A High Dimensional QSAR Study on the Aldose Reductase Inhibitory Activity of Some Flavones: Topological Descriptors in Modeling the Activity†

Yenamandra S. Prabhakar,*,‡ Manish K. Gupta,‡ Nobendu Roy,§ and Yenamandra Venkateswarlu||

Medicinal and Process Chemistry Division, Central Drug Research Institute, Lucknow-226 001, India, and Strand Genomics Ltd., Bangalore-560 080, India, and Natural Products Laboratory, Indian Institute of Chemical Technology, Hyderabad-500 007, India

Received February 18, 2005

The quantitative structure−activity relationships (QSAR) of the Aldose Reductase (AR) inhibitory activity of 48 flavones were studied using Free−Wilson, Combinatorial Protocol in Multiple Linear Regression (CP-MLR), and Partial Least Squares (PLS) procedures. For the latter two procedures 152 Molconn-Z parameters and six indicators corresponding to the hydroxyls of flavones were used as molecular descriptors. Independently, all procedures suggested the significance of hydroxyls in modulating the activity of these compounds. The CP-MLR procedure identified 26 descriptors to model the activity. They suggested that structures rich in aromatic CH fragments, with a limited number of aliphatic fragments such as −CH2−, −CH<, and free hydroxyls at 7-, 3′-, and 4′-positions of the 2-arylbenzpyran-4-one core would be preferred for the activity. The PLS analysis agreed with the information content and the relative significance of the descriptors identified in the CP-MLR for modeling the activity. The study offers the scope to modulate the inhibitory activity of these compounds.

INTRODUCTION

In healthy people, glucose is metabolized through the Embden-Meyerhoff pathway. In cases of diabetes mellitus, with the increased levels of glucose in insulin-insensitive tissues the Aldose Reductase (AR) in the polyol pathway facilitates the conversion of glucose to sorbitol. In this cascade of events the accumulated sorbitol is attributed to be responsible for cataract, neuropathy, and retinopathy in diabetic cases.[1,2] Thus, the inhibition of AR in the polyol pathway may prevent and lead to the cure of the complications arising out of diabetes mellitus. In this background, Matsuda and co-workers[3] studied the AR inhibitory activity of a large number of flavones and related compounds from traditional antidiabetic remedies. Here, many of these compounds shared 2-arylbenzpyran-4-one as a scaffold for different chemical groups surrounding this moiety. This offers the scope to investigate the AR inhibitory activity of these compounds in relation to the functional group environment surrounding this core moiety. In this connection, application of graph theory to chemical structures results in several topological, topographical, and related descriptors characteristic to the molecular graphs from different perspectives. POLLY,[4] Molconn-Z,[5] CODESSA,[6] DRAGON,[7] TOPSMODE,[8] etc. are some well-known or recent programs embedded with graph theoretical concepts for characterizing the chemical structure and compute a large number of descriptors for modeling and other studies. When dealing with a large number of descriptors in the modeling studies, for the optimum utilization of contents of the generated data sets, it is necessary to identify different models as well as information rich descriptors corresponding to the phenomenon under investigation. The Genetic Function Approximation (GFA),[9] MUtation and SElection Uncover Models (MUSEUM),[10] and Combinatorial Protocol in Multiple Linear Regression (CP-MLR)[11] are a few approaches which address the evolution of multiple models and in doing so identifying contributing descriptors in quantitative structure−activity relationship (QSAR) and quantitative structure−property relationship (QSPR) studies. Here we have considered the CP-MLR approach[11−14] to discover the structure−activity models and contributing descriptors for the AR inhibitory activity of flavones[3] (Figure 1) in terms of different graph theoretical descriptors obtained from Molconn-Z software.[5] In this, while each equation (model) addresses different substructural regions and attributes in the predictive and diagnostic aspects of the chosen phenomenon, the identified descriptors from the cross-section of the models

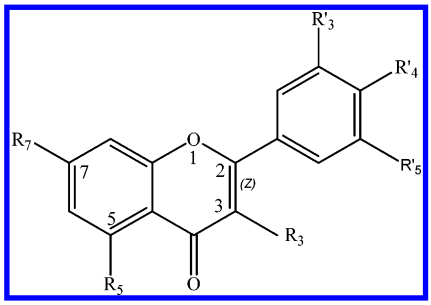

Figure 1. General structure of the flavones associated with rat lens aldose reductase inhibitory activity. In this, R3 and R7 may be H, OH, OMe, or O-glycoside and R5, R′3, R′4, and R′5 may be H, OH, or OMe.

* Corresponding author phone: +91-522-2612411; fax: +91-522-2624305; e-mail: yenpra@yahoo.com.

† This article is based on material presented in the Fourth Indo−US Workshop on Mathematical Chemistry, January 8-12, 2005, University of Pune, Pune, India organized by Natural Resources Research Institute, University of Minnesota Duluth, Duluth, Minnesota, USA and University of Pune, Pune, India. C.D.R.I. Communication No. 6719.

‡ Central Drug Research Institute.

§ Strand Genomics Ltd.

|| Indian Institute of Chemical Technology.

10.1021/ci050060u CCC: $33.50
Published on Web 04/28/2005

provide the scope to understand the predictive and diagnostic aspects of different substructural regions beyond the individual models. With this perspective we carried out a QSAR study on the flavones[3] (Figure 1) to investigate the different structural attributes and their information content in rationalizing the activity of these analogues. The results are presented here.

MATERIALS AND METHODS

Data Set. The structural information of flavones and their reported rat lens aldose reductase (AR) inhibitory activity (IC_{50}, i.e., concentration, in moles per liter, required to produce 50% inhibition) transformed in the form of logarithm of the inverse of inhibitory concentration ($-\log IC_{50}$) have been listed in Table 1.[3] Both the Free−Wilson[15] (nonparametric approach) and Molconn-Z descriptors[5] have been considered for the analysis of the inhibitory activity of these compounds. In the Free−Wilson method, the 2-arylbenzpyran-4-one moiety has been considered as a parent skeleton with the rest of the groups as substituents. In this way the substituent groups of these compounds (Table 1) have been represented by 20 Free−Wilson descriptors (compounds-to-parameters ratio is 2.28). The Molconn-Z descriptors[5] of these compounds have been computed from the SMILES notations of the structural drawings created in CS ChemBats3D Ultra.[16] This resulted in 152 nonzero and nonidentical descriptors characterizing the atom/group/path-type counts, simple/valence Chi indices, shape/complexity indices, H-bond donor/acceptor counts, and electrotopological state (E-state) sums of these compounds.[17] The CP-MLR protocol has been applied on this data set to identify the all-possible models that could emerge from these descriptors. All those descriptors which have taken part in the CP-MLR models have been further analyzed in comparison with the leftover descriptors using the partial least squares (PLS) procedure[18,19] by deriving MLR-like PLS models. The computational procedure involved in CP-MLR[11] and the model validation is briefly described below.

Computational Procedure. CP-MLR is a 'filter' based variable selection procedure involving selected subset regressions for model development in QSAR and QSPR studies.[11] In this procedure a combinatorial strategy with appropriately placed 'filters' has been interfaced with MLR to result in the extraction of diverse structure−activity models, each having a unique combination of descriptors from the data set under study. Models are discovered in this procedure within the predefined limits of a minimum and a maximum number of descriptors per model—termed as 'model search perimeter'. Here the 'filters' are significance evaluators of the variables in regression at different stages of model development. Of these, filter-1 is set in terms of inter-parameter correlation cutoff criteria for variables to stay as a subset (filter-1, default value 0.3). The second filter is set in terms of *t*-values of regression coefficients of variables associated with a subset (filter-2, default value 2.0). The third filter is set in terms of a predefined threshold level of r-bar (square-root of adjusted multiple correlation coefficient of regression equation) (filter-3, default value 0.74) to evaluate the advantage of a variable in models with varying degrees of freedom. Finally, to exclude false or artificial correlations, the external consistency of the variables of a model have been addressed in terms of cross-validated R^2 (Q^2) criteria with a leave-one-out (LOO) cross-validation procedure as a default option (filter-4, default limits $0.3 \leq Q^2 \leq 1.0$). In addition to cross-validation, each identified model has been reassessed for the chance correlations, if any, by repeated randomization of the biological response.[12,20] The data sets with a randomized response vector have been subjected to multiple regression analysis. The emerging regression equations, if any, with correlation coefficients better than or equal to the one corresponding to unscrambled response data were counted. Every model has been subjected to 100 such simulation runs. This has been used as a measure to express the percent chance correlation of the model under examination. In this study all the models of the AR inhibitory activity of flavones have been identified under the default filter thresholds conditions i.e., filter-1 as 0.3, filter-2 as 2.0, filter-3 as 0.74, and filter-4 as $0.3 \leq Q^2 \leq 1.0$ of CP-MLR protocol. As the total number of descriptors involved in this study is large, only those descriptors participating in the models have been addressed in the discussion.

RESULTS AND DISCUSSION

2-Arylbenzpyran-4-one scaffold has attracted the modeling perception of other workers as well for its AR inhibitory activity.[21,22] These modeling studies involved selected topological, classical, and quantum chemical descriptors of the compounds. According to Stefanic-Petek and co-workers,[22] the hydrophobicity, size, and charge of the 2-phenyl substituents of the flavones are important parameters for modeling the AR inhibitory activity. Also, among the quantum chemical descriptors, the net electronic charges and total electron surface density of 2-phenyl substituents are found to influence the activity.[22] In the case of Amic and co-workers' study with topological and electronic descriptors, most of the compounds in the training set are coumarins.[21] In this, the sum of π-charges, Wiener number, molecular topological index, dipole moment, path lengths of different orders, and presence or absence of hydroxyls at different positions of these compounds are found to influence the activity. In this background the compounds listed in Table 1 were analyzed using both the nonparametric (Free−Wilson[15]) and parametric (Molconn-Z descriptors[5]) approaches. Table 2 presents the summary of Free−Wilson analysis in the form of coefficients (or contributions) of the functional groups to the AR inhibitory activity of these compounds (Table 1). In carrying out this analysis, no compound from these analogues has been excluded in order to obtain an overall estimation of all the functional groups' contributions to the activity. Among the functional groups with a prevalence of two or more, the coefficients corresponding to the R_5-position and the OMe of R_3-position are of little significance and consequence to the activity of these compounds. The coefficients of the remaining functional groups (Table 2) signify their relative importance in modulating the AR inhibitory activity of the 2-arylbenzpyran-4-one skeleton. This analysis has resulted in the genesis of a new compound, compound 57 (Table 1), whose in silico activity has been found to be as good as the best compound (compound 56) among the analogues under consideration. In the parametric approach, the 152 Molconn-Z descriptors of these compounds under the default filter threshold conditions of the CP-MLR protocol with a model search

Table 1. Observed and Modeled Rat Lens Aldose Reductase Inhibitory Activity of Flavones (Figure 1)

							$-\log IC_{50}$						
compd[a]	R_3	R_5	R_7	R'_3	R'_4	R'_5	obs[b]	eq 2[c]	eq 3	eq 4	eq 5	eq 6	PLS[d]
1	H	H	H	H	H	H		5.34	5.19	4.76	4.78	4.73	4.89
2	H	H	OH	H	H	H	5.00	5.45	5.38	5.09	4.98	4.89	4.96
3	H	OH	OH	H	H	H	5.07	5.57	5.48	5.33	5.10	5.07	5.01
4	H	OH	OMe	H	H	H		5.09	5.05	4.83	4.74	4.81	4.79
5	H	H	OH	H	OH	H	5.42	5.57	5.61	5.45	5.55	5.55	5.62
6	H	H	H	OH	OH	H	6.43	5.80	5.88	5.80	6.11	6.15	6.18
7	H	H	OH	OH	OH	H	6.52	5.91	6.06	6.09	6.26	6.28	6.40
8	H	OH	OH	H	OH	H	5.66	5.68	5.70	5.68	5.65	5.73	5.65
9	H	OH	OMe	H	OMe	H		4.73	4.83	4.67	4.59	4.66	4.71
10	H	OH	SU1	H	OH	H	4.64	5.20	4.99	5.21	5.01	4.89	4.89
11	H	OH	SU2	H	OMe	H	5.33	4.73	4.73	4.55	4.55	4.56	4.42
12	H	OH	OH	OH	OH	H	6.35	6.03	6.14	6.29	6.33	6.43	6.40
13	H	OH	OH	OH	OMe	H	5.07	5.55	5.70	5.76	5.61	5.60	5.67
14	H	OH	OMe	OH	OMe	H	4.92	5.07	5.26	5.24	5.22	5.31	5.45
15	H	OH	OMe	OMe	OMe	H	4.14	4.36	4.50	4.65	4.63	4.85	4.83
16	H	OMe	OMe	OMe	OMe	H		4.36	4.51	4.66	4.64	4.43	4.88
17	H	OH	SU1	OH	OH	H	6.00	5.55	5.40	5.66	5.49	5.40	5.47
18	H	OH	SU3	OH	OH	H	5.51	5.55	5.90	5.63	6.13	5.79	5.93
19	H	OH	SU1	OH	OMe	H	4.64	5.07	4.97	5.13	4.78	4.57	5.05
20	OH	H	H	H	H	H		4.97	4.82	4.45	4.48	4.23	4.02
21	OH	OH	OMe	H	H	H		4.73	4.63	4.47	4.39	4.09	4.15
22	OH	OH	OH	H	OH	H	5.00	5.32	5.27	5.31	5.29	5.01	4.98
23	SU3	OH	OH	H	OH	H	5.29	5.32	5.46	5.28	5.62	5.25	5.49
24	OH	OH	OH	OH	OH	H	5.66	5.67	5.68	5.88	5.93	5.68	5.70
25	OH	OH	OMe	OH	OH	H	5.57	5.19	5.25	5.35	5.54	5.39	5.22
26	OH	OH	OH	OH	OMe	H	4.96	5.19	5.25	5.35	5.21	4.85	4.98
27	OMe	OH	OMe	OH	OH	H	6.09	5.19	5.26	5.35	5.55	5.61	5.48
28	OH	OH	OMe	OH	OMe	H	5.22	4.71	4.81	4.82	4.82	4.56	4.77
29	OMe	OH	OMe	OH	OMe	H	4.47	4.71	4.82	4.82	4.83	4.77	5.02
30	OH	OH	OMe	OMe	OMe	H	4.14	4.00	4.05	4.23	4.23	4.10	4.15
31	OMe	OH	OMe	OMe	OMe	H	4.60	4.00	4.05	4.24	4.23	4.32	4.39
32	OMe	OMe	OMe	OMe	OMe	H		3.53	3.62	3.70	3.84	3.60	4.31
33	SU1	OH	OH	OH	OH	H	5.35	5.67	5.29	5.72	5.36	5.28	5.57
34	SU4	OH	OH	OH	OH	H	5.52	5.67	5.29	5.72	5.36	5.28	5.57
35	SU5	OH	OH	OH	OH	H	6.82	5.67	5.81	5.74	6.04	6.18	5.95
36	SU6	OH	OH	OH	OH	H	6.75	5.67	5.52	5.77	5.67	6.35	6.08
37	SUI	OH	SU1	OH	OH	H	4.08	5.19	4.43	5.03	4.36	4.11	4.30
38	SU2	OH	OH	OH	OH	H	5.05	5.67	5.43	5.50	5.53	5.69	5.45
39	SU2	OH	OMe	OH	OH	H	4.68	5.19	5.01	4.97	5.15	5.41	5.28
40	SU2	OH	OMe	OH	OMe	H	4.39	4.71	4.58	4.44	4.44	4.59	4.73
41	SU2	OH	OMe	OMe	OMe	H	4.06	4.00	3.83	3.86	3.87	4.15	4.16
42	OH	H	OH	OH	OH	H	5.43	5.55	5.63	5.70	5.89	5.74	5.81
43	OH	OH	OH	OH	OH	OH	4.54	5.30	5.28	5.11	5.10	5.04	5.11
44	OH	OH	OH	OH	OMe	OH	4.72	5.30	5.29	5.13	4.80	4.53	4.56
45	OH	OH	OMe	OH	OH	OH	4.68	4.83	4.85	4.60	4.74	4.78	4.67
46	OMe	OH	OMe	OH	OH	OH	4.92	4.83	4.85	4.62	4.76	5.01	4.96
47	OH	OH	OMe	OH	OMe	OH	4.62	4.83	4.85	4.62	4.43	4.26	4.38
48	OH	OH	OMe	OMe	OMe	OH	4.36	4.12	4.09	4.39	4.26	4.05	4.07
49	OMe	OMe	OMe	OMe	OMe	OMe		3.16	3.21	2.94	3.02	2.75	3.56
50	SU5	OH	OH	OH	OH	OH	5.42	5.30	5.38	5.09	5.37	5.73	5.52
51	SU5	OH	OH	OH	OMe	OH	5.42	5.30	5.39	5.11	5.05	5.20	4.97
52	SU5	OH	OMe	OH	OMe	OH	4.32	4.83	4.95	4.59	4.67	4.92	4.65
53	SU5	OH	OH	OMe	OMe	OH	4.68	4.60	4.62	4.76	4.74	4.83	4.57
54	SU5	OH	OMe	OMe	OMe	OH	4.15	4.12	4.19	4.23	4.35	4.54	4.24
55	SU5	OH	OMe	OMe	OMe	OMe	4.15	3.64	3.75	3.42	3.62	3.89	3.70
56	SU7	OH	OH	OH	OH	OH	7.09	6.51	6.99	6.57	6.69	6.71	6.49
57[e]	SU7	OH	OH	OH	OH	H		6.87	7.42	7.14	7.28	7.06	6.84

[a] In these compounds the substituent groups corresponding to the SUgar moieties have been abbreviated as SU suffixed with a number as SU1 for O-β-D-glucopyranosyl; SU2 for O-β-D-glucopyranosyl(6→1)-O-α-L-rhamnopyranosyl; SU3 for :O-β-D-glucopyranosiduronic acid; SU4 for O-β-D-galacotopyranosyl; SU5 for O-α-L-rhamnopyranosyl; SU6 for O-α-L-arabinopyranosyl; and SU7 for O-(2′-galloyl)-α-L-rhamnopyranosyl. [b] Reference 3. [c] Activity from corresponding equation. [d] From four-component PLS model derived using the 26 contributing descriptors identified in the CP-MLR. [e] New compound.

perimeter as up to three descriptors has resulted in the identification of only one model (eq 1).

$$-\log IC_{50} = 11.230 - 6.716(1.013)\mathbf{Qv} - 10.065(3.488)(\mathbf{redundancy}) - 0.288(0.046)\mathbf{n4Pae[1,3]}$$
$$n = 48,\ r = 0.770,\ Q^2 = 0.513,\ s = 0.512,\ F = 21.35 \quad (1)$$

In eq 1 and in all other regression equations, n is the number of compounds, r is the correlation coefficient, Q^2 is cross-validated R^2 from the leave-one-out (LOO) procedure, s is the standard error of the estimate, and F is the F-ratio

Table 2. Free−Wilson Method Derived Functional Groups Contributions to the Aldose Reductase Inhibitory Activity ($-\log IC_{50}$) of Flavones (Figure 1)[a]

substituent	R_3	R_5	R_7	R'_3	R'_4	R'_5
H	0.160(0.126)15[b]	−0.044(−)6	0.503(−)1	−0.523(−)8	−0.076(−)2	0.175(0.050)32
OH	−0.170(0.107)12	0.006(0.026)42	0.175(0.076)23	0.189(0.049)32	0.306(0.069)26	−0.470(−)13
OMe	−0.028(0.212)4		−0.157(0.11)18	−0.232(0.168)8	−0.390(0.096)20	
SU1[c]	−0.946(0.303)2		−0.499(0.22)4			
SU2	−0.750(0.206)4		0.757(0.459)1			
SU3	0.010(−)1		−0.470(0.419)1			
SU4	−0.471(0.400)1					
SU5	0.373(0.184)7					
SU6	0.751(0.400)1					
SU7	1.738(0.422)1					
constant[d]	5.143					

[a] Free−Wilson regression statistics − total number of compounds 48, compound-to-parameter ratio 2.28, correlation coefficient 0.926, standard error 0.387, *F*-value 8.10. [b] Functional group contribution (standard error) and total number of occurrence in the corresponding substituent position; a dash in the parentheses indicates that the group contribution is from the restriction equation. [c] See footnote of Table 1 for SU1 to SU7. [d] Constant corresponds to the contribution of core of the structure (Figure 1) to the activity.

between the variances of calculated and observed activities. The values given in the parentheses are the standard errors of the regression coefficients. The in-depth characteristics of all descriptors identified in this study are available in ref 5 and links provided therein. In the discussion of models only the relevant descriptors have been briefly addressed in association with the activity. In eq 1, **Qv** is a topology parameter representing the general polarity, **n4Pae[1,3]** is the count of atom pairs where an atom with one connection (vertx **a**lpha) is separated by four bonds from another atom with three connections (vertx **e**psilon) and the '**redundancy**' is '(1.0-[Si/log10(nvx)])' in which Si is the Shannon information index, and nvx is the number of non-hydrogen atoms in molecule. Extension of the model search perimeter up to five descriptors resulted merely in two more models, having four-descriptors each, with **ndssC** (count of '>C=' fragments), **ESdssC** (electrotopological state of '>C=' fragments), **redundancy**, **n2Pag[1,2]** (count of alpha-gamma vertices), and **n3Pad[1,3]** (count of alpha-delta vertices) as participating descriptors. This prompted us to look for additional variables to further enrich the Molconn-Z descriptors of these compounds. In these compounds hydroxyl is one nontrivial and common functional group for all substituent positions. Also, in the Free−Wilson analysis of these compounds, the group coefficients corresponding to the hydroxyl at different positions, except for the one at R_5-position, are statistically worth making note of. In view of these observations to give additional weight to the hydroxyl groups at the R_3, R_5, R_7, R'_3, R'_4, and R'_5 positions of the 2-arylbenzpyran-4-one skeleton, six indicator parameters (I_3, I_5, I_7, I'_3, I'_4, and I'_5, respectively) have been defined and are included in the 152 Molconn-Z descriptors data set. These indicators take a value of 'one' if the respective substituent position is occupied by hydroxyl and 'zero' otherwise. It is worth mentioning the rigor of the CP-MLR procedure that even though in the Free−Wilson analysis five out of six of these hydroxyls are found to be associated with significant coefficients, in this procedure the indicators corresponding to these hydroxyls do not form an 'all-indicator-model' in any combination. However, the revised data set with 158 descriptors in the CP-MLR analysis, under identical filter conditions with the model search perimeter as up to five-descriptors, has resulted in the identification of 35 models (3 three-descriptor models, 21 four-descriptor models, and 11 five-descriptor models) for AR inhibitory activity of these compounds. All 35 models shared 26 descriptors among themselves. In all the models the *t*-values of regression coefficients are significant at the 95% level. In the randomization study (100 simulations per model) none of the identified models have shown any chance correlation. The essence of all these models has been provided in Table 3 in the form of identified descriptors' averaged regression coefficients together with their standard deviations across the models and the total incidence corresponding to all the models. This, while providing the averages of the estimated regression coefficients of all the identified descriptors, also shows their variance across the models emerged from the study. To maintain brevity, the complete regression equations have been shown for selected models with three-, four-, and five-descriptors each (eqs 2−4).

$$-\log IC_{50} = 1.714 + 0.362(0.062)\mathbf{nHaaCH} + 0.478(0.093)\mathbf{n2Pag[1,2]} + 0.709(0.164)\mathbf{I'_3}$$

$$n = 48,\ r = 0.780,\ Q^2 = 0.541,\ s = 0.502,\ F = 22.72 \quad (2)$$

$$-\log IC_{50} = 1.136 + 0.308(0.045)\mathbf{ESHaaCH} + 0.726(0.210)\mathbf{ESssCH2} + 0.446(0.084)\mathbf{n2Pag[1,2]} + 0.774(0.151)\mathbf{I'_3}$$

$$n = 48,\ r = 0.826,\ Q^2 = 0.615,\ s = 0.457,\ F = 23.15 \quad (3)$$

$$-\log IC_{50} = 1.414 + 0.285(0.046)\mathbf{ESHaaCH} + 0.034(0.013)\mathbf{ESsssCH} - 6.940(3.455)(\mathbf{redundancy}) + 0.446(0.084)\mathbf{n2Pag[1,2]} + 0.774(0.151)\mathbf{I'_3}$$

$$n = 48,\ r = 0.817,\ Q^2 = 0.569,\ s = 0.474,\ F = 16.81 \quad (4)$$

In these models **nHaaCH** is the count of aromatic CH; **ESHaaCH, ESssCH2**, and **ESsssCH** are electrotopological states (E-states), respectively, of aromatic CH, −CH2−, and >CH− fragments; **n2Pag[1,2]** is the count of atom pairs where an atom with one connection (vertx **a**lpha) is separated by two bonds from another atom with two connections (vertx **g**amma); and $\mathbf{I'_3}$ is the indicator parameter for the R'_3-substituent position of the core moiety. The descriptors **nHaaCH** and **ESHaaCH** are intercorrelated to a very large

Table 3. Identified Descriptors Contribution in Modeling the Rat Lens Aldose Reductase Inhibitory Activity of Flavones (Figure 1)

VarName[a]	av coeff (SD) incidence[b]	MLR-like PLS coeff (fc)[c]	VarName[a]	av coeff (SD) incidence[b]	MLR-like PLS coeff (fc)[c]
χchs6	5.942(−)1	2.089(0.038)	**Qv**	−6.716(−)1	−1.176(0.048)
nHaaCH	0.360(0.009)3	0.045(0.030)	**redundancy**	−8.509(0.951)14	−7.306(0.085)
nssCH	−0.410(−)1	−0.097(0.025)	**rad**	−0.107(−)1	0.018(0.013)
ndssC	0.751(0.042)8	0.182(0.023)	**mulrad**	−0.255(0.018)2	−0.143(0.046)
ESHaaCH	0.286(0.011)12	0.058(0.048)	**n2Pag[1,2]**	0.465(0.062)30	0.184(0.079)
ESssCH$_2$	0.619(0.151)2	0.316(0.054)	**n3Pad[1,3]**	−0.220(0.007)8	−0.046(0.046)
ESsssCH	0.031(0.005)2	0.004(0.012)	**n3Pad[2,3]**	−0.039(−)1	−0.003(0.006)
ESdssC	−0.459(0.023)6	−0.038(0.007)	**n4Pae[1,2]**	−0.095(0.003) 3	0.015(0.017)
ESSHBa	−0.005(−)1	0.001(0.018)	**n4Pae[1,3]**	−0.214(0.036)10	−0.101(0.091)
EShmax	−5.870(0.034)2	−0.336(0.013)	**I$_3$**	−0.404(0.046)3	−0.238(0.054)
tets1	−0.0006(−)1	0.00002(0.003)	**I$_7$**	0.393(0.010)3	0.098(0.026)
tets2	−0.005(−)1	0.00009(0.001)	**I′$_3$**	0.606(0.123)27	0.478(0.119)
IDCbar	−0.511(−)1	0.032(0.005)	**I′$_4$**	0.775(0.053)4	0.346(0.091)

[a] The descriptors are identified from models up to five parameters emerged from CP-MLR protocol with filter-1 as 0.3; filter-2 as 2.0; filter-3 as 0.74; and filter-4 as $0.3 \le Q^2 \le 1.0$, number of compounds in the study are 48; **χchs6** - simple chain (ring) of lengths (orders) 6; **nHaaCH** - count of aromatic CH; **nssCH$_2$** - count of −CH$_2$−; **ndssC** - count of =C<; **ESHaaCH** - electrotopological states (E-state) of aromatic CH; **ESssCH2** - E-state of −CH2−; **ESsssCH** - E-state of >CH−; **ESdssC** - E-state of =C<; **ESSHBa** - total E-state for strong hydrogen-bond acceptors; **EShmax** - E-state of maximum of hydrogen; **tets1** - total topological index based on E-state (normalized by path length); **tets2** - total topological index based on electrotopological state (normalized by the square of path length); **IDCbar** - Bonchev−Trinajstic information index; **Qv** - polarity descriptor; **redundancy** is (1.0 − [Si/log10(nvx)]); **rad** - graph radius; **mulrad** - multiplicity of graph radius; **nxPab[y,z]** represents the count (**n**) of atom pairs where an atom (vertx **a**) with **y** connection is separated by **x** bonds from another atom (vertx **b**) with **z** connections. Here **a** stands for alfa (α) vertx and **b** may represent any one vertx corresponding to **g** (gamma, γ) or **d** (delta, δ), or **e** (epsilon, ϵ). **I** - indicator parameter for 3-, 7-, 3′-, and 4′- positions of Figure 1. [b] The average regression coefficient of the descriptor corresponding to all models, its standard deviation (SD) and the total number of its incidence. The arithmetic sign of the coefficient represents the actual sign of the regression coefficient in the models. [c] MLR-like regression coefficient from the four-component PLS model of the identified descriptor; (fc) is fraction contribution of the regression coefficient to the activity. The constant term of this equation is 5.460.

extent (r=0.969) and carry similar information. These descriptors suggest that increased aromatic CHs in the compounds are favorable for AR inhibitory activity. Also, the count of −CH2− (**nssCH2)** has participated in a model to explain the activity (Table 3). This descriptor and **ESssCH2** are intercorrelated (r=0.982). These two descriptors and **ESsssCH** collectively suggest that an increase of the −CH2−/−CH< content in the structures is not in favor of the activity. The other descriptors of interest in the study are **ndssC** and **ESdssC** which are intended respectively for the count and electrotopological state of '>C=' fragment in the compounds. In these compounds this fragment is part of carboxylic function in glucouronic acid moiety. The regression coefficients of these descriptors suggest that it is not favorable for the activity. Among the vertex pair counts, **n2Pag[1,2]**, **n4Pae[1,2]**, **n3Pad[1,3]**, **n4Pae [1,3]**, and **n3Pad[2,3]** have taken part in the model formation (Table 3). Of these **n2Pag[1,2]** has participated in most of the models and suggests that an atom with one connection which is separated by two bonds from another atom with two connections is preferred for the activity. In other words this points toward the substituent groups' neighborhood in these compounds. Of the six indicators included in the data set, four (corresponding to 3-, 7-, 3′-, and 4′-positions) have taken part in the model formation (Table 3). Among these indicators the one meant for the 3-position suggests that a free hydroxyl at this center is not favorable for the activity. While the indictor parameters address the influence of individual hydroxyls present in the molecule, the coefficient of **ESSHBa** (electrotopological state index of total of strong H-bond acceptors) in Table 3 suggests the cumulative influence of this component on the activity. Also the descriptor **EShmax**, maximum of hydrogen electrotopological state, has participated in the models. This parameter addresses the activity of the compounds in relation to the polarity of hydrogens present in the compounds. The coefficient of this descriptor suggests that compounds with less polarized hydrogens would be better for the activity. The other descriptors of some interest in the models are **rad** (graph radius) and **mulrad** (multiplicity of graph radius). The coefficients of these descriptors suggest the necessity of compact graphs for better activity. The total topological state indices based on electrotopological states i.e., **tets1** (normalized by path length) and **tets2** (normalized by the square of path length) and the Bonchev−Trinajstic information index **IDCbar** have also participated in some models. However they have only trivial influence on the activity. The identified descriptors hold many more diverse models among them. The following are two such models with six and seven descriptors each obtained from the 26 identified descriptors (Table 3) in CP-MLR by redefining filter-1 as 0.79 (eqs 5 and 6). In eq 6, **χchs6** is a simple sixth order chain (ring) Chi index.

$$-\log IC_{50} = 1.623 + 0.276(0.045)\mathbf{ESHaaCH} + 0.942(0.201)\mathbf{ESsssCH2} - 8.182(3.033)(\mathbf{redundancy}) + 0.398(0.099)\mathbf{n2Pag[1,2]} + 0.601(0.147)\mathbf{I'_3} + 0.333(0.166)\mathbf{I'_4}$$

$$n = 48,\ r = 0.867,\ Q^2 = 0.675,\ s = 0.413,\ F = 20.78 \quad (5)$$

$$-\log IC_{50} = 2.599 + 10.938(2.315)\,\mathbf{\chi chs6} + 0.937(0.284)\mathbf{EsssCH2} - 8.812(2.292)(\mathbf{redundancy}) + 0.290(0.093)\mathbf{n2Pag[1,2]} - 0.211(0.042)\mathbf{n4Pae[1,3]} + 0.460(0.139)\mathbf{I'_3} + 0.540(0.150)\mathbf{I'_4}$$

$$n = 48,\ r = 0.882,\ Q^2 = 0.691,\ s = 0.396,\ F = 20.10 \quad (6)$$

The information content of the identified descriptors (Table 3) versus the leftover descriptors has been further investigated by splitting the composite original data set of 158 descriptors

into two groups—one with the 26 identified descriptors listed in Table 3 (data set-1) and the other with the 132 leftover descriptors (data set-2). PLS analysis has been carried out on both data sets to develop 'single window structure−activity models' comprising identified descriptors (data set-1) as well as leftover descriptors (data set-2). In the PLS analysis, the descriptors have been autoscaled to give each one of them equal importance. For the data set of the identified descriptors (data set-1), in the cross-validation procedure[18,19] four PLS components were found to be enough to explain the activity. The coefficients of the MLR-like equation of the PLS model of the identified descriptors are listed in Table 3.[17] This PLS model has explained 74% of the total variance in the activity of the compounds with a cross-validated R^2 value of 0.671 (r^2=0.740, s=0.413, F=30.69). To draw a simple comparison between the information contents of PLS components of identified descriptors (data set-1) and leftover descriptors (data set-2), a four-component PLS model for the activity has been developed from the leftover descriptors (data set-2) also. Here the PLS model with 132 leftover descriptors (data set-2) has explained only 58.7% of the total variance in the activity of the compounds (r^2=0.587, s=0.522, F=15.28) in comparison to the 74% of that of 26 identified descriptors. Also, the four-component PLS model of composite original data set with 158 descriptors (r^2=0.767, s=0.392, F=35.32) is only marginally better than that of the 26 identified descriptors. In other words, the 26 identified descriptors (Table 3) are enriched with the information content corresponding to the activity in comparison to the 132 leftover descriptors. All these results of PLS analyses are in general agreement with that of CP-MLR analysis. Among the identified descriptors, the coefficients of the MLR-like PLS model suggest that the descriptors **n2Pag[1,2]**, **n4Pae [1,3]**, **redundancy**, $\mathbf{I'_3}$, and $\mathbf{I'_4}$ have significant influence on the activity, whereas the descriptors **tets1**, **tets2**, **IDCbar**, and **n3Pad[2,3]** have only a marginal influence on the activity. The trend followed by the descriptors' incidence in CP-MLR models also points in this direction and conveys their relative importance. Independent of Free−Wilson analysis, all the models predict compound 57 (Table 1) as a high active compound.

CONCLUDING REMARKS

The core template of the compounds of this study (Table 1) and that of the compounds of Stefanic-Petek and co-workers[22] are structurally similar. Stefanic-Petek and co-workers[22] rationales for the AR inhibitory activity of the flavones involved the substituent contributions in terms of physicochemical and electronic properties of the 2-phenyl moiety of 2-arylbenzpyran-4-one scaffold. However, the rationales developed for the compounds listed in Table 1 are more comprehensive as they involve the complete structure in explaining the activity. Here, the flavones' Molconn-Z descriptors in association with the indicator parameters for the hydroxyl groups have yielded good models for the AR inhibitory activity of these compounds. Throughout the study the magnitude and sign of the regression coefficients of identified descriptors are stable. This adds importance to all the models and the participating descriptors. Also, the results of CP-MLR and PLS analysis support each other. Among the Molconn-Z descriptors the study identified several of them describing the electrotopological states, atom-cluster, and atom pair counts to model the activity. In atom-pair counts, the regression coefficient of **n2Pag[1,2]** suggests that structures rich in such atom pairs where an atom with one connection is separated by two bonds from another atom with two connections are preferred for the activity. This clearly points toward the structures with short chains or simple ring substitutions. The regression coefficients associated with the atom type count of aromatic CH fragments and its electrotopological state suggests their necessity for the activity. The regression coefficients of electrotopological states of aliphatic structural fragments CH2 and −CH< suggest for their minimization in the structures for better activity. The descriptors **ndssC** and **ESdssC** suggest that >C= is not in favor of activity. In these compounds it is part of the carboxylic function in the glucouronic acid moiety, and this offers the scope for modification to generate new compounds. The regression coefficients of the indicator parameters are in agreement with the Free−Wilson group contributions of hydroxyl groups. The coefficients suggest the favorability of free hydroxyls at 7-, 3′-, and 4′- positions of the 2-arylbenzpyran-4-one core structure (Figure 1). The AR inhibitory activities of the compounds have been predicted well by all the models (Table 1). The compound identified in the Free−Wilson analysis is also found to be active in the the CP-MLR as well as the PLS models. The identified descriptors of the study offer the scope to modify the flavones to modulate their AR inhibitory activity.

ACKNOWLEDGMENT

Evaluation version of Molconn-Z program for the descriptors generation is thankfully acknowledged.

Supporting Information Available: The complete data set (Table SI1); Free−Wilson restriction equations (Table SI2); and PLS loadings, weights, and sensitivity of independent and dependent descriptors (Table SI3). This material is available free of charge via the Internet at htpp://pubs.acs.org.

REFERENCES AND NOTES

(1) Terashima, H.; Hama, K.; Yamamoto, R.; Tsuboshima, M.; Kikkawa, R.; Hatanaka, I.; Shigeta, Y. Effect of New Aldose Reductase Inhibitors on Various Tissues in vitro. *J. Pharmacol. Exp. Ther.* **1984**, *229*, 226−230.

(2) Chung, S. S. M.; Ho, E. S. M.; Lam, K. S. L.; Chung, S. K. Contribution of Polyol Pathway to Diabetes-Induced Oxidative Stress. *J. Am. Soc. Nephrol.* **2003**, *14*, S233−S236.

(3) Matsuda, H.; Morikawa, T.; Toguchida, I.; Yoshikawa, M. Structural Requirement of the Flavonoids and Related Compound for Aldose Reductase Inhibitory Activity. *Chem. Pharm. Bull.* **2002**, *50*, 788−795.

(4) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. POLLY, University of Minnesota: Duluth, MN, 1988.

(5) Molconn-Z, ver. 2.07, eduSoft Lc, a Virginia Corporation, Ashland, VA 23005 U.S.A. www.edusoft-lc.com.

(6) (a) Katritzky, A. R.; Lobnov, V.; Karelson, M. *CODESSA* (*Comprehensive Descriptors for Structural and Statistical Analysis*); University of Florida: Gainesville, FL, 1994. (b) Katritzky, A. R.; Perumal, S.; Petrukhin, R.; Kleinpeter, E. CODESSA-Based Theoretical QSPR Model for Hydantoin HPLC-RT Lipophilicities. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 569−574.

(7) DRAGON software by Todeschini, R.; Consonni, V. Milano, Italy. http://disat.unimib.it/chm/Dragon.htm.

(8) Gonzalez, M. P.; Helguera, A. M. TOPS-MODE verces DRAGON Descriptors to Predict Permeability Coefficients through Low-Density Polymer. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 665−672.

(9) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure Activity Relationship and Quantitative Structure Property Relationship. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(10) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.
(11) Prabhakar, Y. S. A Combinatorial Approach to the Variable Selection in Multiple Linear Regression Analysis of Selwood et al. data set- A Case Study. *QSAR Comb. Sci.* **2003**, *22*, 583−595.
(12) Prabhakar, Y. S.; Solomon, V. R.; Rawal, R. K.; Gupta, M. K.; Katti, S. B. CP-MLR/PLS Directed Structure−Activity Modeling of the HIV-1 RT Inhibitory Activity of 2,3-Diaryl-1,3-Thiazolidin-4-Ones. *QSAR Comb. Sci.* **2004**, *23*, 234−244.
(13) Prabhakar, Y. S. A Combinatorial Protocol in Multiple Linear Regression to Model Gas Chromatographic Response Factor of Organophosphonate Esters. *Internet Electron. J. Mol. Des.* **2004**, *3*, 150−162, http://www.biochempress.com.
(14) Gupta, M. K.; Sagar, R.; Shaw, A. K.; Prabhakar, Y. S. CP-MLR Directed QSAR Studies on the Antimycobacterial Activity of Functionalised Alkenols − Topological Descriptors in Modeling the Activity. *Bioorg. Med. Chem.* **2005**, *13*, 343−351.
(15) Free, S. M., Jr.; Wilson, J. W. A Mathematical Approach to Structure−Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399.
(16) ChemDraw Ultra 6.0 and ChemBats3D Ultra, Cambridge Soft Corporation, Cambridge, U.S.A.
(17) The complete data set, Free−Wilson restriction equations, and PLS loadings, weights, and sensitivity of independent and dependent descriptors will be provided on request.
(18) Wold, S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Component Models. *Technometrics* **1978**, *20*, 397−405.
(19) Stahle, L. Wold, S. In *Progress in Medicinal Chemistry*; Eillis, G. P., West, W. B., Eds.; Elsevier Science Publishers: B.V. Amsterdam, 1988; Vol. 25, Chapter 6, pp 291−338.
(20) So, S. S.; Karplus, M. Three-Dimensional Quantitative Structure−Activity Relationship from Molecular Similarity Matrices and Genetic Neural Networks. 2. Applications. *J. Med. Chem.* **1997**, *40*, 4360−4371.
(21) Amic, D.; Davidovic-Amic, D.; Beslo, D.; Lucic, B.; Trinajstic, N. The Use of the Ordered Orthogonalized Multivariate Linear Regression in a Structure−Activity Study of Coumarin and Flavonoid Derivatives as Inhibitors of Aldose Reductase. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 581−586.
(22) Stefanic-Petek, A.; Krbavcic, A.; Solmajer, T.; QSAR of Flavonoids: 4. Differential Inhibition of Aldose Reductase and $p56^{lck}$ Protein Tyrosine Kinase. *Croat. Chem. Acta* **2002**, *75*, 517−529.

CI050060U