Available online at www.sciencedirect.com

SCIENCE DIRECT®

Bioorganic & Medicinal Chemistry

Bioorganic & Medicinal Chemistry 13 (2005) 3269–3277

Quantitative structure–activity relationship to predict differential inhibition of aldose reductase by flavonoid compounds

Michael Fernández,[a,b] Julio Caballero,[a,b] Aliuska Morales Helguera,[c,d] Eduardo A. Castro[e] and Maykel Pérez González[d,f,*]

[a]*Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740 Matanzas, Cuba*
[b]*Probiotic Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740 Matanzas, Cuba*
[c]*Department of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*
[d]*Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*
[e]*INIFTA, División Química Teórica, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Diag. 113 y 64, Suc.4, C.C. 16, La Plata 1900, Argentina*
[f]*Unit of Service, Drug Design Department, Experimental Sugar Cane Station 'Villa Clara-Cienfuegos', Ranchuelo, 53100 Villa Clara, Cuba*

Received 14 December 2004; revised 11 February 2005; accepted 14 February 2005
Available online 19 March 2005

Abstract—Inhibitory activity against aldose reductase enzyme of flavonoid derivatives were modelled using 11 kinds of molecular descriptors from Dragon software. Model with four Galvez Charge Indices described 67% of data variance and overtaken other models using the same number of variables. Galvez indices showed to contain important information on the relationship between the inhibitor structures and its activity by describing the molecular topology and charge transfer through the molecule. In addition, artificial neural networks were trained using charge indices from the linear models but the obtaining networks overfitted the data having low predictive power.

1. Introduction

Several diseases are caused by the inappropriate expression and malfunction of different enzymatic systems. Such is the case of cataract complications in diabetes patients produced by the over-expression of aldose reductase (AR) enzyme (EC 1.1.1.21).[1] This enzyme normally reduces glucose to sorbitol using nicotinamide-adeninedinucleotide phosphate (NADPH) as a cofactor, at same time another enzyme, sorbitol dehydrogenase, oxidizes sorbitol to fructose. However, in diabetes conditions, glucose level in this pathway is increased and sorbitol is produced faster than oxidized to fructose. The accumulation of sorbitol in lens, nerve and retina provokes a hyperosmotic effect causing lens swelling and opacities that ultimate leads to cataract formation.[1] In this sense, a possible prevention or treatment of cataract is the inhibition of AR enzyme.

Flavonoids, phenylbenzo-pyrones (phenylchromones) based on a common three-ring nucleus, are a group of about 4000 naturally occurring compounds that are ubiquitous in all vascular plants and important constituents of the human diet. These low molecular weight substances have important effects in plant biochemistry and physiology, acting as antioxidants, enzyme inhibitors, precursors of toxic substances and pigments and light screens.[2] These compounds have long been recognized to possess anti-inflammatory, antioxidant, antiallergic, hepatoprotective, antithrombotic, antiviral and anticarcinogenic activities.[3–5] Several key bioflavonoids and flavonoids derivatives have been also found with inhibitory activity against AR enzyme.[6–8]

In a study of 30 flavones, 4 isoflavones and 13 coumarins, many potent inhibitors were found, but 5,7,3′,4′-tetrahydroxy-3,6-dimethoxyflavone and 6,3′,4′-trihy-

Keywords: QSAR; Aldose reductase; Inhibitory activity; Charge indices; Artificial neural networks.

* Corresponding author. Tel.: +53 4228 1473; fax: +53 4228 1130; e-mail: mpgonzalez76@yahoo.es

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.02.038

droxy-5,7,8-trimethoxyflavone were especially active.[7] In a subsequent study of 3′,4′-dihydroxyflavones, another potent inhibitor was discovered: 3′,4′-dihydroxy-5,6,7,8-tetramethoxyflavone.[8] Varma and Kinoshita reported the role in the AR inhibitory activity of the *ortho* orientation of the hydroxyl groups in positions 3′ and 4′ of ring B of flavonoids.[9] Other studies have been focused on the effect of number of hydroxyls group in the flavonoid structure and glycosylation on the AR inhibitory activity.[10] Molecular modelling, orbital calculations and kinetics studies have settled AR inhibition site contains a nucleophilic residue that reversible interacts with AR inhibitors in a non-competitive way.[11–15]

Recently a quantitative structure–activity relationship (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.[16] However the obtained model lack of statistical significance showing low correlation coefficients and no predictive ability for the models were reported by the authors.

In this paper we reported QSAR studies for predicting AR enzyme inhibitory activity using the above mentioned set of 75 flavonoid derivatives.[16] Whole descriptor sets of Dragon[17] computer software were used for performing multilinear regression analysis and models with 3, 4 and 5 variables selected by Genetic algorithm procedure were obtained for each group of descriptors. The ability of each kind of descriptors to describe the data was compared taking into account the statistic significance and the predictive power of the models. The best overall-performing linear models contained Galvez descriptors,[18,19] these variable subsets were used for training Artificial Neural Networks in order to obtain improved non-linear models.

2. Materials and methods

2.1. Data set

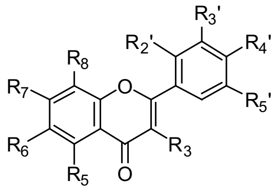
In the present study a data set of 75 flavonoid derivatives for which their activities are reported in the literature by Štefanič-Petek et al.[16] was used. In such a report, inhibitory effects on AR enzyme were measured spectrophotometrically, reaction was initiated by addition of the flavonoids derivatives and the rate of NADPH oxidation was determined by following the decrease in absorbance at 390 nm.[8] Molecular structure, numbering of the substituents and activities of the flavonoid derivatives (natural and synthetic), are summarized in Table 1.

IC_{50} refers to the micromolar concentration of the compound required for 50% inhibition of the enzyme activity.

2.2. Molecular descriptors

In this way we carry out geometry optimization calculations for each compound of this study using the quan-

Table 1. Chemical structures and AR observed and predicted inhibitory activities of flavonoids used in the current study



Compound	Substituent[a]	$\log 1/IC_{50}$	
		Observed	Predicted
1	5,7,3′,4′-OH; 3,6-OCH_3	7.52	5.98
2	3′,4′-OH; 5,6,7,8-OCH_3	7.49	6.80
3	6,3′,4′-OH; 5,7,8-OCH_3	7.47	6.59
4	5,7,3′,4′-OH; 6-OCH_3; 8-CH_2Ph	7.47	6.19
5	5,3′,4′-OH; 6,7,8-OCH_3	7.41	6.45
6	3′,4′-OH; 5,7,8-OCH_3	7.35	6.62
7	5,6,7,3′,4′-OH; 3-OCH_3	7.24	7.24
8	5,6,3′,4′-OH; 7,8-OCH_3	7.19	7.00
9	7,3′,4′-OH; 5,8-OCH_3	7.13	6.50
10	5,3′,4′-OH; 7,8-OCH_3	7.11	6.43
11	3′,4′-OH; 5,6,7-OCH_3	7.04	6.27
12	5,6,7,3′,4′-OH; 8-OCH_3	6.92	7.26
13	6,3′,4′-OH; 5,7-OCH_3	6.85	6.13
14	4′-OH; 5,6,7,8-OCH_3	6.79	6.12
15	8,3′,4′-OH; 5,7-OCH_3	6.79	6.43
16	3′,4′-OH; 3,5,7,8-OCH_3	6.77	7.06
17	5,6,7,3′,4′-OH	6.69	6.97
18	5,3′,4′-OH; 6,7-OCH_3	6.66	6.06
19	5,8,3′,4′-OH; 7-OCH_3	6.64	6.38
20	5,7,3′,4′-OH; 3,8-OCH_3	6.62	6.51
21	6,4′-OH; 5,7,8-OCH_3	6.6	5.97
22	3′,4′-OH; 5,6,7-OCH_3	6.57	6.27
23	5,7,3′,4′-OH; 8-OCH_3	6.55	6.10
24	7,3′,4′-OH; 3,5,8-OCH_3	6.55	6.99
25	8-OCH_3; 5,6,7,3′, 4′-$OCOCH_3$	6.52	6.36
26	5,6,3′,4′-OH; 7-OCH_3	6.52	6.55
27	6,3′,4′-OH; 3,5,7-OCH_3	6.52	6.35
28	5,3′,4′-OH; 3,6,7-OCH_3	6.46	6.21
29	5,7,4′-OH; 6,8-OCH_3	6.39	5.28
30	5,4′-OH; 6,7,8-OCH_3	6.27	5.83
31	5,6,3′,4′-OH; 3,7-OCH_3	6.09	6.82
32	3,5,7,3′,4′-OH	6.09	5.84
33	5,6,4′-OH; 7,8-OCH_3	6.07	6.32
34	5,6,7,4′-OH; 8-OCH_3	5.92	6.58
35	5,6,7,4′-OH; 8,3′-OCH_3	5.92	6.49
36	5,4′-OH; 6,7-OCH_3	5.85	5.19
37	5,7,3′,4′-OH; 3-O-Rh	5.69	5.97
38	2-COOH; 7-OH	5.66	5.85
39	2-$COOCH_2CH_3$; 7-OH	5.64	5.54
40	2-$COOCH_2Ph$; 7-OH	5.6	5.50
41	2-$COOCH(CH_3)_2$; 7-OH	5.51	5.61
42	5,7,4′-OH; 6,8,3′-OCH_3	5.35	5.24
43	6,4′-OH; 5,7,8,3′-OCH_3	5.2	5.83
44	5,4′-OH; 6,7,3′-OCH_3	5.17	5.29
45	5,7-OH; 6,8,4′-OCH_3	5.14	5.10
46	5,6,7-OH; 8-OCH_3	5.09	5.33
47	5,6-OH; 7,8-OCH_3	5.08	5.18
48	3′,4′-OH; 5,6,7-OCH_3; 3-$COCH_3$	5.05	5.90
49	5,3′-OH; 6,7-OCH_3; 4′-O-Glc	5.02	4.93
50	4-OH	4.92	4.31

Table 1 (*continued*)

Compound	Substituent[a]	$\log 1/IC_{50}$	
		Observed	Predicted
51	5-OH; 6,7,3′-OCH_3; 4′-O-Glc	4.88	4.88
52	5-OH; 6,7-OCH_3; 4′-O-Glc	4.79	3.98
53	5,7,3′,4′-OH; 3-O-Glc	4.78	5.41
54	5,7-OH; 6,8,3′-OCH_3; 4′-O-Glc	4.74	4.74
55	4′-OH; 5,6,7,8,3′-OCH_3	4.73	4.96
56	5,4′-OH; 6,8,3′-OCH_3; 7-O-Glc	4.68	5.63
57	4-OH; 7-OCH_3; 3-Ph	4.67	3.91
58	5,7-OH; 6,8,3′, 4′-OCH_3	4.53	5.17
59	3-Ph; 4-OH	4.48	3.43
60	3-OH; 6-OCH_3	4.48	5.98
61	3-CN	4.48	4.39
62	3-COOH	4.42	3.94
63	5,4′-OH; 6,7,8,3′-OCH_3	4.34	5.69
64	3-OH	4.34	4.89
65	3,8-COOH; 5-OCH_3	4.25	4.19
66	4-OH; 3,7-OCH_3	4.15	4.93
67	7-OCH_3; 4-CH_3	4.15	5.02
68	3,5,7,4′-OH; 3′-OCH_3	4.00	5.14
69	4-CH_3	4.00	4.31
70	3-CH_3; 4-OH	4.00	4.89
71	5,6,4′-OH; 7,8,3′-OCH_3	3.96	6.23
72	6-OH; 5,7,8-OCH_3	3.54	4.83
73	5,5′-OH; 7,2′,4′-OCH_3	3.5	3.38
74	7-OH; 5-OCH_3	3.00	2.85
75	5,4′-OH; 7,2′,5′-OCH_3	3.00	3.52

[a] Rh = rhamnose, Glc = glucose.

tum chemical semi-empirical method AM1[20] included in Mopac 6.0 computer software.[21]

Dragon[17] computer software was employed to calculate the molecular descriptors. Dragon descriptors include different groups: Constitutional descriptors, Topological indices, Molecular walks counts,[22,23] BCUT descriptors,[24] Galvez topological charge indices,[18,19] 2D autocorrelations, Randić molecular profiles,[25] geometrical descriptors, RDF descriptors,[26,27] 3D-MoRSE descriptors,[28] Weighted Holistic Invariant Molecular (WHIM) descriptors[29,30] and GETAWAY descriptors.[31]

The total number of descriptors was 1108. Descriptors with constant values inside each group of descriptors were discarded. For the remaining descriptors pairwise correlation analysis (see below) for all kinds of descriptors was performed. The following descriptors exclusion methods were used to reduce, in a first step, the collinearity and correlation between descriptors.

2.3. Pairwise correlation analysis

The procedure consists of elimination of one of the descriptors from each pair with the modulus of the correlation coefficients higher than a predefined value $R_{\max}$ (0.90). The procedure must be carried out with care. Indeed, let $R_{ij} = R(d_i, d_j)$ be the correlation coefficient between descriptors d_i and d_j. Then from $R_{ij} > R_{\max}$ and $R_{jk} > R_{\max}$ does not follow that $R_{ik} > R_{\max}$. So in this case, if d_j is eliminated, d_k must be retained.

In this work, we have used the following algorithm of the pairwise correlation analysis:

1. Sort descriptors by variance and exclude all descriptors with the variance lower than the predefined value. Let D be the descriptor with the highest variance.
2. Calculate correlation coefficient with between D and all other descriptors.
3. Exclude descriptor having the modulus of the correlation coefficient with D higher than $R_{\max}$.
4. Let D be the next descriptor with the highest variance. Go to step (2). If there are no descriptors left, stop.

2.4. Statistical methods

2.4.1. Genetic algorithm (GA) analysis. Eleven models were developed using the descriptors obtained by Dragon[17] computer software. All statistical analysis and data exploration was carried out using the Statistic 6.0.[32] The most significance parameters were identified from the data set using GA analysis.

GA is a class of methods based on biological evolution rules. The first step is to create a population of linear regression models. These regression models mate with each other, mutate, cross-over, reproduce, and then evolve through successive generations towards an optimum solution. The GA simulation conditions were 10,000 generations and 300 populations. The models were linear combinations of 3, 4 and 5 descriptors. The GA procedure was repeated *n*-times to confirm that the selected descriptors are the most optimal descriptor set for describing the modelled property.

Examining the regression coefficients, the standard deviations, the significances and the number of variables in the equation determined the quality of the models.

2.4.2. Artificial neural networks (ANNs) approach. The ANNs is a computer-based model in which a number of processing elements, also called neurons, units or nodes are interconnected by links in a netlike structure forming 'layers'.[33,34] A variable value is assigned to every neuron. The neurons can be one of three different kinds:

1. Input neurons, which receive their values by direct assignation and are associated with independent variables, with the exception of the bias neuron. They form the input layer.
2. Hidden neurons, which collect values from other neurons, giving a result that is passed to a non-input neuron.
3. Output neurons, which collect values from other units. They correspond to different dependent variables, forming the output layer.

The links between units have associated values, named weights that condition the values assigned to the neurons. There exist additional weights assigned to bias values that act as neuron value offsets. The weights are adjusted through a training process in order to minimize network error. Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output.

The characteristics of the ANNs have been found to be suitable for data processing, in which the functional relationship between the input and the output is not previously defined. Structure–activity relationships are often non-linear and very complex and neural networks are able to approximate any kind of analytical continuous function, according to Kolmogorov's theorem.[35]

ANN approach was used for obtaining non-linear models for the AR inhibitory activities of the studied flavonoids derivatives, expecting to improve linear models performance. ANN models and ANN-related statistics calculations were implemented in Matlab 6.5[36] computer software. ANNs used here had variable architectures: number of neurons in the input layer was equal to the number of variables, the number of hidden neurons was varied from 2 to 6 and one neuron was placed in the output layers. Three, four and five molecular descriptors, giving the best linear models, were used as networks inputs and the AR inhibitory activities as target outputs. Network errors were back-propagated and minimized through the networks using a conjugate gradient algorithm with Polak–Ribiere updates. Each network was trained for 5000 epochs or until the network mean square error was lower than 0.02.

2.5. Validation of the models

Linear and non-linear models obtained were validated by calculating q^2 values. The q^2 values are calculated from 'leave-one-out' (LOO) and leave-group-out (LGO) testiness, also known as cross-validation. A data point is removed from the set, and the regression recalculated; the predicted value for that point is then compared to its actual value. This is repeated until each datum has been omitted once; the sum of squares of these deletion residuals can then be used to calculate q^2, an equivalent statistic to R^2. The q^2 values can be considered a measure of the predictive power of a regression equation: whereas R^2 can always be increased artificially by adding more parameters (descriptors), q^2 decreases if a model is overparameterized,[36] and is therefore a more meaningful summary statistic for QSAR models.

3. Results and discussion

As we previously pointed out, one of the objectives of the current work is to compare several kinds of molecular descriptors and methods for describing the property under study. Consequently, we have developed 11 models using the same data set that was included in a previous QSAR study.[16] The symbols, definition and results obtained with the use of Constitutional, Topological, BCUT, Galvez topological charge indices, 2D autocorrelations, Randić molecular profiles, Geometrical, RDF, 3D-MoRSE, WHIM and GETAWAY descriptors are given in Tables 2 and 3.

In this work, the model selection was subjected to the principle of parsimony.[37] Then, we choose a function with high statistical signification but having so few descriptors as possible.

The best QSAR model obtained is given below together with the statistical parameters of the regression.

$$\begin{aligned} -\log(\mathrm{IC}_{50}) = {} & 0.68 + 2.39 \cdot \mathrm{GGI2} - 4.34 \cdot \mathrm{GGI3} \\ & + 51.93 \cdot \mathrm{JGI4} + 69.57 \cdot \mathrm{JGI6} \end{aligned} \qquad (1)$$

$$N = 75 \quad S = 0.712 \quad R = 0.821 \quad F = 36.219$$
$$p < 10^{-5} \quad q^2 = 0.635 \quad S_{\mathrm{cv}} = 0.754$$

where $-\log(\mathrm{IC}_{50})$ is the studied property, N is the number of compounds included in the model, R is the correlation coefficient, S is the standard deviation of the regression, F is the Fisher ratio, q^2 is the correlation coefficient of the cross-validation, p is the significance of the variables in the model and S_{cv} is the standard deviation of the cross-validation. Table 4 shows the values of molecular descriptors used for obtaining the previous equation.

3.1. Interpretation and comparison with other approach and methods

In order to compare the models statistic parameters, we carried out the models using for all of them the same number of variables. As can be seen in Table 3 there are remarkable differences concerning the explanation of the experimental variance given by these models compared to the Charges Indices ones. Firstly, five variables models were unable to explain more than 69% of the data variance and since they did not bring any further significance improvement in comparison with four variable models can be discarded. While the four variables Galvez Charges Indices QSAR model (model 1) explains more than 67% of activities the rest of the models are unable to explain more than 60% of such variance. Moreover, important statistic parameters such as the Fischer ratio (F) and the standard deviation (S) are of higher quality in the case of Galvez Charges Indices model.

Galvez Charge Indices GGIk and JGIk are defined as

$$\mathrm{GGIk} = \sum_{i=1,j=i+1}^{i=N-1,j=N} \left|\mathrm{CT}_{ij}\right| \delta(k, D_{ij}) \qquad (2)$$

$$\mathrm{JGIk} = \frac{\mathrm{GGIk}}{(N-1)} \qquad (3)$$

where N is the number of vertices (atoms different to hydrogen) in the molecular graph. $\mathrm{CT}_{ij} = m_{ij} - m_{ji}$. '$m$' stands for the elements of the M Matrix, $M = A \times D^*$,

Table 2. Symbols of the descriptors used in the models and their definitions

Symbols	Descriptor definition
RBF	Rotatable bond fraction
nAB	Number of aromatic bonds
nH	Number of hydrogen bonds
Ss	Sum of Kier–Hall electrotopological states
nBO	Number of non-hydrogen bonds
RBN	Number of rotatable bonds
nBnz	Number of benzene-like rings
PW2	Path/walk 2–Randić shape index
IDDE	Mean information content of the distance equality
IC4	Information content index (neighbourhood symmetry of 4-order)
X4A	Average connectivity index chi-4
VEA1	Eigenvector coefficient sum from adjacency matrix
BELm1	Lowest eigenvalue n.1 of burden matrix/weighted by atomic masses
BELv1	Lowest eigenvalue n.1 of burden matrix/weighted by atomic van der Waals volumes
BELv6	Lowest eigenvalue n.6 of burden matrix/weighted by atomic van der Waals volumes
BELm5	Lowest eigenvalue n.5 of burden matrix/weighted by atomic masses
GGI2	Topological charges index of order 2
GGI3	Topological charges index of order 3
GGI4	Topological charges index of order 4
JGI2	Mean topological charge index of order 2
JGI4	Mean topological charge index of order 4
JGI6	Mean topological charge index of order 6
MATS6m	Moran autocorrelation—lag 6/weighted by atomic masses
GATS1m	Geary autocorrelations—lag 1/weighted by atomic masses
GATS2m	Geary autocorrelations—lag 2/weighted by atomic masses
MATS3e	Moran autocorrelations—lag 3/weighted by atomic Sanderson electronegativities
DP03	Molecular profile no 03
DP11	Molecular profile no 11
DP18	Molecular profile no 18
SP10	Shape profile no 10
SP11	Shape profile no 11
SP15	Shape profile no 15
SP16	Shape profile no 16
AGDD	Average geometric distance degree
G1	Gravitational index G1
G2	Gravitational index G2 (bond restricted)
SPH	Spherosity
SPAN	Span R
MEcc	Molecular eccentricity
RDF095u	Radial distribution function—9.5/unweighted
RDF045m	Radial distribution function—4.5/weighted by atomic masses
RDF045v	Radial distribution function—4.5/weighted by atomic van der Waals volumes
RDF025e	Radial distribution function—2.5/weighted by Sanderson electronegativities
RDF095e	Radial distribution function—9.5/weighted by Sanderson electronegativities
RDF010p	Radial distribution function—1.0/weighted by atomic polarizabilities
Mor20u	3D-MoRSE—signal 20/unweighted
Mor07m	3D-MoRSE—signal 07/weighted by atomic masses
Mor24p	3D-MoRSE—signal 24/weighted by atomic polarizabilities
Mor20v	3D-MoRSE—signal 20/weighted by atomic van der Waals volumes
Mor24v	3D-MoRSE—signal 24/weighted by atomic van der Waals volumes
Mor07e	3D-MoRSE—signal 07/weighted by Sanderson electronegativities
Mor19e	3D-MoRSE—signal 19/weighted by Sanderson electronegativities
P1u	First component shape directional WHIM index/unweighted
P1m	First component shape directional WHIM index/weighted by atomic masses
Dp	*D* total accessibility index/weighted by atomic polarizabilities
L2s	Second component size directional WHIM index/weighted by atomic electrotopological states
Kv	*K* global shape index/weighted by atomic van der Waals volumes
HATS3m	Leverage-weighted autocorrelation of lag 3/weighted by atomic masses
HATS4m	Leverage-weighted autocorrelation of lag 4/weighted by atomic masses
R3u	*R* autocorrelation of lag 3/unweighted
H5v	*H* autocorrelation of lag 5/weighted by atomic van der Waals volumes
RCON	Randić-type R matrix connectivity

A is the adjacency ($N \times N$) matrix of the molecular graph, D^* is the inverse square distance matrix, in which their diagonal entries are assigned as 0, and δ is Kronecker's delta. Thus, GGIk represents the sum of all

Table 3. Statistical parameters of the linear regression models obtained with the 11 kinds of descriptors used in this study

Model	Variables	R	S	F	$p<$	q^2	S_{cv}
Constitutional	RBF, nAB, nH	0.607	0.984	13.774	10^{-5}	0.328	1.015
Constitutional	Ss, nBO, RBN, nBnz	0.631	0.968	11.554	10^{-5}	0.343	1.010
Topological	PW2, IDDE, IC4	0.633	0.959	15.800	10^{-5}	0.351	0.997
Topological	X4A, PW2, IDDE, VEA1	0.685	0.908	15.497	10^{-5}	0.401	0.965
BCUT	BELm1, BELv1, BELv6	0.617	0.974	14.536	10^{-5}	0.309	1.029
BCUT	BELm1, BELm5, BELv1, BELv6	0.637	0.961	11.967	10^{-5}	0.327	1.023
Galvez Charges Indices	GGI3, GGI4, JGI2	0.783	0.770	37.485	10^{-5}	0.579	0.803
Galvez Charges Indices	GGI2, GGI3, JGI4, JGI6	**0.821**	**0.712**	**36.219**	10^{-5}	**0.635**	**0.754**
2D autocorrelations	MATS6m, GATS1m, GATS2m	0.667	0.923	18.945	10^{-5}	0.391	0.966
2D autocorrelations	MATS6m, MATS3e, GATS1m, GATS2m	0.712	0.876	17.938	10^{-5}	0.439	0.934
Randic molecular profiles	DP03, SP10, SP11	0.665	0.924	18.789	10^{-5}	0.389	0.967
Randic molecular profiles	DP11, DP18, SP15, SP16	0.701	0.889	16.906	10^{-5}	0.430	0.942
Geometrical	AGDD, G2, SPH	0.614	0.977	14.345	10^{-5}	0.330	1.013
Geometrical	AGDD, G1, SPAN, MEcc	0.682	0.912	15.190	10^{-5}	0.402	0.964
RDF	RDF045m, RDF045v, RDF010p	0.685	0.902	20.904	10^{-5}	0.423	0.941
RDF	RDF095u, RDF025e, RDF095e, RDF010p	0.737	0.843	20.794	10^{-5}	0.480	0.899
3D-Morse	Mor20u, Mor07m, Mor24p	0.704	0.880	23.197	10^{-5}	0.448	0.920
3D-Morse	Mor20v, Mor24v, Mor07e, Mor19e	0.752	0.822	22.743	10^{-5}	0.500	0.882
WHIM	P1u, P1e, Dp	0.74	0.832	28.690	10^{-5}	0.505	0.871
WHIM	P1m, L2s, Kv, Dp	0.756	0.816	23.386	10^{-5}	0.518	0.865
GETAWAY	HATS3m, HATS4m, R3u	0.741	0.832	28.789	10^{-5}	0.495	0.880
GETAWAY	HATS3m, HATS4m, H5v, RCON	0.774	0.759	29.786	10^{-5}	0.545	0.841

the CT_{ij} terms, with $D_{ij} = k$, being D_{ij} the entries of the topological distance matrix (D). In the Charge Indices terms, the presence of heteroatoms is taken into account by introducing their electronegativity values (according to Pauling's scale taking chlorine as standard value = 2) in the corresponding entry of the main diagonal of the adjacency matrix. These indexes represent a strictly topological quantity plausably correlating with the charge distribution inside the molecule. This distribution is an important property, which conditions the behaviour of many physicochemical and biological properties.[18,19]

The Galvez Charge Indices model not only overtakes the other ten models in the statistical parameters of the regression but more importantly in the stability to the inclusion–exclusion of compounds as measured by the correlation coefficient and standard deviation of the cross-validation. As can be seen in Table 3 the value of the determination coefficient of leave-one-out cross-validation for the model obtained with the Galvez Charge Indices ($q^2 = 0.635$) was the highest for all analyses model proving the predict power of this approach and the stability of the model.

These results have shown that the Galvez Charge Indices descriptors not only explain the experimental data, but seem to be the best one in doing so.

On the other hand, several reports have been published about QSAR of the inhibition of aldose reductase involving flavonoids. In the Štefanič-Petek et al. report[16] the following equation was used in order to describe the inhibition of AR enzyme by the set of compounds used in the current study.

$$A = 14.05 - 4.21 \cdot \delta_{4'} + 3.94 \cdot \sum \delta_{2'-4'} - 6.92 \cdot \sum \delta_{a_{3'}} - 1.74 \cdot \sum \delta_{a_3}$$
$$N = 75 \quad R = 0.790 \quad F = 21.687 \qquad (4)$$

where A is $-\log(\mathrm{IC}_{50})$, N is the number of compounds included in the model, R is the correlation coefficient and F is the Fisher ratio. The variables in the equation correspond to quantum chemical descriptors; δ_n is the net atomic charge on nth carbon atom and $\sum\delta$ is the total electron surface density.

As we compare Eq. 4 (model 2) with our model, represented by Eq. 1, it is possible to observe that they show similar R (0.821 vs 0.790) using the same number of descriptors. A superficial analysis of this comparison would take to the conclusion that both QSAR models are equivalent from statistical point of view. Nevertheless, a deep analysis of the problem shows that the model 2 presents a high decrease of the F Fisher ratio (40%). Furthermore, the authors do not show in their analysis any validation for this model (i.e., cross-validation or test set). This validation process is very important to demonstrate that the QSAR model can be used for predictive purposes. A QSAR model lacks all kind of value if it presents a good fit but it is unable to make predictions. For that reason the statistical fit should not be confused with the ability of a model to make predictions. In this connection, we developed the validation process of the model 2 and found that presents a squared regression coefficient of the cross-validation (q^2) of about 0.563. This result indicates that model 2 possesses a very limited prediction capability in comparison to model 1, so this result confirms the superiority of our model.

Although affinity labelling studies show non- or uncompetitive behaviour of AR enzyme inhibitors in the NADPH-dependent reduction of glucose to sorbitol,[38] recently evidences supporting the active site as the inhibitor binding site have been reported.[39] This enzyme operates on a large scale of structurally different substrates. To achieve this pronounced promiscuity, the enzyme

Table 4. Values of molecular descriptors used for obtaining model 1

Compounds	Descriptors			
	GGI2	GGI3	JGI4	JGI6
1	4.000	2.813	0.104	0.037
2	4.000	2.688	0.116	0.032
3	4.000	2.688	0.112	0.032
4	4.000	2.625	0.095	0.035
5	4.000	2.688	0.112	0.030
6	3.556	2.250	0.099	0.030
7	4.000	2.563	0.106	0.038
8	4.000	2.563	0.112	0.030
9	3.556	2.250	0.098	0.029
10	3.556	2.250	0.098	0.028
11	3.556	2.375	0.104	0.029
12	4.000	2.438	0.108	0.029
13	3.556	2.375	0.100	0.030
14	3.556	2.500	0.109	0.031
15	3.556	2.250	0.098	0.028
16	4.000	2.688	0.109	0.041
17	3.556	2.125	0.098	0.028
18	3.556	2.375	0.100	0.029
19	3.556	2.250	0.097	0.028
20	4.000	2.688	0.105	0.036
21	3.556	2.500	0.106	0.031
22	3.556	2.375	0.104	0.029
23	3.556	2.250	0.093	0.027
24	4.000	2.688	0.109	0.040
25	4.000	3.063	0.104	0.058
26	3.556	2.250	0.099	0.029
27	4.000	2.813	0.107	0.040
28	4.000	2.813	0.107	0.038
29	3.556	2.500	0.098	0.027
30	3.556	2.500	0.106	0.029
31	4.000	2.688	0.107	0.039
32	3.333	2.375	0.102	0.032
33	3.556	2.375	0.105	0.029
34	3.556	2.250	0.101	0.028
35	4.000	2.563	0.105	0.028
36	3.111	2.188	0.092	0.026
37	5.111	3.438	0.097	0.043
38	2.000	1.188	0.087	0.015
39	2.222	1.313	0.076	0.019
40	2.444	1.250	0.065	0.015
41	2.222	1.438	0.081	0.024
42	4.000	2.813	0.103	0.027
43	4.000	2.813	0.109	0.031
44	3.556	2.500	0.097	0.028
45	3.556	2.500	0.096	0.026
46	3.111	2.188	0.096	0.025
47	3.111	2.313	0.101	0.027
48	4.000	3.063	0.114	0.044
49	5.333	3.688	0.101	0.033
50	1.778	0.938	0.056	0.008
51	5.333	3.688	0.100	0.033
52	4.889	3.500	0.094	0.028
53	5.111	3.563	0.098	0.042
54	5.778	4.000	0.103	0.033
55	3.333	2.625	0.110	0.029
56	5.778	3.875	0.103	0.038
57	2.444	1.875	0.080	0.020
58	4.000	2.813	0.103	0.026
59	2.222	1.688	0.076	0.012
60	2.444	1.438	0.086	0.018
61	2.222	1.500	0.076	0.014
62	2.222	1.750	0.087	0.015
63	4.000	2.813	0.109	0.029
64	2.222	1.375	0.078	0.012
65	2.889	2.500	0.101	0.032
66	2.444	1.688	0.080	0.023
67	2.000	1.125	0.063	0.017
68	3.333	2.500	0.099	0.032
69	1.778	0.938	0.056	0.008
70	2.222	1.375	0.078	0.012
71	4.000	2.688	0.109	0.029
72	3.111	2.438	0.102	0.029
73	3.111	2.625	0.091	0.028
74	2.000	1.813	0.076	0.019
75	3.111	2.625	0.091	0.030

can adapt rather flexibly to its substrates. Likewise, it has a similar adaptability for the binding of inhibitors.[40] In the active site, inhibitors bind to Tyr48, His110 and Trp111 residues by hydrogen bonds and accommodate their hydrophobic portion between hydrophobic residues (Trp111, Leu300) in a 'flexible pocket'.[40]

In our data set two main structural features appear to be relevant for displaying a high inhibitory activity. One of them is the present of the carbonyl group on the aromatic rings, Rastelli et al.[41] have reported that carbonyl groups on AR inhibitor structures can form hydrogen bonds with Tyr48, and His110 residues in the active site. On the other hand, the absence of the OH substituent at position 4′ drastically decreases the inhibitory activity as can be observed for the majority of the less active derivatives. In this sense, Constantino et al.[42] suggested that such residue interacts with Thr113 when the phenyl substituent is well accommodated on the hydrophobic pocket of the AR enzyme active site formed by Trp111 and Leu300. Moreover, hydroxyl and methoxyl substituents at positions 5, 6, 7 and 8 modulate AR inhibitory activity. Among the flavonoid derivatives here studied, compound 71 could be consider as an outliers taking into account its high residual when are predicted by our linear model. Indeed, the low AR inhibitory activity reported for Štefanič-Petek et al.[16] for this compound does not match with its structural features.

Galvez Charge Indices have been successfully employed for encoding chemical structures in previous QSAR studies of biological activities.[43,44] These descriptors contain important information on the relationship between the compound structures and its activities by describing the molecular topology and the charge transfer through the molecule.[18,19] Our data set do not vary appreciably in size and/or shape, mainly relative positions of hydroxyl and methoxyl groups change around the three-phenyl ring skeleton. Then, electrotopological conformation is the main feature that it is varying from one compound to another. That is why, among all kind of tested descriptors, Galvez Charge Indices are the 'best' explaining the variance of the data.

The participation in our linear model of descriptors of lag four and six may be viewed in terms of association of activity information content with structural

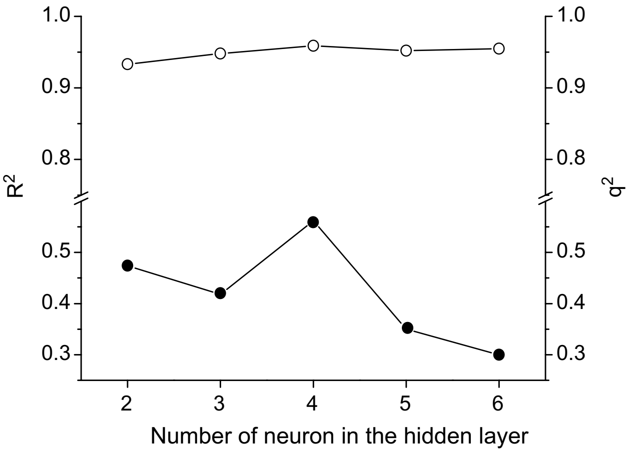


Figure 1. R^2 (○) and q^2 (●) values of LOO cross-validation for the five variables ANN models versus number of neurons in the hidden layer.

fragments of such size. Noteworthy, even lag fragments (two, four, six) have a positive contribution to the inhibitory activity meanwhile odd lag fragment (three) has a negative effect. These facts could be associated with the importance of keeping the three rings basic structure of the inhibitors for matching properly on the AR active site. In this regard, one can note that highest values of descriptor of lag three (over 3.00) correspond to bulky pyranose-ring substituted flavonoid derivatives. However, further deciphering of the information content of these descriptors is very complex as their computations involve integration of the structural fragments and due to this it is not possible to traverse backward from a higher state to a lower one.[45]

Finally, as have been previously explained an aim of this work is to apply ANN models for predicting AR inhibitory activity in order to improve linear model performance. In this connection, ANN models were generated with the Galvez descriptors that typified the best linear models. Networks were training with numbers of hidden neurons varying from 2 to 6.

Non-linear models with three and four variables as network inputs had R^2 values ranging from 0.80 to 0.85, but lacked of predictive power showing q^2 values lower than 0.5. However, five inputs ANNs showed higher statistic significance and stability. Figure 1 depicts R^2 and q^2 values for models with architecture 5-2-1, 5-3-1, 5-4-1, 5-5-1 and 5-6-1.

The R^2 values were slightly increased with the augment of the number of neurons in the hidden layer until a maximum value of about 0.960 was reached. On the other hand, cross-validation q^2 reached 0.612 maximum values for the 5-4-1 architecture. Concluding, even when the best ANN model had a better statistical fit in comparison with the linear one, the cross-validation analysis showed similar q^2 value.

In this case, we have a clear example of overfitting. In QSAR, the use of models that are more flexible than it needs should be avoided. ANNs obtained here are a good example, this methodology is capable to well accommodate some non-linear relationships and so is more flexible than a simple linear regression. For that reason, its use on a data set that acceptably conforms to a simple linear regression will add a level of complexity without any corresponding benefit in performance or, even worse, with poorer performance than the simpler model.

4. Concluding remarks

We have shown that the Galvez Charge Indices approach is able to describe the inhibitory effects on AR enzyme of natural and synthetic flavonoids analogues. In fact, we have developed a model for predicting this effect on a data set of 75 compounds, which is both statistically and chemically sounded. This model explains more than 67% of the variance in the experimental activity with an acceptable predictive power. These features are significantly better than that obtained from other ten different methodologies using multilinear regression analysis and also non-remarkable better performance was found when using ANNs for re-fitting the linear models.

Acknowledgements

Maykel Perez Gonzalez would like to express his sincere gratitude to Professor Jorge Galvez for useful comments and sending valuable information for the development of this work.

References and notes

1. Kinoshita, J. H. *Invest. Ophthalmol.* **1974**, *13*, 713.
2. McClure, J. W. Physiology of Flavonoids in Plants. In *Plant Flavonoids in Biology and Medicine: Biochemical, Pharmacological, and Structure–Activity Relationships*; Cody, V., Middleton, E., Harborne, J. B., Eds.; Alan R. Liss: New York, 1986; pp 77–85.
3. Middleton, E.; Kandaswami, C. The Impact of Plant Flavonoids on Mammalian Biology: Implications for Immunity, Inflammation and Cancer. In *The Flavonoids: Advances in Research Since 1986*; Harborne, I. R., Ed.; Chapman and Hall: London, 1993; pp 619–645.
4. Carroll, K. K.; Guthrie, N.; So, F. V.; Chambers, A. F. Anticancer Properties of Flavonoids, with Emphasis on Citrus Flavonoids. In *Flavonoids in Health and Disease*; Rice-Evans, C. A., Packer, L., Eds.; Marcel Dekker: New York, 1998; pp 437–446.
5. Hertog, M. G. L.; Hollman, P. C. H.; Katan, M. B.; Kromhout, D. *Nutr. Cancer* **1993**, *20*, 21.
6. Iwu, M. M.; Igboko, O. A.; Okunji, C. O.; Tempesta, M. S. *J. Pharm. Pharmacol.* **1989**, *42*, 290.
7. Varma, S. D. Inhibition of Aldose Reductase by Flavonoids: Possible Attenuation of Diabetic Complications. In *Plant Flavonoids in Biology and Medicine: Biochemical, Pharmacological, and Structure–Activity Relationships*; Cody, V., Middleton, E., Harborne, J. B., Eds.; Alan R. Liss: New York, 1986; pp 343–358.
8. Okuda, J.; Miwa, I.; Inagaki, K.; Horie, T.; Nakayama, M. *Chem. Pharm. Bull.* **1984**, *32*, 767.
9. Varma, S. D.; Kinoshita, J. H. *Biochem. Pharmacol.* **1976**, *25*, 2505.

10. Dvornik, D.; Simard-Duquesne, N.; Krami, M.; Ŝestanj, K.; Gabbay, K. H.; Kinoshita, J. H.; Varma, S. D.; Merola, L. O. *Science* **1973**, *182*, 1146.
11. Varma, S. D.; Mikuni, I.; Kinoshita, J. H. *Science* **1975**, *188*, 1215.
12. Kador, P. F.; Sharpless, N. E. *Biophys. Chem.* **1978**, *8*, 81.
13. Kador, P. F.; Sharpless, N. E. *Mol. Pharmacol.* **1983**, *24*, 521.
14. Kinoshita, J. H.; Kador, P. F.; Datiles, M. *J. Am. Med. Assoc.* **1981**, *246*, 257.
15. Matsuda, H.; Morikawa, T.; Toguchida, I.; Yoshikawa, M. *Chem. Pharm. Bull.* **2002**, *50*, 788.
16. Štefanič-Petek, A.; Krbavčič, A.; Šolmajer, T. *Croat. Chem. Acta* **2002**, *75*, 517.
17. Todeschini, R.; Consonni, V.; Pavan, M. Dragon. Software version 2.1, 2002.
18. Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520.
19. Gálvez, J.; Garcia-Domenech, R.; de Julihn-Ortiz, J. V.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 272.
20. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
21. MOPAC version 6.0. Frank J. Seiler Research Laboratory, US Air Force Academy, Colorado Springs, CO, 1993.
22. Rüker, G.; Rüker, C. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683.
23. Gutman, I.; Rüker, C.; Rüker, G. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 739.
24. Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28.
25. Randic, M. *New J. Chem.* **1995**, *19*, 781.
26. Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *J. Vib. Spectrosc.* **1999**, *19*, 151.
27. Hemmer, M. C.; Gasteiger, J. *J. Anal. Chim. Acta* **2000**, *420*, 145.
28. Schuur, J. H.; Setzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334.
29. Todeschini, R.; Vighi, M.; Provenzani, R.; Finizio, A.; Gramatica, P. *Chemosphere* **1996**, *32*, 1527.
30. Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 79.
31. Consonni, V.; Todeschini, P.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682.
32. STATISTICA version 6.0. Statsoft, Inc, 2002.
33. Sumpter, B. G.; Getino, C.; Noid, D. W. *Annu. Rev. Phys. Chem.* **1994**, *45*, 439.
34. Aoyama, T.; Suzuki, Y.; Ichikawa, H. *J. Med. Chem.* **1990**, *33*, 905.
35. Kolmogorov, A. N. *Dokl. Akad. Nauk SSSR* **1957**, *114*, 953.
36. Matlab version 6.5. The MathWorks, Inc., 2002.
37. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1.
38. Kador, P. F.; Lee, Y. S.; Rodriguez, L.; Sato, S.; Malik, A. B.; Ghany, Y. S.; Miller, D. D. *Bioorg. Med. Chem.* **1995**, *3*, 1313.
39. Constantino, L.; Rastelli, G.; Vescovini, K.; Cignarella, G.; Vianello, P.; Del Corso, A.; Cappiello, M.; Mura, U.; Barloco, D. *J. Med. Chem.* **1996**, *39*, 4396.
40. Wilson, D. K.; Tarle, I.; Petrash, J. M.; Quiocho, F. A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 9847.
41. Rastelli, G.; Ferrari, A.; Costantino, M. L.; Gamberini, M. C. *Bioorg. Med. Chem.* **2002**, *10*, 1437.
42. Costantino, L.; Del Corso, A.; Rastelli, G.; Petrash, J. M.; Mura, U. *Eur. J. Med. Chem.* **2001**, *36*, 697.
43. Kios-Santamarina, I.; Garcia-Domenech, R.; Gálvez *J. Bioorg. Med. Chem. Lett.* **1998**, *18*, 477.
44. Calabuig, C.; Antón-Fos, G. M.; Gálvez, J.; García-Doménech, R. *Int. J. Pharm.* **2004**, *278*, 111.
45. Gupta, M. K.; Sagar, R.; Shaw, A. K.; Prabhakar, Y. S. *Bioorg. Med. Chem.* **2005**, *13*, 343.