}$$
$$R_{loo} = 0.917, \quad S_{loo} = 0.499, \quad R_{l\text{-}30\%\text{-}o} = 0.7739, \quad S_{l\text{-}30\%\text{-}o} = 1.106$$
$$\text{RMSE}_{\text{Test Set}} = 4.1607$$

All the linear models have acceptable predictive quality and present two- and three-dimensional descriptors. Each equation presents different descriptors because their particular combination is optimal to predict the IC_{50} activity. For instance, in Eq. 4 a descriptor of one kind (R4p) may represent the dependence of the polarizability on the IC_{50} and in Eq. 3 this dependence could be represented by a different kind of descriptor (BELp4) that encodes polarizability in combination with a third descriptor (E1p).

By examining the statistical parameters calculated from the training and test sets we conclude that the ERM produces better results than both the GA and RM when exploring large sets of descriptors. Table 2 shows a summary of the linear models with 1 to $d_{opt} + 1$ parameters for RM and d_{opt} parameters for ERM and GA. The details of the molecular descriptors of Table 2 are presented in Table 3. Because of this the rest of the analysis will be performed on Eq. 4.

With the purpose of demonstrating that Eq. 4 does not result from happenstance, we resort to a widely used approach to establish the model robustness: the so-called y-randomization.[31] It consists of scrambling the experimental property **p** in such a way that activities do not correspond to the respective compounds. After analyzing 1,000,000 cases of y-randomization, the smallest value $S = 0.8254$ obtained from this process resulted to be considerably greater than the one corresponding to the true calibration $S = 0.3789$. This result suggests that the model is robust, that the calibration is not a fortuitous correlation, and that we have derived a reliable structure–activity relationship.

The plot of predicted versus experimental $-\log IC_{50}$ shown in Figure 4 suggests that the 55 flavone derivatives follow a straight line. Table 1 shows the predicted inhibitory potencies given by Eq. 4 for the training and test sets. The behavior of the residuals in terms of the predictions illustrated in Figure 5 shows a normal distributions for both sets. This figure omits molecule **59**, which exhibits a residual exceeding $3S = 1.14$. This deviation may be either a statistical defect of our model or a physical consequence of the measurement. Although it is not possible to answer this question by means of present QSAR analysis, it is worth taking into consideration that this molecule presents a residual of the same order and sign in the rest of the models used to determine d_{opt}, which suggests that there may be an error in the corresponding data. We revised the optimization of molecule **59** and the calculation of its descriptors without finding any anomalies. A possible explanation may be the lack of other molecules in the training set with only two hydroxyls as it is the case of molecule **59**. However, molecule **57** in the test set exhibits a similar structure without presenting the same discrepancy. A conclusive investigation on this point requires a revision of the empirical data not available at present.

Table 2
Linear QSAR models for the training set of $-\log IC_{50}$ ($N = 55$)

Model	Descriptors used	*R*	*S*	FIT
M1	LUMO	0.616	0.931	0.579
M2	DISPp, C-027	0.739	0.804	1.059
M3	Mor32m, H-048, >0.2	0.826	0.679	1.715
M4	MATS4e, E1u, HATS6e, R4m	0.878	0.582	2.379
M5	GATS4e, DISPe, E1u, HATS5m, R4m	0.900	0.536	2.607
M6	BELp4, GGI8, MATS4e, Mor22e, E1p, R4v (Eq. 3)	0.936	0.436	3.744
M7	SPP, DISPe, RDF140m, E1p, H4m, Dipole Moment, LUMO	0.950	0.392	4.181
M6B	**E1u**, **MATS4e**, **BELm2**, **HATS6e**, **DISPe**, **R4p** (Eq. 4)	**0.952**	**0.379**	**5.140**
M6C	TIC0, MATS4e, H7e, E1u, BELe4, R3v (Eq. 5)	0.937	0.433	3.822

The best relationship appears in bold.

Table 3
Symbols for molecular descriptors involved in different models

Molecular descriptor	Type	Description
LUMO	Quantum-chemical	Lowest unoccupied molecular orbital energy (eV)
DISPp	Geometrical	d COMMA2 value/weighted by atomic polarizabilities
C-027	Atom-centred fragments	C-027 corresponds to: R–CH–X
Mor32m	3D-MoRSE	3D-MoRSE – signal 32/weighted by atomic masses
H-048	Atom-centred fragments	H attached to C2(sp3)/C1(sp2)/C0(sp)
>0.2	Topological	Number of atoms with charge higher than 0.2
MATS4e	2D Autocorrelations	Moran autocorrelation – lag 4/weighted by atomic Sanderson electronegativities
E1u	WHIM	1st component accessibility directional WHIM index/unweighted
HATS6e	GETAWAY	Leverage-weighted autocorrelation of lag 6/weighted by atomic Sanderson electronegativities
R4m	GETAWAY	*R* Autocorrelation of lag 4/weighted by atomic masses
GATS4e	2D Autocorrelations	Geary autocorrelation – lag 4/weighted by atomic Sanderson electronegativities
DISPe	Geometrical	d COMMA2 value/weighted by atomic Sanderson electronegativities
HATS5m	GETAWAY	leverage-weighted autocorrelation of lag 5/weighted by atomic masses
BELp4	BCUT	Lowest eigenvalue n. 4 of Burden matrix/weighted by atomic polarizabilities
GGI8	Topological	Topological charge index of order 8
Mor22e	3D-MoRSE	3D-MoRSE – signal 22/weighted by atomic Sanderson electronegativities
E1p	WHIM	1st component accessibility directional WHIM index/weighted by atomic polarizabilities
R4v	GETAWAY	*R* Autocorrelation of lag 4/weighted by atomic van der Waals volumes
SPP	Charge	Subpolarity parameter
RDF140m	Radial distribution function	Radial distribution function – 14.0/weighted by atomic masses
H4m	GETAWAY	*H* Autocorrelation of lag 4/weighted by atomic masses
Dipole moment	Quantum-chemical	Total molecular dipole moment (Debyes)
BELm2	BCUT	lowest eigenvalue n. 2 of Burden matrix/weighted by atomic masses
R4p	GETAWAY	*R* Autocorrelation of lag 4/weighted by atomic polarizabilities
TIC0	Topological	total information content index (neighborhood symmetry of 0-order)
H7e	GETAWAY	*H* Autocorrelation of lag 7/weighted by atomic Sanderson electronegativities
BELe4	BCUT	lowest eigenvalue n. 4 of Burden matrix/weighted by atomic Sanderson electronegativities
R3v	GETAWAY	*R* Autocorrelation of lag 3/weighted by atomic van der Waals volumes

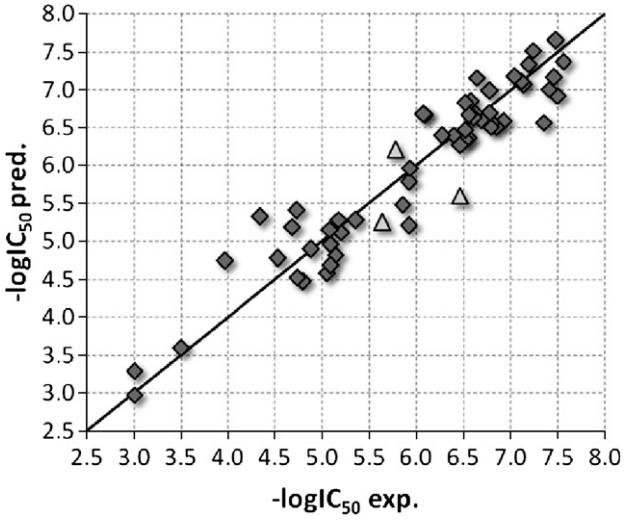

Figure 4. Predicted (Eq. 4) versus experimental $-\log IC_{50}$ for the training (rhombus) and test (triangles) sets.

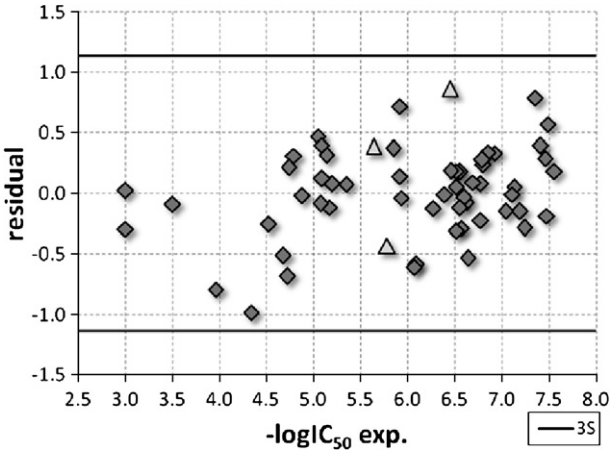

Figure 5. Dispersion plot of the residuals for the training and test sets according to Eq. 4.

Table 4
Correlation matrix for descriptors of Eq. 4 ($N = 55$)

	E1u	MATS4e	BELm2	HATS6e	DISPe	R4p
E1u	1	0.1226	0.2569	0.3177	0.0431	0.1928
MATS4e		1	0.2742	0.3455	0.0274	0.0674
BELm2			1	0.0482	0.3041	0.5992
HATS6e				1	0.0137	0.0578
DISPe					1	0.259
R4p						1

The correlation matrix shown in Table 4 reveals that the descriptors of the linear model are not seriously inter-correlated ($R_{ij} < 0.599$), which justifies the appearance of all the parameters in the equation. The predictive power of the linear model is satisfactory as revealed by its stability upon the inclusion or exclusion of compounds, measured by the statistical parameters $R_{\text{loo}} = 0.934$ and *l-n%-o* $R_{l\text{-}30\%\text{-}o} = 0.803$. According to the literature, $R_{l\text{-}n\%\text{-}o}$ must be greater than 0.71 in order to have a validated model.[32]

As pointed out earlier, we cannot compare our model with previously reported ones because they are based on data with errors in the structure of the molecules.

The molecular descriptors appearing in the linear Eq. 4 merge two- and three-dimensional aspects of the molecular structure, and can be classified as follows: (i) a WHIM descriptor: E1u, 1st component accessibility directional WHIM index/unweighted; (ii) a 2D autocorrelation: MATS4e, Moran autocorrelation – lag 4/weighted by atomic Sanderson electronegativities; (iii) a BCUT descriptor: BELm2, lowest eigenvalue n. 2 of Burden matrix/weighted by atomic masses; (iv) two GETAWAY descriptors: *HATS6e*, leverage-weighted autocorrelation of lag 6/weighted by atomic Sanderson electronegativities, and R4p, R autocorrelation of lag 4/weighted by atomic polarizabilities; and finally, a geometrical descriptor: DISPe, d COMMA2 value/weighted by atomic Sanderson electronegativities.

WHIM (weighted holistic invariant molecular descriptors) descriptors are based on statistical indices calculated on the projections of atoms along principal axes.[33] The aim is to capture 3D information regarding size, shape, symmetry, and atom distributions with respect to invariant reference frames. To calculate them a weighted covariance matrix is obtained from different weighting schemes for the atoms: the unweighted case, atomic mass, van der Waals volume, Sanderson atomic electronegativity, atomic polarizability, and electrotopological state indices. Depending on the weighting scheme different covariances matrices and hence different principal axes are obtained. Essentially the WHIM descriptors provide a variety of principal axes with respect to a defined atomic property. For each weighting scheme, a set of statistical indices are calculated on the atoms projected onto the principal axes (i.e., principal components). Descriptor E1u is a first component accessibility directional WHIM descriptor, which is univariate statistical index calculated on the scores of the individual principal components.

Different structural variables introduced by Broto, Moreau, and Geary[34,35] correspond to bi-dimensional autocorrelations between pairs of atoms in the molecule, and were defined in order to reflect the contribution of a considered atomic property to the experimental observations under investigation ($-\log \text{IC}_{50}$). The atomic properties that can be adopted to differentiate the nature of atoms are the mass (m), polarizability (p), electronegativity (e), or the volume (v). These indices can be readily calculated, that is, by summing products of atomic weights (employing atomic properties such as atomic polarizabilities, and molecular volumes) of the terminal atoms of all the paths of a prescribed length. For the case of MATS4e, the path connecting a pair of atoms has length 4 and involves the atomic Sanderson electronegativities as weighting scheme to distinguish their nature.

BCUT descriptors are the eigenvalues of a modified connectivity matrix, the Burden matrix (**B**).[36,37] The matrix is an **H** depleted molecular graph defined as follows: diagonal elements are atomic numbers of the elements (Z_i); off-diagonal elements (B_{ij}), representing bonded atoms i and j are equal to $\pi^* \times 10^{-1}$, where π^* is the conventional bond order (i.e., 1, 2, 3, 1.5 for single, double, triple, and aromatic bonds, respectively); off-diagonal elements corresponding to terminal bonds are increased by 0.01 and all other matrix elements are set to 0.001. The ordered sequence of the **n** smallest eigenvalues of **B** was proposed as a molecular descriptor based on the assumption that the lowest eigenvalues contain contributions from all the atoms and thus reflects topology of the molecule. The BCUT descriptors are an extension of the Burden eigenvalues and consider three classes of matrices whose diagonal elements correspond to atomic charge related values, atomic polarizability related values, and atomic H bond abilities. A variety of definitions have been used for the off-diagonal terms and both 2D and 3D approaches are considered. The highest and lowest eigenvalues of these matrices have been shown to be discriminating descriptors. BELm2 is the second lowest eigenvalue of **B** involving the atomic masses as weighting scheme.

The GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) type of descriptors[38] were designed with the main purpose of matching the 3D-molecular geometry. These numerical variables are derived from the elements h_{ij} of the molecular influence matrix (**H**), obtained through the values of atomic cartesian coordinates. The diagonal elements of **H** (h_{ii}) are called leverages, and are considered to represent the influence of each molecule atom in determining the whole shape of the molecule. For instance, the mantle atoms always have higher h_{ii} values than atoms near the molecule center, while each off-diagonal element h_{ij} represents the degree of accessibility of the jth atom to interactions with the ith atom. The influence/distance matrix (**R**) involves a combination of the elements of **H** matrix with those of the geometric matrix (**G**). Descriptor R4p involved in Eq. 4 is of the R-GETAWAY type, and represents an **R** index of maximal contribution to the autocorrelation in lag 4 (topological distance) and involves the atomic polarizabilities as weighting scheme to distinguish their nature. Descriptor HATS6e is a 3D-autocorrelation in lag 6 obtained from the Molecular influence Matrix involving the atomic Sanderson electronegativities as weighting scheme.

Geometrical descriptors are different kinds of conformationally dependent descriptors based on the molecular geometry. Comparative molecular moment analysis (CoMMA)[39] utilizes moments of the molecular mass and charge distributions up to and including second order in the development of molecular similarity descriptors. As a consequence, two Cartesian reference frames are then defined with respect to each molecular structure. One frame is the principal inertial axes calculated with respect to the center-of-mass. For neutrally charged molecular species, the other reference frame is the principal quadrupolar axes calculated with respect to the molecular 'center-of-dipole'.

The standardization of the regression coefficients of Eq. 4 allows assigning greater importance to the molecular descriptors that exhibit larger absolute standardized coefficients.[26] In our case we have

$$\begin{aligned}\text{MATS4e}(1.11) > \text{R4p}(0.83) > \text{BELm2}(0.62) > \text{HATS6e}(0.59)\\ > \text{E1u}(0.37) > \text{DISPe}(0.24)\end{aligned} \quad (6)$$

where the standardized coefficients are shown in parentheses. The ranking of contributions given by Eq. 6 suggest that the bi-dimensional autocorrelations *MATS4e* and the GETAWAY descriptor R4p are the most relevant variables for present set of flavonoids. MATS4e indicates that the activity could have a significant depen-

Table 5
Predicted (Eq. 4) $-\log IC_{50}$ of new 2-, 7-substituted benzopyranes

No.	Substituents	$-\log IC_{50}$ Pred.
Estimation set (flavones, Fig. 1)		
61	7-OCH3	7.566
62	7-Cl	6.873
63	7-Br	6.570
Estimation set (cromones, Fig. 2)		
64	2-(2-furyl)	−2.184
65	2-(β-naphtyl)	11.064
66	2-(α-naphtyl)	9.455
67	7-Br, 2-(β-naphtyl)	7.031
68	7-Cl, 2-(α-naphtyl)	5.881
69	7-CH_3, 2-(α-naphtyl)	8.828
70	7-Br, 2-(α-naphtyl)	5.405
71	7-OCH_3, 2-(β-naphtyl)	8.440
72	7-OCH_3, 2-(α-naphtyl)	6.259
73	7-Cl, 2-(β-naphtyl)	7.415
74	7-Cl, 2-(2-furyl)	−4.220
75	7-F, 2-(α-naphtyl)	6.883
76	7-CH_3, 2-(β-naphtyl)	10.415

Note: Substituents indication on structures **61–63** are based on a flavone backbone (Fig. 1) and structures **64–76** on a cromone backbone (Fig. 2).

dence on the electronegativity of the atoms that form the molecule. The most relevant 3D-descriptor R4p is expected to have great dependence on conformational changes, since it encodes information on pairs of atoms considerably far from each other (lag of 4). For this reason, it is possible to argue that the affinity constants for present set of flavone derivatives have great dependence on conformational changes.

By means of the QSAR Eq. 4 we estimated the aldose reductase inhibition activity $-\log IC_{50}$ of our synthesized derivatives, the results are shown in Table 5. Our calculation suggests that those flavonoids with a naphtyl group may exhibit great activity and are, consequently, good candidates for further study. On the other hand, molecules with a furanyl group may probably exhibit low activity and could in principle be rejected as candidates.

4. Conclusions

In this paper, we constructed a predictive QSAR model of inhibitory activity against AR enzyme for 55 flavonoids using six molecular descriptors that take into account 2D- and 3D-aspects of the molecular structure. By means of this QSAR model, we estimated the AR inhibitory activity of some recently synthesized flavonoids displaying 2-, 7-substitutions in the benzopyrane backbone, whose activity has not yet been obtained experimentally. The main result of our investigation is that the presence of a naphtyl group substituting the benzopyrane nucleus greatly increases that activity, while the presence of a furanyl group manifestly decreases it.

Acknowledgment

This work was supported by the National Council of Scientific and Technological Research (CONICET).

References and notes

1. Gabbay, K. H. *N. Engl. J. Med.* **1973**, *288*, 831.
2. Kinoshita, J. H.; Varma, S. D.; Fukui, H. N. *Jpn. J. Ophthalmol.* **1976**, *20*, 399.
3. Kinoshita, J. H. *Invest. Ophthalmol. Vis. Sci.* **1974**, *13*, 713.
4. Varma, S. D.; Kinoshita, J. H. *Biochem. Pharmacol.* **1976**, *25*, 2505.
5. Iwu, M. M.; Igboko, O. A.; Okunji, C. O.; Tempesta, M. S. *Pharm. Pharmacol.* **1990**, *42*, 290.
6. Okuda, J.; Miwa, I.; Inagaki, K.; Horie, T.; Nakayama, M. *Chem. Pharm. Bull.* **1984**, *32*, 767.
7. Varma, S. D. In *Plant Flavonoids in Biology and Medicine: Biochemical, Pharmacological, and Structure–Activity Relationships*, Liss, A. R., Ed., New York, 1986, p 343.
8. Haraguchi, H.; Hayashi, R.; Ishizu, T.; Yagi, A. *Planta Med.* **2003**, *69*, 853.
9. Hansch, C.; Leo, A.. *Exploring QSAR Fundamentals Applications in Chemistry Biology*; American Chemical Society: Washington, DC, 1995.
10. Štefanič-Petek, A.; Krbavčič, A.; Šolmajer, T. *Croat. Chem. Acta* **2002**, *75*, 517.
11. Fernández, M.; Morales, J. C.; Helguera, A.; Castro, E. A.; Pérez Gonzáez, M. *Bioorg. Med. Chem.* **2005**, *13*, 3269.
12. Duchowicz, P. R.; Castro, E. A.; Fernández, F. M.; González, M. P. *Chem. Phys. Lett.* **2005**, *412*, 376.
13. Duchowicz, P. R.; Castro, E. A.; Fernández, F. M. *MATCH Commun. Math. Comput. Chem.* **2006**, *55*, 179.
14. Duchowicz, P. R.; Fernández, M.; Caballero, J.; Castro, E. A.; Fernández, F. M. *Bioorg. Med. Chem.* **2006**, *14*, 5876.
15. Helguera, A. M.; Duchowicz, P. R.; Pérez, M. A. C.; Castro, E. A.; Cordeiro, M. N. D. S.; González, M. P. *Chemometr. Intell. Lab.* **2006**, *81*, 180.
16. Mercader, A. G.; Duchowicz, P. R.; Fernandez, F. M.; Castro, E. A. *Chemometr. Intel. Lab. Syst.* **2008**, *92*, 138.
17. So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.
18. Bennardi, D. O.; Romanelli, G. P.; Jios, J. L.; Vazquez, P. G.; Caceres, C. V.; Autino, J. C. *Heterocycl. Commun.* **2007**, *13*, 79.
19. Okuda, J.; Miwa, I.; Inagaki, K.; Horie, T.; Nakayama, M. *Biochem. Pharmacol.* **1982**, *31*, 3807.
20. Inagaki, K.; Miwa, I.; Okuda, J. *Arch. Biochem. Biophys.* **1982**, *216*, 337.
21. Kador, P. F.; Merola, L. O.; Kinoshita, J. H. *Docum. Ophthalmol. Proc. Ser.* **1979**, *18*, 117.
22. Sung Lim, S.; Hoon Jung, S.; Ji, J.; Shin, K. H.; Keum, S. R. *J. Pharm. Pharmacol.* **2001**, *53*, 653.
23. HYPERCHEM. 6.03 (Hypercube), http://www.hyper.com.
24. DRAGON. 5.0 Evaluation Version, http://www.disat.unimib.it/chm.
25. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley VCH: Weinheim, Germany, 2000.
26. Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons: New York, 1981.
27. Melanie, M. *A Bradford Book*; The MIT Press: Cambridge, Massachusetts London, England, 1998. p 3.
28. Hawkins, D. M.; Basak, S. C.; Mills, D. *J. Chem. Inf. Model.* **2003**, *43*, 579.
29. Liu, H.; Gramatica, P. *Bioorg. Med. Chem.* **2007**, *15*, 5251.
30. Matlab. 5.0 The MathWorks Inc., http://www.mathworks.com/.
31. Wold, S.; Eriksson, L.. In *Chemometrics Methods in Molecular Design*; Waterbeemd, H. v. d., Ed.; Weinheim: VCH, 1995; vol. 20, p 309.
32. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269.
33. Todeschini, R.; Gramatica, P. *Quant. Struct.-Act. Relationships* **1997**, *16*, 113.
34. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 757.
35. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359.
36. Burden, F. R. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225.
37. Frank, R. B. *Quant. Struct.-Act. Relationships* **1997**, *16*, 309.
38. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Model.* **2002**, *42*, 693.
39. Silverman, B. D.; Platt, D. E. *J. Med. Chem.* **1996**, *39*, 2129.