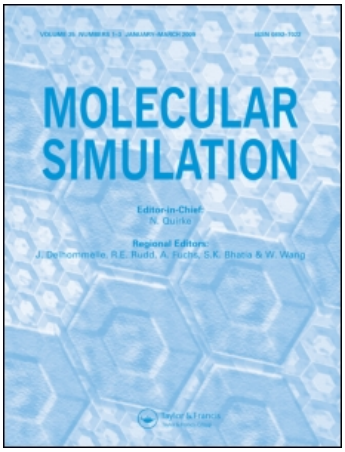


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Exploring QSAR of hydroxyphenylureas as antioxidants using physicochemical and electrotopological state atom parameters

Supratim Ray[a]; Partha Pratim Roy[a]; Chandana Sengupta[a]; Kunal Roy[a]
[a] Drug Theoretics and Cheminformatics Lab, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India

Online publication date: 11 May 2010

To cite this Article Ray, Supratim , Roy, Partha Pratim , Sengupta, Chandana and Roy, Kunal(2010) 'Exploring QSAR of hydroxyphenylureas as antioxidants using physicochemical and electrotopological state atom parameters', Molecular Simulation, 36: 6, 484 — 492
To link to this Article: DOI: 10.1080/08927021003664058
URL: http://dx.doi.org/10.1080/08927021003664058

Molecular Simulation
Vol. 36, No. 6, May 2010, 484–492

Taylor & Francis
Taylor & Francis Group

Exploring QSAR of hydroxyphenylureas as antioxidants using physicochemical and electrotopological state atom parameters

Supratim Ray, Partha Pratim Roy, Chandana Sengupta and Kunal Roy*[1]

Drug Theoretics and Cheminformatics Lab, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

(*Received 17 November 2009; final version received 30 January 2010*)

In the present study, free radical scavenging activity of 36 substituted hydroxyphenylurea derivatives was subjected to classical quantitative structure–activity relationship (QSAR) analyses using physicochemical (hydrophobicity and molar refractivity) and electrotopological state atom parameters. For the development of the QSAR models, statistical techniques such as stepwise multiple linear regression and genetic function approximation (GFA) were used. The developed models indicate an important contribution of the phenolic hydroxyl group of hydroxyphenylureas, apart from that of the urea moiety and piperazine nucleus of the side chain, to the free radical scavenging activity. The presence of substituents at the phenyl ring influences the electron density distribution over the phenolic ring system and modulates the activity. Hydrophobicity is found to contribute positively to the free radical scavenging activity. Based on internal validation (Q^2), external validation (R^2_{pred}) and overall validation criteria ($r^2_{m\,(\text{overall})}$), a GFA model with spline options was found to be the best model ($Q^2 = 0.957$, $R^2_{\text{pred}} = 0.966$, $r^2_{m\,(\text{overall})} = 0.914$).

Keywords: QSAR; antioxidant; hydroxyphenylurea; GFA

1. Introduction

Antioxidants are of great interest because of their involvement in protecting cells against reactive oxygen species produced through a free radical-mediated reaction mechanism. Free radicals are chemical compounds with unpaired electrons. They can emerge spontaneously and/or generate through catalysis of enzymes or transition metals in chemical or biological systems [1]. Owing to the great potential of free radicals to react with various compounds by electron transfer, proton transfer, H-atom abstraction or addition reaction, they are considered responsible for a series of undesired processes such as ageing, material degradation, food deterioration and many diseases [2,3]. Therefore, much interest has been focused on finding antioxidants to prevent the radical-induced impairments in chemical, food and pharmaceutical industries [4,5]. Antioxidants are widely used as ingredients in dietary supplements in the hope of maintaining health and preventing diseases such as cancer and coronary heart disease. Polyphenols, vitamin E derivatives and flavonoids have recently gained significant interest among various antioxidants [6]. Some novel hydroxyphenylurea derivatives also possess antioxidant activity. High inhibitory activity of hydroxyphenylurea derivatives against lipid peroxidation was also reported [7]. Electronic and steric effects of substituents on the phenolic hydroxyl group were shown to be of importance in governing the inhibitory potency [7]. Some derivatives also possess inhibitory potency evaluated against acyl-CoA: cholesterol acyltransferase enzyme system. The quantitative structure–activity relationship (QSAR) analysis revealed that their inhibitory activities were controlled by the hydrophobicity of the whole molecule, the substitution pattern of the urea moiety and the existence of carboxylic acid [8].

Based on these considerations, in the present paper, we model the free radical scavenging activity of a set of hydroxyphenylureas assayed in peripheral multinuclear neutrophil cells [9] using electrotopological state (E-state) and physicochemical parameters to develop predictive QSAR models for the said class of antioxidants. The purpose of this modelling is to identify the features of the molecules contributing to the antioxidant activity and thereby to help in designing new antioxidant molecules of this class. The developed models may also be used to compute antioxidant activity of newly designed molecules so that loss due to synthesis and experimental testing of less potent molecules can be avoided.

2. Material and methods

2.1 Data-set and descriptors

In the present study, substituted hydroxyphenylurea derivatives ($n = 36$) were used as the model data-set. The number of compounds available under this particular category of antioxidants being limited, it may appear

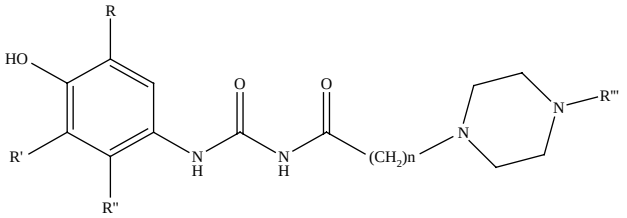
*Corresponding author. Email: kunalroy_in@yahoo.com

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927021003664058
http://www.informaworld.com

difficult to develop such models; however, this again justifies the requirement of development of such predictive models to design novel potent antioxidants. The developed models have been sufficiently validated using multiple strategies to check the reliability of the models. The free radical scavenging activity data [rate of change per unit added volume of free radicals, K_{app} (expressed as $\log(1/K_{app})$)] were subjected to QSAR analyses using physicochemical [$\log P$ (partition coefficient) and MR (molar refractivity)] and E-state atom parameters. The $\log P$ and MR of the compounds were calculated using the method of Ghose and Crippen [10] with ChemDraw Ultra software version 5.0 (http://www.cambridgesoft.com/). E-state parameters were calculated using in-house software SRETSA developed partly by one of the authors and standardised on known data-sets.[2] The program uses, as input, only the connection table in a specific format along with intrinsic state values of different atoms. To the output file thus obtained, the biological activity data were introduced to make it ready for subsequent regression analysis. There are five regions of structural variations: *n*, *R*, *R′*, *R″* and *R‴* positions (Table 1). The common atoms

Table 1. Structural features and calculated $\log P$ values of hydroxyphenylurea compounds.



Compound no.	*n*	*R*	*R′*	*R″*	*R‴*	$\log P$[a]
1	1	H	H	H	CH_3	−0.26
2	1	H	H	H	CH_2Ph	1.48
3	1	H	H	H	Ph	1.82
4	1	H	H	H	$2CH_3OC_6H_4$	1.69
5	1	H	OCH_3	H	CH_3	−0.38
6	1	H	OCH_3	H	CH_2Ph	1.35
7	1	H	OCH_3	H	Ph	1.69
8	1	H	OCH_3	H	$2CH_3OC_6H_4$	1.57
9	1	H	OC_2H_5	H	CH_3	−0.04
10	1	H	OC_2H_5	H	CH_2Ph	1.69
11	1	H	OC_2H_5	H	Ph	2.03
12	1	H	OC_2H_5	H	$2CH_3OC_6H_4$	1.9
13	1	CH_3	H	CH_3	CH_3	0.72
14	1	CH_3	H	CH_3	CH_2Ph	2.45
15	1	CH_3	H	CH_3	Ph	2.79
16	1	CH_3	H	CH_3	$2CH_3OC_6H_4$	2.67
17	2	H	H	H	CH_3	0.04
18	2	H	H	H	CH_2Ph	1.77
19	2	H	H	H	Ph	2.11
20	2	H	H	H	$2CH_3OC_6H_4$	1.99
21	2	H	OCH_3	H	CH_3	−0.09
22	2	H	OCH_3	H	CH_2Ph	1.64
23	2	H	OCH_3	H	Ph	1.99
24	2	H	OCH_3	H	$2CH_3OC_6H_4$	1.86
25	2	H	OC_2H_5	H	CH_3	0.25
26	2	H	OC_2H_5	H	CH_2Ph	1.98
27	2	H	OC_2H_5	H	Ph	2.32
28	2	H	OC_2H_5	H	$2CH_3OC_6H_4$	2.2
29	2	OCH_3	OCH_3	H	CH_3	−0.22
30	2	OCH_3	OCH_3	H	CH_2Ph	1.52
31	2	OCH_3	OCH_3	H	Ph	1.86
32	2	OCH_3	OCH_3	H	$2CH_3OC_6H_4$	1.73
33	2	CH_3	H	CH_3	CH_3	1.01
34	2	CH_3	H	CH_3	CH_2Ph	2.74
35	2	CH_3	H	CH_3	Ph	3.09
36	2	CH_3	H	CH_3	$2CH_3OC_6H_4$	2.96

Note: [a]Calculated according to [10].

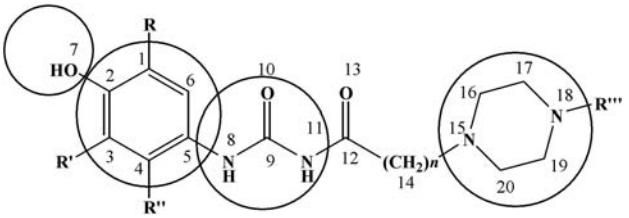


Figure 1. Common atoms of hydroxyphenylureas (with arbitrary numbering and important sites).

for all of the compounds (with arbitrary numbering) are shown in Figure 1.

2.1.1 *E-state atom index*

Structural specificity of a drug molecule is exhibited at an atomic or fragmental level instead of the whole molecule. In the drug–receptor interaction phenomenon, a portion of the molecule (pharmacophore) may play a more important role than the other segments. Although basic information for the constitution of topological indices is derived from the atom level (count of atoms, bonds, paths of bonds, etc.), most of the indices are applied to the whole molecule after summing up all the components over the whole molecule. Thus, QSAR studies at the atomic or fragmental level are justified in the present context [11].

The E-state atom index developed by Kier and Hall [12] is an atom-level descriptor encoding both the electronic character and topological environment of each skeletal atom in a molecule. The E-state of a skeletal atom is formulated as an intrinsic value I_i plus a perturbation term ΔI_i, arising from the electronic interaction within the molecular topological environment of each atom in the molecule.

The intrinsic value has been defined as the ratio of a measure of the electronic state (Kier–Hall valence state electronegativity) to the local connectedness. The count of valence electrons, which are the most reactive and involved in chemical reactions and bond formations, is considered in the expression of I to encode the electronic feature. To reflect differences in electronegativity among the atoms, principal quantum number is employed in the expression of I. The topological attribute is included using the adjacency count of the atom. The intrinsic value of an atom i is defined as

$$I_i = \frac{(2/N)^2 \delta^{\mathrm{v}} + 1}{\delta}, \tag{1}$$

where N stands for the principal quantum number and δ^{v} and δ indicate the count of valence electrons and sigma electrons, respectively, associated with the atom i in the hydrogen-suppressed graph. The intrinsic E-state calculated according to Equation (1) produces different values of an atom in different degrees of substitution (branching). The values are also different for different atoms having differences in electronegativity. The intrinsic values increase with an increase in electronegativity or electron richness and decrease with an increase in branching (substitution).

The perturbation factor for the intrinsic state of atom i is defined as

$$\Delta I_i = \sum_{j \neq i} \frac{I_i - I_j}{r_{ij}^2}, \tag{2}$$

where r_{ij} stands for the graph separation factor, i.e. the count of skeletal atoms in the shortest path connecting the atoms i and j including both atoms.

The summation of the intrinsic state of an atom and the influence of the field is called E-state of the atom,

$$S_i = I_i + \sum_{j \neq i} \Delta I_{ij}. \tag{3}$$

It is a representation of molecular structure information as it varies with changes in structural features including branching, cyclicity, heteroatom variation and changes in relative positions of different groups. The E-state considers both bonded and non-bonded interactions. The bonded component depends simply on differences in electronegativity among the adjacent atoms. The non-bonded interaction may be through an inductive effect across the skeleton and is a function of graph separation factor and electronegativity differences. Thus, the E-state represents electronic distribution information modified by both local and global topologies. The information encoded in the E-state value for an atom is the electronic accessibility at that atom.

The E-state index has been projected as a useful tool in the context of QSAR studies and reported to have power to identify atoms or fragments in the molecules, which are important for the biological activity [13–15]. The present group of authors have also used E-state parameters to explore QSARs of different pharmaceutically important classes of compounds [16–23]. In continuation of such efforts, the present study shows the utility of E-state parameters in QSAR studies by exploring QSAR analyses of free radical scavenging activity of substituted hydroxyphenylurea derivatives using physicochemical ($\log P$ and MR) and E-state atom parameters.

2.2 *Model development*

To begin the model development process, the whole dataset ($n = 36$) was divided into training ($n = 27$, 75% of the total number of compounds) and test ($n = 9$, 25% of the total number of compounds) sets using k-means clustering technique [24] applied on a standardised descriptor matrix.

QSAR models were developed using the training-set compounds (optimised by Q^2), and then the developed models were validated (externally) using the test-set compounds. For the development of the QSAR models, the statistical techniques used were stepwise multiple linear regression (MLR) and genetic function approximation (GFA). Stepwise MLR was performed using MINITAB software[3] on a Pentium 4 PC and GFA was performed using Cerius2 version 4.10 software[4] under QSAR+ environment on a Silicon Graphics O2 workstation running under the IRIX 6.5 operating system.

In stepwise regression, a multiple-term linear equation was built step by step [25]. The basic procedures involve: (1) identifying an initial model, (2) iteratively 'stepping', i.e. repeatedly altering the model of the previous step by adding or removing a predictor variable in accordance with the 'stepping criteria' ($F = 4$ for inclusion; $F = 3.9$ for exclusion) in our case and (3) terminating the search when stepping is no longer possible given the stepping criteria, or when a specified maximum number of steps has been reached. Specifically, at each step, all variables are reviewed and evaluated to determine which one will contribute most to the equation. That variable will then be included in the model, and the process is started again. A limitation of the stepwise regression search approach is that it presumes that there is a single 'best' subset of variables and seeks to identify it. There is often no unique 'best' subset, and all possible regression models with a similar number of X variables as in the stepwise regression solution should be fitted subsequently to study whether some other subsets of X variables might be better.

The GFA technique [26,27] was used to generate a population of equations rather than one single equation for a correlation between biological activity and descriptors. The GFA involves the combination of the multivariate adaptive regression splines algorithm with the genetic algorithm to evolve a population of equations that best fit the training-set data. It provides an error measure, called the lack-of-fit (LOF) score, that automatically penalises models with too many features. It also inspires the use of splines as a powerful tool for nonlinear modelling. A distinctive feature of GFA is that it produces a population of models (e.g. 100), instead of generating a single model, as do most other statistical methods. The range of variations in this population gives added information on the quality of fit and importance of the descriptors.

GFA can build models not only with linear polynomials but also with higher order polynomials, splines and Gaussians. By using spline-based terms, GFA can perform a form of automatic outlier removal and classification. The splines used are truncated power splines and denoted by angular brackets. For example, $\langle f(x) - a \rangle$ is equal to zero if the value of $(f(x) - a)$ is negative, else it is equal to $(f(x) - a)$. The constant a is called the knot of the spline. A spline partitions the data samples into two classes, depending on the value of some features. The value of the spline is zero for one of the classes and non-zero for the other classes. Splines are interpreted as performing either range identification or outlier removal. If there are many members in the non-zero partition, then the spline is identifying a range of effect. If there are only a few members of the non-zero set, this indicates that the spline is identifying outliers [27].

2.3 Statistical qualities

The statistical qualities of the equations were judged by the parameters such as squared correlation coefficient (R^2) and variance ratio (F) at specified degrees of freedom (df) [28]. For GFA-derived equations, LOF was taken as an objective function to select an equation. The generated QSAR equations were validated by leave-one-out cross-validation R^2 (Q^2) and predicted residual sum of squares (PRESS) [29–31] and were then used for the prediction of free radical scavenging activity of the test-set compounds. The prediction qualities of the models were judged by statistical parameters such as predictive R^2 (R^2_{pred}), squared correlation coefficient between observed and predicted values of the test-set compounds with (r^2) and without (r^2_0) the intercept. It was previously shown that the use of R^2_{pred} and r^2 might not be sufficient to indicate the external validation characteristics [32]. Thus, an additional parameter $r^2_{m\,(\text{test})}$ [defined as $r^2(1 - \sqrt{r^2 - r^2_0})$], which penalises a model for large differences between observed and predicted values of the test-set compounds, was also calculated. Two other variants [33] of the r^2_m parameter, $r^2_{m\,(\text{LOO})}$ and $r^2_{m\,(\text{overall})}$, were also calculated. The parameter $r^2_{m\,(\text{overall})}$ is based on prediction of both training- (LOO prediction) and test-set compounds. It was previously shown [33] that $r^2_{m\,(\text{LOO})}$ and $r^2_{m\,(\text{test})}$ penalise a model more strictly than Q^2 and R^2_{pred}, respectively. Additionally, selected models were subjected to randomisation tests to check the reliability of the models. Use of multiple strategies of validation ascertains acceptability of the models in spite of the limited number of molecules in the training and test sets.

3. Results and discussion

Membership of compounds in different clusters generated using k-means clustering technique is shown in Table 2. The test-set size was set to approximately 25% to the total data-set size [34] and the test-set members are given in Table 3. Statistical qualities of all important models are listed in Table 4. In all the models, differences between R^2 and Q^2 values for the models are not very high (less than 0.3) [35]. The values of selected E-state parameters

Table 2. *k*-Means clustering of compounds using standardised descriptors.

Cluster no.	No. of compounds in different clusters	Compounds (Sl. nos.) in each cluster										
1	3	**1**	**5**	**17**								
2	6	**2**	**3**	**4**	**7**	**18**	**19**					
3	11	**6**	**8**	**10**	**11**	**14**	**15**	**20**	**22**	**23**	**27**	**35**
4	10	**12**	**16**	**24**	**26**	**28**	**30**	**31**	**32**	**34**	**36**	
5	6	**9**	**13**	**21**	**25**	**29**	**33**					

of different compounds are listed in Table S1 of the Supplementary Material.

3.1 Stepwise regression

Setting the 'stepping criteria' ($F = 4$ for inclusion; $F = 3.9$ for exclusion), the following equation was obtained:

$$\log(1/K_{\text{app}}) = 35.529(\pm 2.581) - 18.847(\pm 1.561)S_6 - 1.394(\pm 0.168)S_4 + 0.102(\pm 0.015)\log P + 11.420(\pm 1.920)S_5 - 0.907(\pm 0.285)S_7,$$

$$n_{\text{Training}} = 27, \quad R^2 = 0.972, \quad R_a^2 = 0.965, \qquad (4)$$

$$F = 146.06(\text{df } 5, 21), \quad Q^2 = 0.953, \quad \text{PRESS} = 0.124,$$

$$n_{\text{Test}} = 9, R_{\text{pred}}^2 = 0.916, r_{m\,(\text{test})}^2 = 0.912,$$

$$r_{m\,(\text{LOO})}^2 = 0.885, \quad r_{m\,(\text{overall})}^2 = 0.896.$$

According to the standardised regression coefficients, the relative order of importance of the descriptors is as follows: $S_6 > S_4 > \log P > S_5 > S_7$. The standard errors of regression coefficients are given within parentheses. Equation (4) could explain 96.5% of the variance (adjusted coefficient of variation) and leave-one-out predicted variance was found to be 95.3%. While Equation (4) was applied for the prediction of test-set compounds, the predictive R^2 value for the test set was found to be 0.916.

The negative coefficients of S_6, S_4 and S_7 indicate that activity decreases with an increase in the E-state value of atoms 6, 4 and 7, respectively, while the positive coefficient of S_5 indicates that activity increases with an increase in the E-state value of atom 5. Compounds with high values of the E-state parameter for atom 6 (S_6) such as **35**, **34**, **33** and **36** showed poor activity while compounds with the corresponding values in the moderate range (such as **1**, **3**, **4**, **17** and **20**) showed good free radical scavenging activity. The presence of the $-CH_3$ substituent at *R* position increases the E-state value at position 6 (e.g. compounds **33**–**36** having comparatively lower activity values), while the presence of the methoxy substituent at *R* position decreases the E-state value at position 6 (e.g. compounds **29**–**32** having comparatively higher activity values). Again, compounds with lower values of S_7 (e.g. **1**, **3**, **4**, **17** and **19**) show activity in the higher range, and this indicates the importance of the phenolic hydroxyl

Table 3. Observed and calculated free radical scavenging activity ($\log K_{\text{app}}$) of hydroxyphenylurea compounds.

Sl. no.	Obs[a]	Cal[b]	Cal[c]	Cal[d]
Training set				
1	1.796	1.729	1.710	1.697
3	1.824	1.744	1.740	1.716
4	1.678	1.709	1.700	1.673
6	1.108	1.156	1.169	1.111
9	0.903	0.951	0.929	0.930
10	0.854	0.890	0.865	0.917
12	0.889	0.930	0.918	0.878
13	0.907	1.003	1.001	0.945
14	0.914	0.956	0.943	0.912
15	1.032	1.029	1.034	0.964
16	1.009	1.000	0.998	0.945
17	1.509	1.608	1.530	1.565
18	1.469	1.547	1.466	1.518
19	1.523	1.622	1.558	1.584
20	1.456	1.588	1.520	1.541
21	1.071	1.094	1.051	1.029
22	1.027	1.034	0.987	0.978
23	1.018	1.109	1.081	1.044
28	0.845	0.809	0.738	0.798
29	1.337	1.361	1.307	1.301
30	1.292	1.301	1.249	1.368
31	1.387	1.375	1.341	1.338
32	1.208	1.340	1.304	1.244
33	0.827	0.888	0.822	0.814
34	0.815	0.840	0.763	0.834
35	0.857	0.915	0.857	0.886
36	0.842	0.884	0.819	0.867
Test set				
2	1.721	1.670	1.647	1.650
5	1.168	1.217	1.233	1.162
7	1.081	1.231	1.261	1.178
8	1.036	1.197	1.223	1.051
11	0.987	0.965	0.958	0.970
24	0.987	1.074	1.041	0.918
25	0.836	0.829	0.748	0.797
26	0.793	0.768	0.684	0.837
27	0.851	0.842	0.776	0.890

Notes: [a]Observed activity [9]. [b]Calculated from Equation (4). [c]Calculated from Equation (5). [d]Calculated from Equation (6).

group for the free radical scavenging activity. The negative coefficient of S_5 indicates that compounds with high values of the parameter (such as **13**, **33**–**35**) have low activity values. The presence of a methyl group at R'' position increases the value of S_5. Compounds having values of the parameter in the moderate range (such as **1**, **3**, **4**, **17**–**20**) have good free radical scavenging activity. In the above equation, *n*-octanol/water partition coefficient ($\log P$) has a positive regression coefficient. Compounds such as **9**, **13**, **21** and **33** with $\log P$ values in the lower range show poor activity. On the contrary, compounds with high $\log P$ values (**14**–**16, 34**–**36**) show poor free radical scavenging activity due to high values of the S_6 parameter for the corresponding compounds. Compounds (**3**, **4**, **18**, **31** and **32**) with moderate values of the parameter ($\log P$) show good activities. The number of methylene fragments in the side chain (n) and nature of R''' substituents have an influence on the hydrophobicity values of the compounds.

3.2 GFA

Variables involved in the best GFA-derived models are shown in the following equations and comparison of statistical qualities of the models is shown in Table 4. Equations (5) and (6) are among the best ones obtained from the GFA (5000 iterations) method using linear and linear-spline terms, respectively:

$$\begin{aligned}\log(1/K_{\text{app}}) &= 30.809(\pm1.148) - 1.49803(\pm0.070)S_7 \\ &\quad - 5.710(\pm0.316)S_8 - 0.617(\pm0.037)S_1 \\ &\quad + 0.126(\pm0.015)\log P,\end{aligned}$$

$$n_{\text{Training}} = 27, \quad \text{LOF} = 0.005, \quad R^2 = 0.973,$$

$$R_a^2 = 0.968, \quad F = 198.91(\text{df}\,4, 22), \tag{5}$$

$$Q^2 = 0.957, \quad \text{PRESS} = 0.113, \quad n_{\text{Test}} = 9,$$

$$R^2_{\text{pred}} = 0.856, \quad r^2_{m\,(\text{test})} = 0.724,$$

$$r^2_{m\,(\text{LOO})} = 0.893, \quad r^2_{m\,(\text{overall})} = 0.862.$$

Equation (5) consists of four terms (S_7, S_8, S_1 and $\log P$). According to the standardised regression coefficients, the relative importance of the descriptors is in the following order: $S_7 > S_8 > S_1 > \log P$. Among the above terms, E-state parameters of atoms 7 (S_7), 8 (S_8) and 1 (S_1) have negative contributions and $\log P$ has a positive coefficient. The negative coefficient of S_8 indicates that compounds with lower S_8 values (such as **1**, **3**, **4** and **29**) have good activity than those with higher S_8 values (such as **14**, **33**–**36**). The S_8 parameter indicates the importance of the anilino nitrogen atom of the hydroxyphenylureas for the free radical scavenging activity. The negative contribution of the parameter S_1 indicates that a higher value E-state index at atom 1 is detrimental for the activity. Compounds with low values of the parameter (such as **29**–**32**) are more active than compounds with the corresponding high values (**9**, **10**, **12** and **28**). The E-state value at atom 1 is influenced by the nature of the R substituent. Compounds **1**, **3**, **4** and **17**–**20** show good free radical scavenging activity in spite of high S_1 values due to low values of the parameter S_7. The standard errors of the regression coefficients are given within parentheses. Equation (5) could explain and predict 96.8 and 95.7% of the variance, respectively. The external validation statistics of Equation (5) are very good. The predictive R^2 value for the test set was 0.856 and the r_m^2 values for the test, training and overall sets were found to be 0.724, 0.893 and 0.862, respectively:

$$\begin{aligned}\log(1/K_{\text{app}}) &= -6.445(\pm0.569) + 1.433(\pm0.063) \\ &\quad \times \langle 9.68473 - S_7 \rangle - 11.912(\pm0.904)S_9 \\ &\quad + 0.461(\pm0.047)S_3 - 4.926(\pm0.683) \\ &\quad \times \langle 0.748479 - S_{20} \rangle,\end{aligned} \tag{6}$$

$$n_{\text{Training}} = 27, \quad \text{LOF} = 0.005, \quad R^2 = 0.972, \quad R_a^2 = 0.967,$$

$$\text{F} = 191.26(\text{df}\,4, 22), \quad Q^2 = 0.957, \quad \text{PRESS} = 0.113,$$

$$n_{\text{Test}} = 9, \quad R^2_{\text{pred}} = 0.966, \quad r^2_{m\,(\text{test})} = 0.921,$$

$$r^2_{m\,(\text{LOO})} = 0.895, \quad r^2_{m\,(\text{overall})} = 0.914.$$

The relative order of importance of the descriptors is as follows: $\langle 9.68473 - S_7 \rangle > S_9 > S_3 > \langle 0.748479 - S_{20} \rangle$. Equation (6) could explain and predict 96.7 and 95.7%, respectively, of the variance. The external validation statistics of Equation (6) are very good ($R^2_{\text{pred}} = 0.966$ and $r^2_{m\,(\text{test})} = 0.921$, $r^2_{m\,(\text{LOO})} = 0.895$ and $r^2_{m\,(\text{overall})} = 0.914$).

Table 4. Statistical comparison of different models.

Type of statistical methods	R^2	Q^2	R^2_{pred}	$r^2_{m\,(\text{test})}$	$r^2_{m\,(\text{LOO})}$	$r^2_{m\,(\text{overall})}$
Stepwise regression	0.972	0.953	0.916	0.912	0.885	0.896
GFA linear	0.973	**0.957**	0.856	0.724	0.893	0.862
GFA spline	0.972	**0.957**	**0.966**	0.921	0.895	**0.914**

Note: The best values of different parameters are shown in bold.

There is no outlier for Equation (6), i.e. the residuals for all compounds are less than twice the standard error of the estimate of the equation.

In Equations (4) and (5), the term S_7 has negative regression coefficients. The positive coefficient of $\langle 9.68473 - S_7 \rangle$ indicates that the E-state index of atom 7 (S_7) is conducive for the binding affinity when the value is less than 9.68473. For the optimum activity, the value S_7 should be less than 9.68473 (in the case of compounds **1**, **3**, **4**, **17**–**20**). It is evident from the above results that compounds (e.g. **1**, **3**, **4**, **17**–**20**) without any substitution at R and R' positions have higher free radical scavenging activity. The results also indicate that a $-CH_3$ substitution at R position (e.g. compounds **14**–**16**) as well as a $-OC_2H_5$ substitution (e.g. compounds **10**, **12** and **28**) at R' position reduces the activity.

The negative coefficient of S_9 indicates that the activity decreases with an increase in the E-state values of atom 9. Compounds with high values of S_9 (e.g. **13**, **33**–**35**) show poor activity. The term S_9 indicates the importance of carbonyl carbon of hydroxyphenylureas for the free radical scavenging activity. The positive coefficient of S_3 indicates that activity increases with an increase in the E-state value of atom 3.

S_{20} is the E-state index of atom number 20. The negative coefficient of the term $\langle 0.748479 - S_{20} \rangle$ indicates that S_{20} has a negative impact when the corresponding value is more than 0.748479. For example, compounds **12**, **28** and **36** having S_{20} values of 0.696, 0.696 and 0.710, respectively, have poor activity. Compound **1** with the highest value of S_{20} shows better activity. The term S_{20} indicates the importance of the piperazine nucleus in the side chain of hydroxyphenylureas towards the free radical scavenging activity.

Equations (5) and (6) were subjected to process and model randomisation tests and the results are shown in Table 5. The results suggest that Equations (5) and (6) are robust and not obtained by chance. Based on the $r^2_{m\,(\text{overall})}$ criterion, which considers both internal and external validations, the GFA model with spline options (Equation (6)) appears to be the best model ($r^2_{m\,(\text{overall})} = 0.914$). Deeb et al. [36] reported QSAR models on the same data-set; however, the final model developed by them contains eight descriptors, while Equation (6) developed by us contains only four descriptors. The statistical quality of the model containing four parameters developed by Deeb et al. [36] is inferior to that of Equation (6). The novelty of the present work is that, using only a limited number of descriptors (E-state values along with $\log P$ and MR), a statistically superior and more robust model has been derived in this study in comparison to the models developed by Deeb et al. [36] using a huge array of descriptors. Moreover, interpretation of the model developed in the present study is simple and straightforward as the model directly indicates the important fragments/sites important for the mediation of the activity. The method of calculation of the descriptors used in this work is easy and not computationally demanding unlike quantum chemical descriptors. The adopted approach of using E-state parameters also does not require alignment of molecules and conformational analysis unlike 3D-QSAR. All these issues make the developed model reproducible and easily usable by other workers for the development of novel antioxidant molecules.

4. Conclusion

The whole data-set ($n = 36$) was divided into a training set (27 compounds) and a test set (9 compounds) based on k-means clustering of the standardised descriptor matrix and models were developed from the training set. The predictive ability of the models was judged from the prediction of the activity of the test-set compounds. All the developed models indicate an important contribution of the phenolic hydroxy group to the free radical scavenging activity. Hydrophobicity contributes positively to the free radical scavenging activity. The number of methylene fragments in the side chain (n) and nature of R''' substituents have an influence on the hydrophobicity values of the compounds. Compounds without any substitution at R, R' and R'' positions show better activity. Specifically, it was also observed that $-CH_3$ and $-OC_2H_5$ at R and R' positions, respectively, decrease the activity. Furthermore, the results indicate the importance of the urea moiety and piperazine nucleus in the side chain for the free radical scavenging activity. A comparison of the statistical quality of different models is given in Table 4.

Table 5. Results of randomisation tests of the model development process and the developed models.

Randomisation type	Process		Model	
Equation number	(5)	(6)	(5)	(6)
Modelling technique	GFA	GFA spline	GFA	GFA spline
R from the non-random model	0.986	0.986	0.986	0.986
Confidence level	90%	90%	99%	99%
Mean value of R from random trials ± standard deviation	0.416 ± 0.153	0.537 ± 0.182	0.377 ± 0.119	0.374 ± 0.122

Statistical qualities of all the developed models are well above the recommended minimum values. In all the models, the differences between R^2 and Q^2 are not very high (less than 0.3) [34]. Based on the $r^2_{m\,(\text{overall})}$ criterion, which considers both internal and external validations, the GFA model with spline options (Equation (6)) appears to be the best model ($r^2_{m\,(\text{overall})} = 0.914$).

Finally, we can conclude that, apart from the presence of the phenolic hydroxyl moiety and hydrophobicity of the compounds, appropriate substitutions to maintain optimal electron density distribution over the benzene nucleus, urea moiety and piperazine nucleus are required for good free radical scavenging activity of the hydroxyphenylureas.

Acknowledgement

C.S.G. and K.R. thankfully acknowledge the AICTE, New Delhi for the financial assistance.

Notes

1. URL: http://sites.google.com/site/kunalroyindia/.
2. SRETSA is a statistical software in Visual Basic, developed by S. Ray and R. Biswas and standardised using known data-sets.
3. MINITAB is a statistical software of Minitab Inc., USA. Available at http://www.minitab.com.
4. Cerius2 version 4.8 is a product of Accelrys, Inc., San Diego, USA. Available at http://www.accelrys.com/cerius2.

References

[1] B. Halliwell and J.M.C. Gutteridge, *Free Radicals in Biology and Medicine*, Oxford University Press, Oxford, 1989, pp. 185–205.
[2] J.L. Marx, *Oxygen free radicals linked to many diseases*, Science 235 (1987), pp. 529–531.
[3] B. Halliwell, *Free radicals, oxygen toxicity and ageing*, in *Age Pigment*, R.S. Sohal, ed., Elsevier/North Holland, New York, 1981, pp. 1–19.
[4] C.E. Thomas, *Approach and rationale for the design of synthetic antioxidants as therapeutic agents*, in *Handbook of Synthetic Antioxidants*, L. Packer and E. Cadenas, eds., Marcel Dekker, New York, 1997, pp. 1–52.
[5] G. Scott, *Antioxidants*, Bull. Chem. Soc. Jpn 61 (1988), pp. 165–170.
[6] A. Kampkotter, C.G. Nkwonkam, R.F. Zurawski, C. Timpel, Y. Chovolou, W. Watjen, and R. Kahl, *Investigations of protective effects of the flavonoids quercetin and rutin on stress resistance in the model organism* Caenorhabditis elegans, Toxicology 234 (2007), pp. 113–123.
[7] K. Nakao, R. Shimizu, H. Kubota, M. Yasuhara, Y. Hashimura, T. Suzuki, T. Fujita, and H. Ohmizu, *Quantitative structure–activity analyses of novel hydroxyphenylurea derivatives as antioxidants*, Bioorg. Med. Chem. 6 (1998), pp. 849–868.
[8] K. Nakao, H. Kubota, M. Yasuhara, K. Saito, T. Suzuki, H. Ohmizu, and R. Shimizu, *Novel hydroxyphenylurea dual inhibitor against acyl-CoA: Cholesterol acyltransferase (ACAT) and low density lipoprotein (LDL) oxidation as antiatherosclerotic agent*, Bioorg. Med. Chem. 9 (2001), pp. 853–861.
[9] K. Youssef, M. El-Sherbeny, F. Elshafire, O. Al-Deeb, H. Farag, and S. Awadalla, *New developments in antioxidant research*, in *Novel Hydroxyphenylurea as Antioxidant and Chemopreventive Agents*, H. Panglossi, ed., Nova Science Publishers, New York, 2006, pp. 61–79.
[10] A.K. Ghose and G.M. Crippen, *Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure–activity relationships. 2. Modelling dispersive and hydrophobic interactions*, J. Chem. Inf. Comput. Sci. 27 (1987), pp. 21–35.
[11] L.H. Hall, B. Mohney, and L.B. Kier, *The electrotopological state: An atom index for QSAR*, Quant. Struct-Act. Relat. 10 (1991), pp. 43–51.
[12] L.B. Kier and L.H. Hall, *An electrotopological state index for atoms in molecules*, Pharm. Res. 7 (1990), pp. 801–807.
[13] L.H. Hall, B. Mohney, and L.B. Kier, *Comparison of electrotopological state indexes with molecular orbital parameters: Inhibition of MAO by hydrazides*, Quant. Struct-Act. Relat. 12 (1993), pp. 44–48.
[14] L.B. Kier and L.H. Hall, *Advances in Drug Research*, Vol. 22, Academic Press, New York, 1992.
[15] L.B. Kier and L.H. Hall, *Molecular Structure Description: The Electrotopological State*, Academic Press, San Diego, CA, 1999.
[16] K. Roy, A.U. De, and C. Sengupta, *QSAR with electrotopological state atom index: Antiadrenergic activity of N,N-dimethyl-2-bromo-2-phenylethylamines*, Indian J. Chem. 38B (1999), pp. 942–947.
[17] K. Roy, D.K. Pal, and C. Sengupta, *QSAR of antineoplastics V: Exploration of receptor interaction sites of antitumor N-(7-indolyl)benzenesulfonamides targeting G1 phase using electrotopological state atom index*, Drug Des. Discov. 17 (2001), pp. 207–218.
[18] K. Roy, A.U. De, and C. Sengupta, *QSAR of human factor Xa inhibitor N^2-aroylanthranilamides using principal component factor analysis*, Drug Des. Discov. 18 (2002), pp. 23–31.
[19] K. Roy, S. Chakraborty, and A. Saha, *Exploring selectivity requirements for COX-2 vs. COX-1 binding of 3,4-diaryloxazolones using E-state index*, Bioorg. Med. Chem. Lett. 13 (2003), pp. 3753–3757.
[20] K. Roy and J.T. Leonard, *QSAR modeling of HIV-1 reverse transcriptase inhibitor 2-amino-6-arylsulfonylbenzonitriles and congeners using molecular connectivity and E-state parameters*, Bioorg. Med. Chem. 12 (2004), pp. 745–754.
[21] C. Sengupta, J.T. Leonard, and K. Roy, *Exploring QSAR of melatonin receptor ligand benzofuran derivatives using E-state index*, Bioorg. Med. Chem. Lett. 14 (2004), pp. 3435–3439.
[22] S. Ray, C. Sengupta, and K. Roy, *QSAR modeling of antiradical and antioxidant activities of flavonoids using electrotopological state (E-State) atom parameters*, Cent. Eur. J. Chem. 5 (2007), pp. 1094–1113.
[23] S. Ray, C. Sengupta, and K. Roy, *QSAR modeling for lipid peroxidation inhibition potential of flavonoids using topological and structural parameters*, Cent. Eur. J. Chem. 6 (2008), pp. 267–276.
[24] J.T. Leonard and K. Roy, *On selection of training and test sets for the development of predictive QSAR models*, QSAR Comb. Sci. 25 (2006), pp. 235–251.
[25] R.B. Darlington, *Regression and Linear Models*, McGraw Hill, New York, 1990.
[26] Y.L. Fan, M. Shi, K.W. Kohn, Y. Pommier, and J.N. Weinstein, *Quantitative structure–antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies*, J. Med. Chem. 44 (2001), pp. 3254–3263.
[27] D. Rogers and A.J. Hopfinger, *Application of genetic function approximation to quantitative structure–activity relationships and quantitative-property relationships*, J. Chem. Inf. Comput. Sci. 34 (1994), pp. 854–866.
[28] G.W. Snedecor and W.G. Cochran, *Statistical Methods*, Oxford & IBH Publishing Co. Pvt. Ltd, New Delhi, 1967.
[29] S. Wold and L. Eriksson, *Multivariate linear modeling,* in *Chemometric Methods in Molecular Design*, H. van de Waterbeemd, ed., VCH, Weinheim, Germany, 1995, pp. 312–317.
[30] A.K. Debnath, *Quantitative structure–activity relationship (QSAR): A versatile tool in drug design*, in *Combinatorial Library Design and Evaluation: Principles, Software Tools, and Applications in Drug Discovery*, A.K. Ghose and V.N. Viswanadhan, eds., Marcel Dekker, New York, 2001, pp. 73–129.
[31] K. Roy, *On some aspects of validation of predictive QSAR models*, Expert Opin. Drug Discov. 2 (2007), pp. 1567–1577.

[32] P.P. Roy and K. Roy, *On some aspects of variable selection for partial least squares regression models*, QSAR Comb. Sci. 27 (2008), pp. 302–313.
[33] P.P. Roy, S. Paul, I. Mitra, and K. Roy, *On two novel parameters for validation of predictive QSAR models*, Molecules 14 (2009), pp. 1660–1701.
[34] P.P. Roy, J.T. Leonard, and K. Roy, *Exploring the impact of the size of training sets for the development of predictive QSAR models*, Chemom. Intell. Lab. Syst. 90 (2008), pp. 31–42.
[35] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, and P. Gramatica, *Methods for reliability and uncertainty assessment and for applicability evaluations of classifications- and regression-based QSARs*, Environ. Health Perspect 111 (2003), pp. 1361–1375.
[36] O. Deeb, K.M. Youssel, and B. Hemmateenejad, *QSAR of novel hydroxyphenylureas as antioxidant agents*, QSAR Comb. Sci. 27 (2008), pp. 417–424.