Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 16 (2008) 5720–5732
Bioorganic & Medicinal Chemistry

Halogenated derivatives QSAR model using spectral moments to predict haloacetic acids (HAA) mutagenicity

Alfonso Pérez-Garrido,[a,*] Maykel Pérez González[b,c,✠] and Amalio Garrido Escudero[a,*]

[a]*Environmental Engineering and Toxicology Department, Catholic University of San Antonio, Guadalupe, Murcia, C.P. 30107, Spain*
[b]*Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, Villa Clara, C.P. 54830, Cuba*
[c]*Department of Organic Chemistry, Vigo University, C.P. 36200, Vigo, Spain*

Received 19 November 2007; revised 29 February 2008; accepted 25 March 2008
Available online 30 March 2008

Abstract—The risk of the presence of haloacetic acids in drinking water as chlorination by-products and the shortage of experimental mutagenicity data for most of them requires a research work. This paper describes a QSAR model to predict direct mutagenicity for these chemicals. The model, able to describe more than 90% of the variance in the experimental activity, was developed with the use of the spectral moment descriptors. The model, using these descriptors with multiplicative effects provides better results than other linear descriptors models based on Geometrical, RDF, WHIM, eigenvalue-based indices, 2D-autocorrelation ones, and information descriptors, taking into account the statistical parameters of the model and the cross-validation results. The structural alerts and the mutagenicity-predicted values from the model output are in agreement with references from other authors. The mutagenicity predicted values for the three haloacetic acids, which have available experimental data (TCAA—Trichloroacetic acid, BDCAA—Bromodichloroacetic acid, and TBAA—Tribromoacetic acid), are reasonably close to their experimental values, specially for the latest two.

1. Introduction

The introduction of water disinfection processes was a significant success in the control of waterborne diseases.[1] Drinking water disinfection is required to remove harmful pollutants, including pathogenic microorganisms. Some studies demonstrated that concentrated extracts of disinfected drinking water were toxic in many in vivo and in vitro bioassays.[2] Most drinking water disinfection by-products form as a result of the reaction between organic matter in raw water and chemical disinfectants like chlorine. These organic compounds come from two major sources: (1) breakdown products of naturally occurring materials (NOM), which include humic acids, microorganisms and their metabolites, and some petroleum-based high molecular weight aliphatic and aromatic hydrocarbons; and (2) products from domestic and commercial activities, including agricultural and urban runoff and wastewater discharges. Drinking water disinfection by-products (DBPs) represent an important class of environmentally hazardous chemicals. They can increase the risk for human health in long-term basis. Epidemiological studies demonstrate that individuals who consume chlorinated drinking water are exposed to a higher risk of developing a cancer of stomach, pancreas, kidney, bladder, and rectum as well as Hodgkins and non-Hodgkins lymphoma.[3–5] The issues of human exposure to DBPs for epidemiological and health risk assessment were recently reviewed.[6]

The haloacetic acids are the second greater group of drinking water disinfection by-products and some of them are rodent liver carcinogens[7–9] and mutagenic in *Salmonella typhimurium*.[10–15] There is a shortage of information about carcinogenic and mutagenic potency for these chemicals. The use of methods which are able to predict such values are important for toxicological risk assessment. This is the reason for our having developed this QSAR model in order to predict the mutagenicity in *S. typhimurium* strain TA100. The mutagenic-

Keywords: QSAR; Spectral moments; Haloacetic acids; Mutagenicity.
* Corresponding authors. Tel.: +34 968 278 755 (A.P-G.); +34 968 278 544 (A.G.E.); e-mail addresses: Aperez@pdi.ucam.edu; AGescudero@pdi.ucam.edu
✠ Deceased.

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.03.070

ity in *S. typhimurium* as determined by the Ames test is used world-wide for initial screening to determine the mutagenic potential of new chemicals and drugs. It is know that there is a high predictive value for rodent carcinogenicity when a mutagenic response is obtained.[16–18]

A Ref. 19 has been found describing a QSAR model to infer the mechanism of action of mono-, di-, and tri-halogenated acids by relating the 1- and 2-atom fragment contributions towards the total toxicity as predicted by TOPKAT to certain types of descriptor. The theory being that there is a good correlation between the two is a reasonable indication of a particular mechanism of action.

There are more than 3200 molecular descriptors that can be used to solve the problem outlined above.[20] A useful kind of descriptors in Medicinal Chemistry have been introduced some years ago.[21–27] The descriptors based on spectral moments are a good example of these ones. Recently, the MARCH-INSIDE (MARkovian CHemicals IN SIlico Design) descriptors were introduced.[28–30] This kind of descriptors have proved to be very useful in studies with proteins,[31,32] anticancer compounds[33], and antimicrobials.[34] The other type of descriptors based on spectral moments are named TOPS-MODE (TOPological Sub-structural MOlecular DEsign) descriptors[35–37] and they are the spectral moments of the bond matrix weighted in the main diagonal with different physicochemical parameters. These descriptors are easy to calculate and they can be used when there are a heterogeneous series of compounds.[22,38]

The successful application of this theoretical approach to the modeling of toxicological[39–41] and ecotoxicological[42,43] properties has also inspired us to perform a more exhaustive study.The intention was to test and validate TOPS-MODE applicability in assessing discovery of leads and mutagenic impact of chemicals. The selection of a data set on mutagenic toxicity is not random.

Descriptors based on spectral moments show a clear interest in new QSAR model research in bioorganic and medicinal chemistry fields. We focus our research work to develop a QSAR model based on spectral moments because of the advantages of easy use and understanding.

2. Materials and methods

2.1. Data set

A data set of 42 halogenated derivatives was collected from literature and U.S. National Toxicological Program (NTP) (Table 1). The updated NTP database is also available at the following web site: http:// ntp-apps.niehs.nih.gov/ ntp_tox/index.cfm. In this data set are included nitrohaloalkanes, haloacids, haloaldehides, halocetones, haloalcohols, haloepoxides, and haloalkanes and the most of them are well-known alkylating agents. A big group of such compounds is present in drinking water as disinfection by-products.

The activity is define as the logarithm of TA100 strain *S. typhimurium* Ames test[44] without activation and with preincubation. It is calculated as the slope of the linear portion of the dose–response curve.[45] The data set includes the values from Plewa et al.[13,15] The preincubation protocol used by these authors (1 h instead of 20 min) is different from that of the NTP protocol. The values from the previous by mentioned authors were adjusted using a correction factor calculated as the ratio between the NTP values (tested following a preincubation period of 20 min) and the Plewa et al. values.

2.2. The TOPS-MODE descriptors

The TOPS-MODE descriptors are based on the calculation of the spectral moments of the so-called bond matrix.[46] The theoretical basis has been described in previous reports.[35,36] Nevertheless, an overview of this descriptor family is going to be given below. The bond matrix is defined as a square and symmetric matrix whose entries are ones or zeros if the corresponding bonds are adjacent or not. The order of this matrix (m) is the number of bonds in the molecular graph, being two bonds adjacent if they are incident to a common atom. The spectral moments of the edge adjacency matrix are defined as the traces. That is the sum of the main diagonal of the different powers of such matrix. Several bond weights such as standard bond distance (Std), standard bond dipole moments (Dip, Dip2), hydrophobicity (H), polar surface area (Pols), polarizability (Pol), molar refractivity (Mol), van der Waals radii (vdW), and Gasteiger–Marsilli charges (Gas) were used for computing the spectral moments of the bond matrix. Since most of the approaches for computing physicochemical properties from fragment are based on atom-additive methods, several transform from atomic to bond contributions were carried out. The way in which these atomic contributions were transformed into bond contributions has been described by Estrada et al.[47]:

$$w(i,j) = \frac{w_i}{\delta_i} + \frac{w_j}{\delta_j} \quad (1)$$

where w_i and δ_i are the atomic weight and vertex degree of the atom i. The calculation of the TOPS-MODE descriptors was carried out with the software MODESLAB 1.0.[48] The input of the software consists of SMILES codes for each compound.[49] We calculated the first 15 spectral moments (μ_1–μ_{15}) for each bond weight and the number of bonds in the molecules (μ_0). Also, we multiplied μ_0 and μ_1 for the first 15 spectral moments obtaining 30 new variables. These variables include very valuable information due to the nonlinear behaviour of the biological process.[26] To apply the current approach to the structure–toxicity relationship, the following steps should be followed: first, to select an adequate training set according to the aim and scope of the research. Second, to draw the molecular graphs for each molecule included in the training set. The third step is to differentiate the molecular bonds with appropriate weights. The fourth one is to compute the spectral moments of the bond matrix for each molecule of the

Table 1. Names, CAS number, mutagenic potency, and reference to the compounds used in this study

Compounds	Name	CAS	Log TA100	Reference
1	2,4,4-Trichloro-3-(dichloromethyl)-2-butenoic acid	97055-37-3	2.821	95
2	(*Z*)-2,4,4-Trichloro-3-formyl-2-butenoic acid	117823-31-1	4.070	95
3	(*Z*)-2,4-Dichloro-3-formyl-2-butenoic acid	—	2.847	95
4	(*Z*)-2-Chloro-3-methyl-4-oxo-2-butenoic acid	—	−0.319	95
5	(*E*)-2,4,4-Trichloro-3-(chloromethyl)-2-butenoic acid	—	3.466	95
6	(*S*)-2,3-Dibromopropanal	5221-17-0	−0.522	96
7	Dichloroacetic acid	79-43-6	−1.829	13
8	Chloroacetic acid	79-11-8	−1.943	13
9	Bromoacetic acid	79-08-3	0.363	13
10	2-Bromopropane	75-26-3	−1.716	NTP
11	2,3-Dichloro-1-propene	78-88-6	−0.0065	NTP
12	(*R*)-1,2-Dichloropropane	78-87-5	−2.473	NTP
13	2,2-Dichloroacetyl Chloride	79-36-7	−1.212	NTP
14	1,3-Dichloro-2-propanol	96-23-1	−1.837	NTP
15	(*S*)-2,3-Dibromo-1-propanol	96-13-9	−0.334	NTP
16	1,2-Dibromoethane	106-93-4	−0.607	NTP
17	(*R*)-2-(Chloromethyl)oxirane	106-89-8	−0.692	NTP
18	2-Chloroacetaldehyde	107-20-0	−0.789	NTP
19	2-Chloroethanol	107-07-3	−2.074	NTP
20	3-Chloro-1-propene	107-05-1	−2.326	NTP
21	(*E*)-1,4-Dichloro-2-butene	110-57-6	0.179	NTP
22	(2*S*,3*S*,4*S*,5*S*)-1,6-Dibromohexane-2,3,4,5-tetraol	488-41-5	−1.607	NTP
23	(*R*)-2-(Fluoromethyl)oxirane	503-09-3	−1.092	NTP
24	2-Bromoethanol	540-51-2	−2.075	NTP
25	(*E*)-1,3-Dichloro-1-propene	542-75-6	0.010	NTP
26	(*S*)-3-Iodo-1,2-propanodiol	554-10-9	−1.260	NTP
27	2-Chloro-2-nitropropane	594-71-8	−1.874	NTP
28	(*R*)-1-Chloro-1-nitropropane	600-25-9	−0.679	NTP
29	(*S*)-2,3-Dichloro-1-propanol	616-23-9	−0.746	NTP
30	3-Bromo-1-propanol	627-18-9	−1.694	NTP
31	Dibromoacetic acid	631-64-1	−1.203	NTP
32	2-Chloroethyl acrylate	2206-89-5	−1.214	NTP
33	(*S*)-2-(2,2,2-Trichloroethyl)oxirane	3083-25-8	−1.193	NTP
34	(*S*)-2-(Trichloromethyl)oxirane	3083-23-6	0.455	NTP
35	(*R*)-2-(Bromomethyl)oxirane	3132-64-7	−0.371	NTP
36	3-Chloro-*N*,*N*-dimethyl-1-propanamine	5407-04-5	−2.400	NTP
37	Bromochloroacetic acid	5589-96-8	−1.207	NTP
38	(*S*)-2,3-Dibromopropyl acrylate	19660-16-3	−1.550	NTP
39	(*R*)-1-Bromo-2-propanol	19686-73-8	−1.793	NTP
40	2-Bromoacrylaldehyde	14925-39-4	−0.207	97
41	3-Bromo-3-buten-2-one	61203-01-8	−0.886	97
42	Iodoacetic acid	64-69-7	0.776	15

data set. The fifth step is to find a quantitative structure–toxicity relationship by using a regression analysis:

$$P = a_0\mu_0 + a_1\mu_1 + a_2\mu_2 + \ldots + a_k\mu_k + b \qquad (2)$$

where P is the studied activity, in our case, the log TA100 partitioning, μ_k is the k-th spectral moment, and the a_k are the coefficients obtained by linear regression. The sixth step is to test the predictive capability of the regression model by cross-validation procedures and an external prediction set. And finally, to compute the contribution of the different substructures to determine their quantitative contribution to the mutagenicity of the studied molecules.

2.3. Structural alerts identification

The identification of structural alerts (fragment contribution) to the toxicity is based on bond contributions. This procedure, implemented in MODESLAB software, consists in transforming a QSAR model into a bond additive scheme. As a result we calculate for each molecule the toxicological property as a sum of bond contributions. Bond contributions are derived from the local spectral moments. They are defined as the diagonal entries of the different powers of the weighted E matrix.

$$\mu_k^T(i) = b_{ii}(T)^k \qquad (3)$$

where $\mu_k^T(i)$ is the k-th local spectral moment of the bond i, $b_{ii}(T)$ are the diagonal entries of the weighted E matrix and T is the type of bond weight.

For a given molecule, we can substitute the values of the local spectral moments computed by Eq. 3 into Eq. 4 and thus gather the total contribution to the toxicity of its different bonds.

$$P = b_0 + \sum_k a_k \cdot \mu_k^T \qquad (4)$$

Since the activity modeled is expressed as log TA100, positive bond contributions increase the TA100 value and increase the mutagenic activity and vice versa. The structural information highlighted by the bond contri-

butions may allow, together with other theoretical and experimental data, a better understanding of the mechanisms of mutagenic action of the involved chemicals. Also, it can be useful for the proposal of new metabolic routes associated with the mutagenesis phenomenon.

2.4. Computational strategies

Calculation of spectral moments was carried out using Modeslab 1.0 software[48] taking the Simplified Molecular Imput Line Entry Specification (SMILES) format[49] of the geometrically optimized compound structure by Cosmic module of Tsar 3.3 software (Accelryc Inc., http://www.accelrys.com). The other family of descriptors like Geometrical (74 descriptors), RDF (150 descriptors), WHIM (99 descriptors), eigenvalue-based indices (44 descriptors), and 2D-autocorrelation (96 descriptors) was calculated used Dragon Web.[50] The variables with constants or close to constants values were deleted. The mathematical models were obtained by means of Multiple Regression Analysis (MRA) as implemented in the Tsar 3.3 software. The variables to be included in the equation were selected using forward stepwise procedure as variable selection strategy.[51]

2.5. Model selection and validation

The statistical significance of the models was determined by examining the squared regression coefficient (R^2), the standard deviation (s), the Fisher ratio (F) and the ratio between the number of cases and the number of adjustable parameters in the model ρ statistic which we assume as the criterion $\rho \geqslant 4$.[52]

In addition, further criteria exist to compare the quality of the models obtained. One of them uses the correlation coefficient R which is barely meaningful because it tends to select as many variables as possible as well as the standard deviation s. The other criterion is the Kubinyi function (FIT), being closely related to the F value, which was created and proved to be useful.[53,54] The best model will be the one that exhibits the high value of this function. The other of these criteria was formulated by Akaike sometime ago.[55,56] Akaike's information criteria (AIC) take into account the statistical goodness of fit and the number of parameters that have to be estimated to achieve that degree of fit. The model that produces the minimum value of these statistics should be considered potentially the most useful. The outliers detection was carried out for the compounds that have large residual.

The robustness of the models and their predictivity were evaluated by Q^2 *leave-one-out* (LOO) cross-validation, an equivalent statistic to R^2, and bootstrapping test (Q^2_{boot}). The stability when a heavy perturbation in the training set is applied was checked by response randomization (Y-scrambling) ($a(R^2)$ and $a(Q^2)$) procedures. These calculations were carried out with the software Mobydigs Computer Software 1.0.[57] To sum up, a good quality of the models was indicated by high values in F, FIT, and ρ, lower values in AIC and s, as well as close to one values in R^2, Q^2, and Q^2_{boot}.

2.6. Orthogonalization of descriptors

The main drawback of collinearity from the point of view of a QSAR model is about the stability of the coefficients in the linear regression model. In the case of the TOPS-MODE descriptors this can be translated into false interpretation of bond contributions. The magnitude and sign of them can be falsified by the effect produced by the existence of collinear variables in the model. We employed the Randic' method of orthogonalization which has been described in detail in several papers.[58–62] Thus, we will give only a general overview here.

The first step for orthogonalizing the molecular descriptors is to select the appropriate order of orthogonalization, which, in this case, is the order of significance of the variables in the model. The first variable (v_1) is taken as the first orthogonal descriptors $\Omega\ v_1$ and the second one is orthogonalized respect to it by taking the residual of its correlation with Ωv_1. The process is repeated until all variables are completely orthogonalized and the orthogonal variables are then used to obtain the new model. For the extraction of the information contained in the orthogonalized descriptors we followed the procedure reported by Estrada et al.[22]

2.7. Applicability domain of the models

Once we obtained the model we defined its applicability domain to make predictions for the rest of haloacetic acids. There are several methods for assessing the applicability domain (AD) of QSAR models[63] but the most common one encompasses determining the leverage values for each compound.[64] To visualize the AD of a QSAR model, the plot of standardized residuals versus leverage values (h) (the Williams plot) can be used for an immediate and simple graphical detection of both the response outliers (i.e., compounds with standardized residuals greater than two standard deviation units) and structurally influential chemicals in a model ($h > h^*$). These calculations were carried out with the software Mobydigs Computer Software 1.0.[57] Figure 8 shows the Williams plot, i.e., for each compound of the training set. From this plot, the applicability domain is established inside a squared area within ±2 standard deviations and a leverage threshold h^* of 0.68 ($h^* = 3\kappa/n$, being κ the number of model parameters and n the number of objects). For making predictions, predicted mutagenicity data must be considered reliable only for those chemicals that fall within the applicability domain on which the model was constructed.[65]

3. Results and discussion

3.1. QSAR model

The best QSAR model obtained with the spectral moments is given as follows together with the statistical parameters of the regression.

$$\begin{aligned}\log TA100 = {} & 8.052\cdot10^{-11}(\pm1.54\cdot10^{-11})\mu_{15}^{\mathrm{Std}} \\ & - 5.302\cdot10^{-2}(\pm1.12\cdot10^{-2})\mu_{1}^{\mathrm{Pols}} \\ & + 5.723\cdot10^{-8}(\pm1.09\cdot10^{-8})\mu_{7}^{\mathrm{Mol}} \\ & - 7.991\cdot10^{-10}(\pm1.26\cdot10^{-10})\mu_{0}\mu_{13}^{\mathrm{Dip}} \\ & + 2.152\cdot10^{-8}(\pm4.39\cdot10^{-9})\mu_{0}\mu_{5}^{\mathrm{Pols}} \\ & + 1.599\cdot10^{-2}(\pm1.59\cdot10^{-3})\mu_{1}\mu_{2}^{\mathrm{Dip2}} \\ & - 3.331\cdot10^{-10}(\pm1.01\cdot10^{-10})\mu_{1}\mu_{14}^{\mathrm{Dip2}} \\ & - 2.208\cdot10^{-8}(\pm4.66\cdot10^{-9})\mu_{1}\mu_{9}^{\mathrm{Pol}} \\ & - 1.163(\pm0.28) \\ N = 42;\ \ R^2 = 0.842;\ \ & Q^2_{(\mathrm{CV\text{-}LOO})} = 0.684; \\ s = 0.679;\ \ F = 21.98;\ \ & \mathrm{AIC} = 0.659;\ \ \mathrm{FIT} = 1.659 \\ Q^2_{\mathrm{boot}} = 0.639;\ \ & a(r^2) = 0.114;\ \ a(Q^2) = -0.49 \end{aligned} \tag{5}$$

where N is the number of compounds included in the model, R the correlation coefficient, s standard deviation of the regression, F the Fisher ratio, Q^2 the correlation coefficient of the cross-validation, AIC the Akaike Information Criterion and FIT the Kubinyi Function. The values of the descriptors are presented in Table 2.

Although this theoretical model has eight variables and acceptable statistical parameters, a step-by-step outlier extraction procedure led to different models with a better statistical profile. In this study, two outliers extracted represented a 4.76% of the whole data. Compounds **4** ((*Z*)-2-chloro-3-methyl-4-oxo-2-butenoic acid) and **9** (Bromoacetic acid) present large residuals and should be considered as outliers. Compound **4** is structurally similar to compounds **1** (2,4,4-trichloro-3-(dichloromethyl)-2-butenoic acid), **2** ((*Z*)-2,4,4-trichloro-3-formyl-2-butenoic acid), **3** ((*Z*)-2,4-dichloro-3-formyl-2-butenoic acid), and **5** ((*E*)-2,4,4-trichloro-3-(chloromethyl)-2-butenoic acid) of the training set. Compound **4** has a main action mechanism throughout adenosine adduct[66] and for that reason it has a low value of the activity respect its peer butenoic acids which have a mechanism of action through guanosine adduct. Even though it is known that AT sites are the primary targets in studied strains TA98 or TA100, only compounds with mechanism of action throughout guanosine adducts (GC sites) are detected by a TA100 strain mutagenicity Ames test. This fact explains the observed differences between both. The other outlier, compound **9**, and, like Iodoacetic acid but in minor amount, could induce its genotoxic damage via an oxidative stress mechanism[67] unlike the rest of brominated derivatives present in the training set. For the same reason the Iodoacetic acid must be an outlier but together with (S)-3-iodo-1,2-propanodiol is the unique Iodine derivatives presents in this training set. On removal of these compounds from the training set, the next equation is obtained:

$$\begin{aligned}\log TA100 = {} & 8.95\cdot10^{-11}(\pm1.26\cdot10^{-11})\mu_{15}^{\mathrm{Std}} \\ & - 5.77\cdot10^{-2}(\pm9.42\cdot10^{-3})\mu_{1}^{\mathrm{Pols}} \\ & + 5.87\cdot10^{-8}(\pm8.79\cdot10^{-9})\mu_{7}^{\mathrm{Mol}} \\ & - 8.57\cdot10^{-10}(\pm1.03\cdot10^{-10})\mu_{0}\mu_{13}^{\mathrm{Dip}} \\ & + 2.30\cdot10^{-8}(\pm3.61\cdot10^{-9})\mu_{0}\mu_{5}^{\mathrm{Pols}} \\ & + 1.75\cdot10^{-2}(\pm1.33\cdot10^{-3})\mu_{1}\mu_{2}^{\mathrm{Dip2}} \\ & - 4.07\cdot10^{-10}(\pm8.39\cdot10^{-11})\mu_{1}\mu_{14}^{\mathrm{Dip2}} \\ & - 2.26\cdot10^{-8}(\pm3.77\cdot10^{-9})\mu_{1}\mu_{9}^{\mathrm{Pol}} \\ & - 1.24(\pm0.23) \\ N = 40;\ \ R^2 = 0.902;\ \ & Q^2_{(\mathrm{CV\text{-}LOO})} = 0.842; \\ s = 0.548;\ \ F = 35.730;\ \ & \mathrm{AIC} = 0.727;\ \ \mathrm{FIT} = 1.529 \\ Q^2_{\mathrm{boot}} = 0.718;\ \ & a(r^2) = 0.125;\ \ a(Q^2) = -0.392 \end{aligned} \tag{6}$$

We detected high correlation coefficients among the descriptor values of model (Eq. 6). Table 3 shows that some of regression coefficients were higher than 0.70, showing that they were closely correlated. Therefore, orthogonalization of the molecular descriptors was conducted.

Orthogonalization of molecular descriptors was undertaken to avoid collinearity among variables and model overfitting. Collinearity of variables should be as low as possible because interrelatedness among different descriptors can result in highly unstable models. The QSAR model obtained with the spectral moments (Eq. 7) after orthogonalization is given below, together with the statistical parameters of regression analysis.

$$\begin{aligned}\log TA100 = {} & 6.52\cdot10^{-3}(\pm5.64\cdot10^{-4})\Omega\mu_{1}\mu_{2}^{\mathrm{Dip2}} \\ & - 1.80\cdot10^{-10}(\pm2.42\cdot10^{-11})^{2}\Omega\mu_{0}\mu_{13}^{\mathrm{Dip}} \\ & + 1.34\cdot10^{-11}(\pm6.35\cdot10^{-12})^{3}\Omega\mu_{15}^{\mathrm{Std}} \\ & + 7.09\cdot10^{-9}(\pm2.22\cdot10^{-9})^{4}\Omega\mu_{7}^{\mathrm{Mol}} \\ & + 4.81\cdot10^{-9}(\pm1.62\cdot10^{-9})^{5}\Omega\mu_{0}\mu_{5}^{\mathrm{Pols}} \\ & - 4.50\cdot10^{-2}(\pm9.05\cdot10^{-3})^{6}\Omega\mu_{1}^{\mathrm{Pols}} \\ & - 1.85\cdot10^{-8}(\pm3.67\cdot10^{-9})^{7}\Omega\mu_{1}\mu_{9}^{\mathrm{Pol}} \\ & - 4.07\cdot10^{-10}(\pm8.39\cdot10^{-11})^{8}\Omega\mu_{1}\mu_{14}^{\mathrm{Dip2}} \\ & - 1.91(\pm0.13) \\ N = 40;\ \ R^2 = 0.902;\ \ & Q^2_{(\mathrm{CV\text{-}LOO})} = 0.842; \\ s = 0.548;\ \ F = 35.730;\ \ & \mathrm{AIC} = 0.727;\ \ \mathrm{FIT} = 1.529 \\ Q^2_{\mathrm{boot}} = 0.718;\ \ & a(r^2) = 0.125;\ \ a(Q^2) = -0.392 \end{aligned} \tag{7}$$

Once the non-desirable collinearity problems among the descriptors were eliminated the model obtained with the TOPS-MODE descriptors was compared with other families.

3.2. Comparison with other descriptors

The spectral moments were compared with other methodologies such as Geometrical,[20] RDF,[68] WHIM,[20]

Table 2. Values of the spectral moments used in the model

Compounds	Name	Log TA100	μ_{15}^{Std}	μ_{1}^{Pols}	μ_{7}^{Mol}	$\mu_{0}\mu_{13}^{Dip}$	$\mu_{0}\mu_{5}^{Pols}$	$\mu_{1}\mu_{2}^{Dip2}$	$\mu_{1}\mu_{14}^{Dip2}$	$\mu_{1}\mu_{9}^{Pol}$
1	2,4,4-Trichloro-3-(dichloromethyl)-2-butenoic acid	2.821	32399259648	57.53	5319063.5	3938757888	120429368	638.02	8035869184	101569274
2	(*Z*)-2,4,4-Trichloro-3-formyl-2-butenoic acid	4.071	17575616512	74.60	3180590.25	2981445888	131358560	559.84	4532435968	49024272
3	(*Z*)-2,4-Dichloro-3-formyl-2-butenoic acid	2.848	10623868928	74.60	2456188	2577081600	131358560	458.56	2109218685	31900258
4	(*Z*)-2-Chloro-3-methyl-4-oxo-2-butenoic acid	−0.320	7693142016	74.60	1102993.125	2400710656	131358560	363.63	1201073250	29644918
5	(*E*)-2,4,4-Trichloro-3-(chloromethyl)-2-butenoic acid	3.466	24558442496	57.53	4594256	3502516480	120429368	528.86	4903678976	79299192
6	(*S*)-2,3-Dibromopropanal	−0.522	29399392256	17.07	19273382	3796207104	13531792	208.51	3007512320	71086728
7	Dichloroacetic acid	−1.830	9651770368	57.53	2302502	1081926919	60212696	184.65	3086136064	21476786
8	Chloroacetic acid	−1.943	3874263808	57.53	1577735.875	832599616	60212696	129.84	1112014735	11595926
9	Bromoacetic acid	0.364	4498661888	57.53	10937820	795104960	60212696	127.76	1063220478	22664120
10	2-Bromopropane	−1.717	28033912832	0.00	8497154	2682537216	10800	53.98	609960128	95866400
11	2,3-Dichloro-1-propene	−0.007	5745217024	0.00	2748616.5	870004416	3440	83.27	362619648	30664026
12	(*R*)-1,2-Dichloropropane	−2.473	33391790080	0.00	2751213.25	3755125504	10800	118.55	1896276782	68256624
13	2,2-Dichloroacetyl Chloride	−1.212	10644528128	17.07	2722970.75	1619424644	9018675	205.95	4067141888	25971036
14	1,3-Dichloro-2-propanol	−1.838	39579848704	40.46	3183758.75	4337495552	77919208	155.83	2450607360	49380968
15	(*S*)-2,3-Dibromo-1-propanol	−0.334	45910818816	40.46	19350196	4466391040	77915872	150.53	2520851200	89414640
16	1,2-Dibromoethane	−0.608	12928412672	0.00	21928546	1119365902	4620	79.61	941540928	64058428
17	(*R*)-2-(Chloromethyl)oxirane	−0.693	61269241856	9.23	1899091.5	5696471552	143557.9219	118.07	2917894912	31245614
18	2-Chloroacetaldehyde	−0.789	2879185920	17.07	1702408.375	847295168	9018675	100.31	706531968	11804425
19	2-Chloroethanol	−2.074	10087707648	40.46	1808295.375	1123018729	56662730	62.17	578029184	20640296
20	3-Chloro-1-propene	−2.326	3641498880	0.00	2375211.25	625260096	3440	38.46	126195640	24974387
21	(*E*)-1,4-Dichloro-2-butene	0.179	7802283520	0.00	3720415.5	1441410646	8360	118.55	519270304	46837468
22	(2*S*,3*S*,4*S*,5*S*)-1,6-Dibromohexane-2,3,4,5-tetraol	−1.607	1.49E+11	161.84	22287698	24518873088	651642432	485.00	9049610240	198085360
23	(*R*)-2-(Fluoromethyl)oxirane	−1.092	58288287744	9.23	601365.25	5640154112	143557.9219	114.14	2778343680	16835970
24	2-Bromoethanol	−2.076	11155117056	40.46	11206682	1071490027	56662730	60.43	544268672	34284384
25	(*E*)-1,3-Dichloro-1-propene	0.011	4120854016	0.00	2833947.5	663904256	3440	83.27	277559616	24379362
26	(*S*)-3-Iodo-1,2-propanodiol	−1.261	42241433600	80.92	187263327	3959013376	169988944	57.37	561084736	512510048
27	2-Chloro-2-nitropropane	−1.875	56509394944	43.14	1141175.375	5407563264	123703072	85.18	1384615900	76588288
28	(*R*)-1-Chloro-1-nitropropane	−0.679	33910702080	43.14	1291310.5	3943764992	123702896	85.18	1031692288	53157732
29	(*S*)-2,3-Dichloro-1-propanol	−0.747	40742600704	40.46	2827282.5	4724036096	77915872	155.83	2710674432	50294844
30	3-Bromo-1-propanol	−1.694	25292779520	40.46	11172864	2804492544	77915872	85.87	879171264	55146760
31	Dibromoacetic acid	−1.203	12629614592	57.53	18618924	1466588211	60212696	179.88	2826977536	55349956
32	2-Chloroethyl acrylate	−1.215	12221314048	26.30	2575745.75	2902270976	21329620	278.79	1935080897	47216572
33	(*S*)-2-(2,2,2-Trichloroethyl)oxirane	−1.194	1.11E+11	9.23	3150522.5	9492755456	192085.2969	333.86	12412784640	99825096
34	(*S*)-2-(Trichloromethyl)oxirane	0.455	1.03E+11	9.23	3065597	6200264704	143557.9219	264.14	12681416704	69385176
35	(*R*)-2-(Bromomethyl)oxirane	−0.371	63128502272	9.23	11247736	5607643648	143557.9219	115.61	2829678592	45953544
36	3-Chloro-*N*,*N*-dimethyl-1-propanamine	−2.400	39723761664	3.24	3492382.5	5760996864	52503.07813	193.93	1979739511	120393544
37	Bromochloroacetic acid	−1.207	11056564224	57.53	10061060	1537919671	60212696	182.26	2953888512	35974208
38	(*S*)-2,3-Dibromopropyl acrylate	−1.551	48278032384	26.30	20117806	8053475840	25907394	443.37	5407938560	149360656
39	(*R*)-1-Bromo-2-propanol	−1.793	33050613760	40.46	11188269	3278631680	77919208	85.87	1019503675	69423184
40	2-Bromoacrylaldehyde	−0.208	462832512	17.07	9807250	305336992	10519860	98.47	200182816	18189858
41	3-Bromo-3-buten-2-one	−0.886	3105279488	17.07	9815323	1785719556	15034232	144.31	553114304	41683616
42	Iodoacetic acid	0.776	5343832576	57.53	187019040	718894052	60212696	81.38	395034432	352284288

Table 3. Correlation matrix of the eight variables of the model

	μ_{15}^{Std}	μ_{1}^{Pols}	μ_{7}^{Mol}	$\mu_0\mu_{13}^{Dip}$	$\mu_0\mu_{5}^{Pols}$	$\mu_1\mu_{2}^{Dip2}$	$\mu_1\mu_{14}^{Dip2}$	$\mu_1\mu_{9}^{Pol}$
μ_{15}^{Std}	1.00	0.28	−0.02	0.90	0.51	0.31	0.75	0.25
μ_{1}^{Pols}		1.00	0.29	0.50	0.89	0.47	0.23	0.39
μ_{7}^{Mol}			1.00	−0.02	0.16	−0.16	−0.16	0.91
$\mu_0\mu_{13}^{Dip}$				1.00	0.75	0.43	0.62	0.27
$\mu_0\mu_{5}^{Pols}$					1.00	0.43	0.29	0.36
$\mu_1\mu_{2}^{Dip2}$						1.00	0.66	0.04
$\mu_1\mu_{14}^{Dip2}$							1.00	0.06
$\mu_1\mu_{9}^{Pol}$								1.00

eigenvalue-based indices,[69] information,[70–73] and 2D-autocorrelation descriptors[74–76] (Table 4).

The models employing the above mentioned descriptors were developed using the same data set excluding outliers (40 compounds in total) and a maximum of eight variables for keeping the ratio between cases and variables greater to 5.[24,39–41,63,77] The comparisons were done based on regression analysis results, the predictive capability of the generated models.

There were substantial differences in the explanations of the experimental variance given by these models, compared with the TOPS-MODE model system. Thus, while the TOPS-MODE model was able to explain 90% of mutagenic potency, the other models could only explain 82.8% of such variance, at best. This implies that the predictive capability of the TOPS-MODE model is better than the other five models, not only in the statistical parameters of regression but also, and more importantly, in terms of stability for inclusion/exclusion of chemicals, as measured by the determination coefficient (Q^2). Thus, the value of the determination coefficient for leave-one-out cross-validation for the model obtained with the spectral moment ($Q^2 = 0.842$) was the highest of all. Moreover the spectral moments possess the best goodness-of-fit if we observe the values of F, s, AIC, and FIT.

3.3. Structural alerts identification (fragments contribution)

As it is shown in Table 5, the variables weighted with bond dipole moment explain the 66.9% of the descriptor values for specific data set of chemicals used in the analysis. The variables weighted with polar surface, polarizability, molar refractivity, and standard bond distance accounted for 10.6%, 8.0%, 3.3%, and 1.4% of the variance respectively. Thus, bond dipole moment was a key molecular driver of this mutagenicity. The lack of hydrophobicity features agrees with some author regarding the mutagenic direct-action mechanism.[78]

One advantage of the present approach for QSTR and QSAR studies is that it can provide information explaining how structural features of molecules can account for their biological activities. Thus it is possible to detect fragments that contribute positively or negatively to a particular biological activity and their effects can be interpreted in terms of physicochemical properties.[79] The structural fragments identified in the present study are shown in Figure 1, and their contributions to mutagenic potency are in Table 6.

Table 5. Contribution of spectral moments to the model

Variables	Global R^2 (step by step)	Contribution to global R^2 for every variable
$\Omega\mu_1\mu_{2}^{Dip2}$	0.421	0.421
$\Omega^2\mu_0\mu_{13}^{Dip}$	0.595	0.174
$\Omega^7\mu_1\mu_{9}^{Pol}$	0.675	0.08
$\Omega^6\mu_{1}^{Pols}$	0.753	0.078
$\Omega^8\mu_1\mu_{14}^{Dip2}$	0.827	0.074
$\Omega^4\mu_{7}^{Mol7}$	0.86	0.033
$\Omega^5\mu_0\mu_{5}^{Pols}$	0.888	0.028
$\Omega^3\mu_{15}^{Std}$	0.902	0.014

Table 4. The statistical parameters of the linear regression models with 40 compounds (except outliers) obtained for the six kinds of descriptors involved in the comparison

Descriptors	Variables	R^2	F	p	s	Q^2	AIC	FIT	Q^2_{boot}	$a(R^2)$	$a(Q^2)$
Spectral moments	μ_{15}^{Std}, μ_{1}^{Pols}, μ_{7}^{Mol}, $\mu_0\mu_{13}^{Dip}$ $\mu_0\mu_{5}^{Pols}$, $\mu_1\mu_{2}^{Dip2}$, $\mu_1\mu_{14}^{Dip2}$, $\mu_1\mu_{9}^{Pol}$	0.902	35.730	$<10^{-5}$	0.548	0.842	0.727	1.529	0.718	0.125	−0.392
RDF	RDF055v, RDF010u, RDF030v, RDF055p RDF025u, RDF025p, RDF020e, RDF035u	0.812	16.790	$<10^{-5}$	0.754	0.682	0.838	1.287	0.175	0.155	−0.489
Geometrical	G(O..Cl), FDI, HOMT, G(O..I) SPH, SPAN, G2, ASP	0.828	18.731	$<10^{-5}$	0.726	0.659	0.766	1.435	0.442	0.16	−0.478
Eigenvalue-based indices	SEigZ, SEige, AEigp, VRA1 LP1, SEigv, AEigm, Eig1e	0.785	14.211	$<10^{-5}$	0.811	0.647	0.958	1.088	0.505	0.143	−0.555
WHIM	L2s, L3u, E3m, G2e G3u, E3e, Dp, G3s	0.778	13.649	$<10^{-5}$	0.824	0.574	0.989	1.045	0.293	0.158	−0.522
Information	IC1, IDDE, IC2, CIC4 HVcpx, Yindex, Uindex, TIC5	0.810	16.617	$<10^{-5}$	0.762	0.562	0.845	1.271	0.430	0.167	−0.553
2D-autocorrelation	GATS5v, ATS6e, MATS5e, ATS1m GATS2e, ATS2m, GATS4m, MATS5p	0.798	15.334	$<10^{-5}$	0.787	0.390	0.902	1.178	0.222	0.15	−0.648

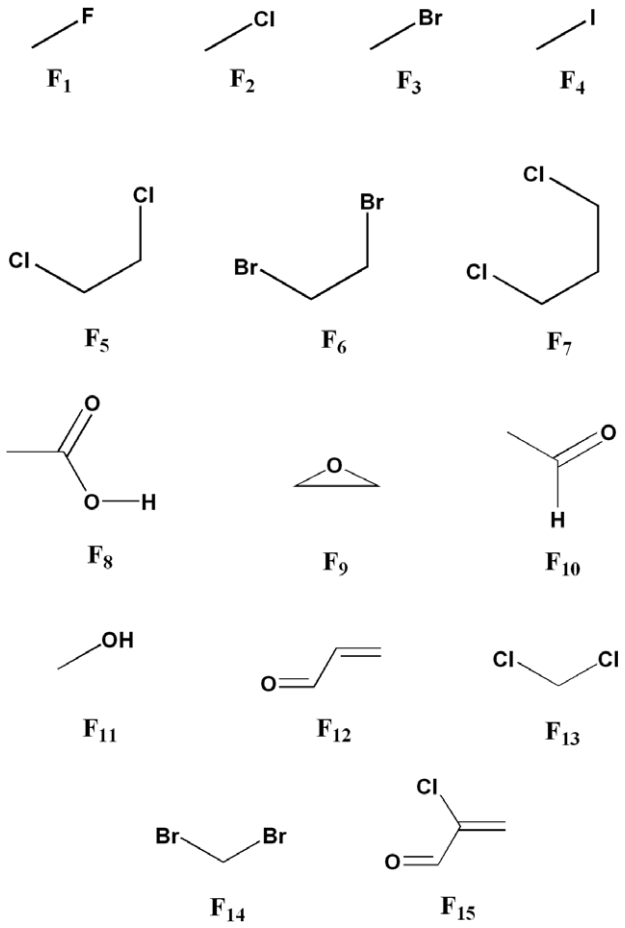

Figure 1. Structural of selected fragment for which their contribution to the mutagenic potency was calculated.

Table 6. The contributions of different structural fragments to the mutagenic activity

Fragments	Contributions
F_1	0.14
F_2	0.45
F_3	0.69
F_4	2.20
F_5	0.90
F_6	1.60
F_7	2.15
F_8	−0.63
F_9	0.67
F_{10}	0.50
F_{11}	−1.09
F_{12}	1.07
F_{13}	0.95
F_{14}	1.47
F_{15}	2.20

The results obtained for fragments F_1, F_2, F_3, and F_4 shows that the mutagenicity order follows the following rule: I > Br > Cl > F. This sequence order is in agreement with the halide order of reactivity and its potential as a leaving group. Some examples of these compounds are in Figure 2. The chemical reactivity of monohaloacetic acids is expected to be similar to that of alkyl halides. The reactivity of methyl halides is primarily dependent on the carbon–halogen bond dissociation energy which is related to the bond strength. Since the bond dissociation energy of the halogen follows the order I < Br < Cl, the strength of carbon–halogen bond increases accordingly. Polarizability and delocalization of the electron could also contribute to making iodine a better leaving group than bromine and much better leaving group than chlorine.[15] This positive contribution of the presence of the halogen substitutent in the mutagenicity was confirmed previously by González et al.[80]

The dihalogenated (Fig. 3), if it is vecinal, increase the mutagenicity since it can act, either directly or after Glutathione conjugation to form the episulfonium ion (powerful electrophile) as cross-linking agents[81,82] (Fig. 4), which we can see through the values of the fragments F_{13}, F_{14}, F_6, and F_7.

Moreover, various other fragments, present in the molecules under study, like epoxide and carbonyl group, agree with the values reported by other authors confirming the known mechanism. Epoxides, aldehydes and carbonyl groups α,β-unsaturated increase this kind of mutagenicity, if we look at the positive values of fragments F_9, F_{10}, and F_{12} (Fig. 5). The epoxides and aldehydes are potential alkylating agents and specially short-chain aldehydes.[83–85] Carbonyl groups α,β-unsaturated have an initial mechanism-type Michael addition[86] (Fig. 6). The chlorine atom presence in position 2 increases the mutagenicity[87] like we can see in fragments F_{12} and F_{15} (Fig. 5), due to the cross-linking potential with another DNA or protein nucleophilic center.[88]

Carboxilic acid and alcohol groups reduce the mutagenicity in the light of the negative values for the fragments F_8 and F_{11} (Fig. 7), since both acids and aliphatic alcohols, the latest one with low numbers of carbon atoms, negative results were in this type of assay.[89–93]

3.4. Applicability domain

The prime overall goal of QSAR research is to develop models that provide accurate predictions for as many chemicals as possible in the universe, particularly for those that have not been tested or for which reliable experimental data are still not available, as well as the properly assessed safety of new chemicals. We defined the applicability domain determining the leverage values for each compound. As seen in Figure 8, the majority of compounds of the training set are inside of this area, however, three halogenated compounds have a leverage greater than h^*, but show standard deviation values within the limit, which implies that they are not to be considered outliers but influential chemicals.[63]

4. Prediction for haloacetic acids

The predictions for the rest of haloacetic acids obtaining with this model are shown in Table 7.

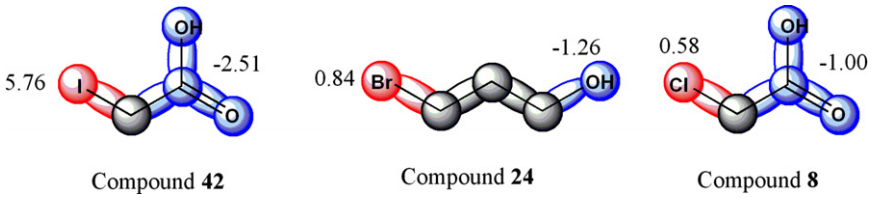

Figure 2. Contribution of the halogens (the fragments in blue have a negative contribution and the fragments in red have a positive contribution).

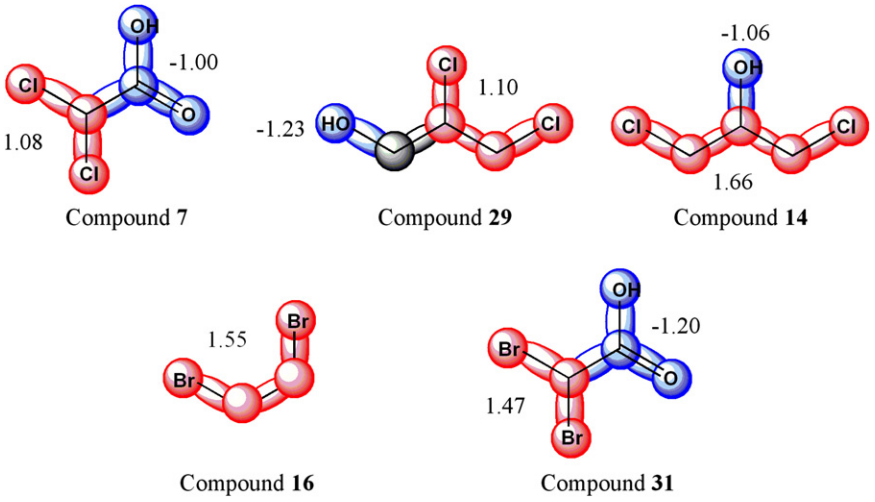

Figure 3. Fragment contribution of some dihalogenated derivatives.

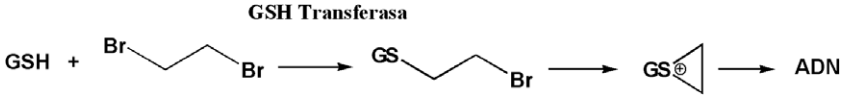

Figure 4. Dihalogenated activation mechanism through glutathione-*S*-transferase.

The compounds marked are the substances which are outside of the applicability domain and then its predictions are not valid. This model predicts a mutagenic activity for most of the dihaloacetic acids greater than trihaloacetic acids. This fact is in agreement with the halogen atom-leaving tendency. This tendency tends to decrease when the halogenation degree increases. The electron-withdrawing effect of the second or third halogen diminishes the leaving potential of the first halogen.[82] The values of the three haloacetic acids: TCAA (TriChloroacetic acid), BDCAA (Bromodichloroacetic acid), and TBAA (Tribromoacetic acid) have an experimental negative value. The model predicts very low values. These results are closer to the experimental results, especially for BDCAA and TBAA. These results can be explained because the model is aimed primarily by the Dipole moments which decrease with increasing the degree of halogenation (Table 8).

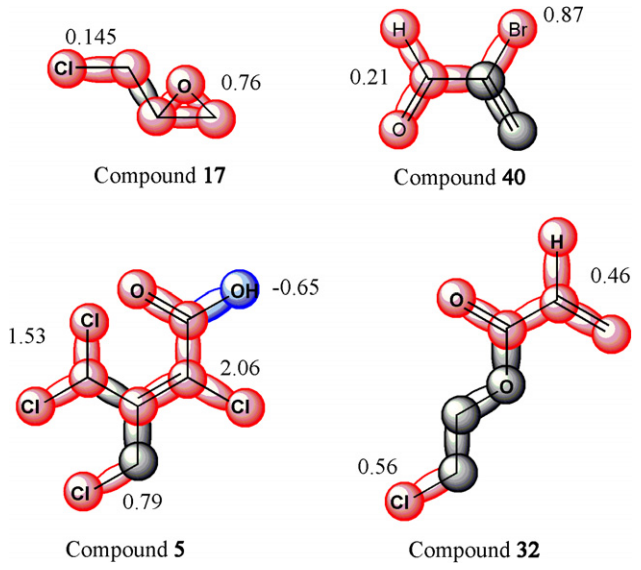

Figure 5. Fragment contribution of some epoxide, aldehyde and carbonyl α,β-unsaturated.

According to the results, we can say that the fluoroiodoacetic (FIAA) and difluoroiodoacetic (DFIAA) acid could show mutagenicity and these substances could show mutagenicity. These compounds can be found most probably in fluorinated waters with a high content of bromide (and iodide).[94]

5. Conclusions

In this paper, we modeled the mutagenicity activity in *S. typhimurium* Ames test TA100 strain without metabolic

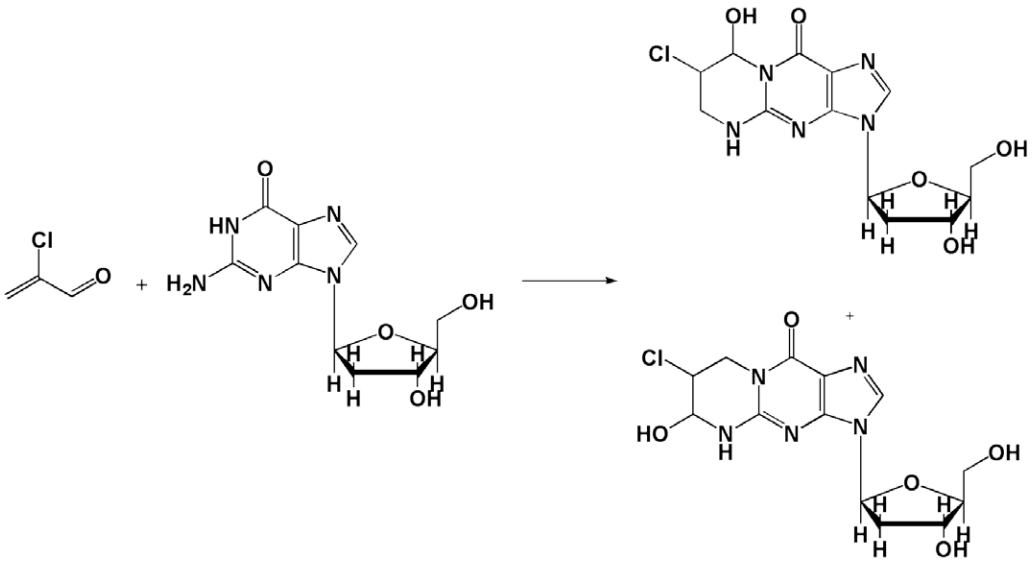

Figure 6. Addition Michael type mechanism for the carbonyl group α,β-unsaturated with chlorine in position 2 with deoxyguanosine.

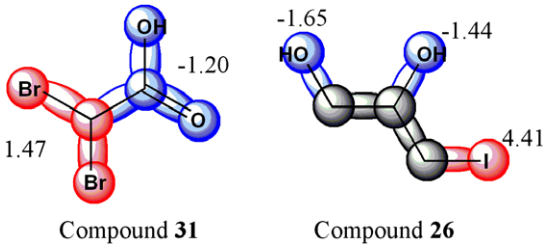

Figure 7. Fragment contribution of some hydroxil and acid moieities.

activation to predict such property for haloacetic acids (HAA). For this purpose, we employed the spectral moments and the variables that were found to be most significant to build the model. These variables were basically dipole moments and polar surface, polarizability, molar refractivity, and standard bond distance.

The spectral moments with multiplicative effects had better results than other linear descriptors such as Geometrical, RDF, WHIM, eigenvalue-based indices, 2D-autocorrelation, and information indices, taking into account the statistical parameters of the model and the cross-validation results.

The structural alerts and the mutagenicity predicted values from the model output are in agreement with references from other authors. This model predicts mutagenicity for fluoroiodoacetic and difluoroiodoace-

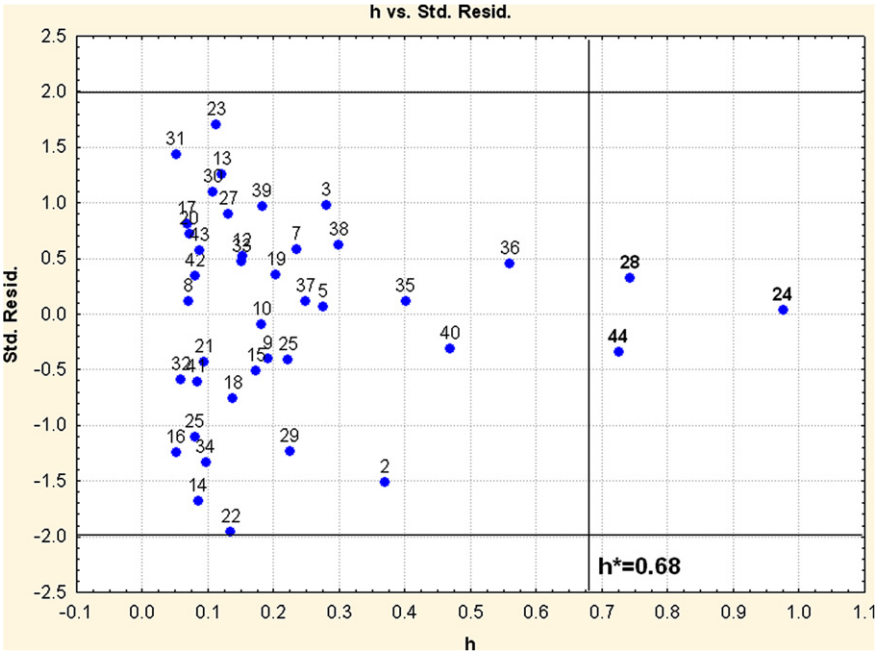

Figure 8. Applicability domain of the model of Eq. 7.

Table 7. Prediction for the haloacetic acids

Compounds	Log $TA100_{Pred.}$	*h*, Leverage	Mutagenicity[a]
Fluoroiodoacetic acid (FIAA)	1.698	0.447	ND
Difluoroiodoacetic acid (DFIAA)	1.008	0.390	ND
Chloroiodoacetic acid (CIAA)	−0.557	0.281	ND
Chlorofluoroiodoacetic acid (CFIAA)	−0.832	0.235	ND
Bromoiodoacetic acid (BIAA)	−1.326	0.371	ND
Bromofluoroiodoacetic acid (BFIAA)	−1.342	0.303	ND
Bromofluoroacetic acid (BFAA)	−1.527	0.164	ND
Trichloroacetic acid (TCAA)	−1.806	0.385	–[b]
Fluoroacetic acid (FAA)	−1.856	0.158	ND
Chlorofluoroacetic acid (CFAA)	−1.951	0.186	ND
Difluoroacetic acid (DFAA)	−1.972	0.191	ND
Difluoroacetic acid (DBFAA)	−2.401	0.336	ND
Bromodichloroacetic acid (BDCAA)	−2.410	0.368	–[b]
Dichlorofluoroacetic acid (DCFAA)	−2.489	0.396	ND
Bromodifluoroacetic acid (BDFAA)	−2.713	0.395	ND
Dichloroiodoacetic acid (DCIAA)	−2.714	0.555	ND
Tribromoacetic acid (TBAA)	−2.756	0.341	–[b]
Bromochlorofluoroacetic acid (BCFAA)	−2.844	0.392	ND
Dibromochloroacetic acid (DBCAA)	−2.949	0.364	ND
Chlorodifluoroacetic acid (CDFAA)	−3.254	0.445	ND
Trifluoroacetic acid (TFAA)	−3.274	0.441	ND
Bromochloroiodoacetic acid (BCIAA)	−4.402	0.695	ND
Dibromoiodoacetic acid (DBIAA)	−5.554	0.798	ND
Fluorodiiodoacetic acid (FDIAA)	−8.450	0.907	ND
Diiodoacetic acid (DIAA)	−9.656	0.926	ND
Chlorodiiodoacetic acid (CDIAA)	−15.668	0.967	ND
Bromodiiodoacetic acid (BDIAA)	−18.840	0.977	ND
Triiodoacetic acid (TIAA)	−45.679	0.996	ND

[a] Mutagenicity values in *S. typhimurium* TA100 strain without metabolic activation extracts of the MDL Toxicity Database. ND no data.
[b] –, Negative assay.

Table 8. Dipole moments of the haloacetic acids

Compounds	Dipole moment[a]
FAA	3.230
DFAA	0.144
TFAA	1.745
CAA	2.716
DCAA	1.284
TCAA	1.564
BAA	2.722
DBAA	1.048
TBAA	1.552
IAA	1.794
DIAA	2.163
TIAA	1.445

[a] Dipole moment calculated with the Vamp module of Tsar 3.3 software.

tic acids and these compounds can be found most probably in fluorinated waters high in bromide (and iodide).

Acknowledgment

The authors acknowledge to MODESLAB 1.0 software owners for delivering a free copy of such program.

References and notes

1. Akin, E. W.; Hoff, J. C.; Lippy, E. C. *Environ. Health Perspect.* **1982**, *46*, 7.
2. Wilcox, P.; Williamson, S. *Environ. Health Perspect.* **1986**, *69*, 141.
3. Morris, R. D.; Audet, A. M.; Angelillo, I. F.; Chalmers, T. C.; Mosteller, F. *Am. J. Public Health* **1992**, *82*, 955.
4. Koivusalo, M.; Jaakkola, J. J.; Vartiainen, T.; Hakulinen, T.; Karjalainen, S.; Pukkala, E.; Tuomisto, J. *Am. J. Public Health* **1994**, *84*, 1223.
5. Bull, R. J.; Birnbaum, L. S.; Cantor, K. P.; Rose, J. B.; Butterworth, B. E.; Pegram, R.; Tuomisto, J. *Fundam. Appl. Toxicol.* **1995**, *28*, 155.
6. Nieuwenhuijsen, M. J.; Toledano, M. B.; Elliott, P. *J. Exposure Anal. Environ. Epidemiol.* **2000**, *10*, 586.
7. Bull, R. J.; Sánchez, I. M.; Nelson, M. A.; Larson, J. L.; Lansing, A. J. *Toxicology* **1990**, *63*, 341–359.
8. DeAngelo, A. B.; Daniel, F. B.; Stober, J. A.; Olson, G. R. *Fundam. Appl. Toxicol.* **1991**, *16*, 337–347.
9. DeAngelo, A. B.; Daniel, F. B.; Most, B. M.; Olson, G. R. *Toxicology* **1996**, *114*, 207–221.
10. Herbert, V.; Gardner, A.; Colman, N. *Am. J. Clin. Nutr.* **1980**, *33*, 1179–1182.
11. Nestman, E. R.; Chu, I.; Kowbel, D. J.; Matula, T. I. *Can. J. Genet. Cytol.* **1980**, *22*, 35–40.
12. DeMarini, D. M.; Perry, E.; Shelton, M. L. *Mutagenesis* **1994**, *9*, 429–437.
13. Kargalioglu, Y.; McMillan, B. J.; Minear, R. A.; Plewa, M. J. *Teratog. Carcinog. Mutagen.* **2002**, *22*, 113–128.
14. Kundu, B.; Richardson, S. D.; Swartz, P. D.; Matthews, P. P.; Richard, A. M.; DeMarini, D. M. *Mutat. Res.* **2004**, *562*, 39–65.
15. Plewa, M. J.; Wagner, E. D.; Richardson, S. D.; Thruston, A. D. J.; Woo, Y. T.; McKague, A. B. *Environ. Sci. Technol.* **2004**, *38*, 4713–4722.
16. McCann, J.; Choi, E.; Yamasaki, E.; Ames, B. N. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 5135–5139.

17. Sugimura, T.; Sato, S.; Nagao, M.; Yahagi, T.; Matsushima, T.; Seino, Y.; Takeuchi, M.; Kawachi, T. Fundamental of Cancer Prevention. In Magee, P., Takayama, S., Sugimura, T., Matsushima, T., Eds.; University Park Press: Baltimore, 1976.
18. Zeiger, E.; Haseman, J. K.; Shelby, M. D.; Margolin, B. H.; Tennant, R. W. *Environ. Mol. Mutagen.* **1990**, *16*, 1–14.
19. Venkatapathy, R.; Bruce, R.; Moudgal, C. Presented at the EPA Science Forum, Mandarin Oriental Hotel, Washington, DC; Available from: <http://www.epa.gov/ord/scienceforum/2004/poster-ord-NtoZ.htm>, 2004.
20. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Mannheim, 2000.
21. Saiz-Urra, L.; González, M. P.; Collado, I. G.; Hernandez-Galan, R. *J. Mol. Graphics Modell.* **2007**, *25*, 680–690.
22. Estrada, E.; Molina, E. *J. Mol. Graphics Modell.* **2006**, *25*, 275–288.
23. González, M. P.; Teran, C.; Teijeira, M. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1291–1296.
24. Morales, A. H.; González, M. P.; Rieumont, J. *Polymer* **2004**, *45*, 2045–2050.
25. Morales, A. H.; Perez, M. A. C.; González, M. P.; Ruiz, R. M.; Díaz, H. G. *Bioorg. Med. Chem. Lett.* **2005**, *13*, 2477–2488.
26. Morales, A. H.; Cabrera, M. A.; Combes, R. D.; González, M. P. *Toxicology* **2006**, *220*, 51.
27. Morales, A. H.; González, M. P.; Cordeiro, M. N. D. S.; Perez, M. A. C. *Toxicol. Appl. Pharmacol.* **2007**, *221*, 189–202.
28. González-Díaz, H.; Gia, O.; Uriarte, E.; Hernadez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Model.* **2003**, *9*, 395–407.
29. González-Díaz, H.; Olazabal, E.; Castanedo, N.; Sánchez, I. H.; Morales, A.; Serrano, H. S.; González, J.; de Armas, R. R. *J. Mol. Model.* **2002**, *8*, 237–245.
30. González-Díaz, H.; Torres-Gomez, L. A.; Guevara, Y.; Almeida, M. S.; Molina, R.; Castanedo, N.; Santana, L.; Uriarte, E. *J. Mol. Model.* **2005**, *11*, 116–123.
31. González-Díaz, H.; Pérez-Castillo, Y.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28*, 1990–1995.
32. González-Díaz, H.; Saíz-Urra, L.; Molina, R.; González-Díaz, Y.; Sánchez-González, A. *J. Comput. Chem.* **2007**, *28*, 1042–1048.
33. González-Díaz, H.; Bonet, I.; Terán, C.; De Clercq, E.; Bello, R.; García, M. M.; Santana, L.; Uriarte, E. *Eur. J. Med. Chem.* **2007**, *42*, 580–585.
34. Prado-Prado, F. J.; González-Díaz, H.; Santana, L.; Uriarte, E. *Bioorgan. Med. Chem.* **2007**, *15*, 897–902.
35. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 844–849.
36. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 320–328.
37. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 23–27.
38. González, M. P.; Díaz, H. G.; Ruiz, R. M.; Cabrera, M. A.; de Armas, R. R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1192–1199.
39. Gonzalez, M. P.; Teran, C.; Fall, Y.; Diaz, L. C.; Helguera, A. M. *Polymer* **2005**, *46*, 2783.
40. González, M. P.; Helguera, A. M.; Molina, R.; García, J. R. *Polymer* **2004**, *45*, 2773.
41. González, M. P.; Dias, L.; Helguera, A. M. *Polymer* **2004**, *15*, 5353.
42. González, M. P.; Helguera, A. M.; Cabrera, M. A.; González, H. *Bioorg. Med. Chem.* **2005**, *13*, 1775.
43. González, M. P.; Díaz, H. G.; Cabrera, M. A.; Ruiz, R. M. *Bioorg. Med. Chem.* **2004**, *12*, 735.
44. Mortelmans, K.; Zeiger, E. *Mutat. Res.* **2000**, *455*, 29–60.
45. Bernstein, L.; Kaldor, J.; McCaan, J.; Pike, M. C. *Mutat. Res.* **1982**, *97*, 267–281.
46. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31–33.
47. Estrada, E.; Uriarte, E.; Gutierrez, Y.; González, H. *SAR QSAR Environ. Res.* **2003**, *14*, 145.
48. Gutierrez, Y.; Estrada, E. 'Modes Lab, version 1.0', 2002.
49. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
50. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. *J. Comput. Aided Mol. Des.* **2005**, *19*, 453–463.
51. Draper, N. R.; Smith, H. *Applied Regression Analysis*, Second ed.; John Wiley and Sons: New York, 1981.
52. Garcia-Domenech, R.; Julian-Ortiz, J. V. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445–449.
53. Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285.
54. Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393.
55. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*; Akademiai Kiado: Budapest, 1973.
56. Akaike, H. *IEEE Trans. Automat. Control* **1974**, *AC-19*, 716.
57. Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A.; Pavan, M. 'Mobydigs Computer Software', 2004.
58. Lucic, B.; Nikolic, S.; Trinajstic, N.; Juric, D. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532.
59. Klein, D.; Randi, M.; Babic, D.; Lucic, B.; Nikolic, S.; Trinajstic, N. *Int. J. Quantum Chem.* **1997**, *63*, 215.
60. Randic, M. *New J. Chem.* **1991**, *15*, 517–525.
61. Randic, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–320.
62. Randic, M. *J. Mol. Struct. (Theochem.)* **1991**, *233*, 45–59.
63. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361.
64. Gramatica, P. *QSAR Comb. Sci.* **2007**, *00*, 1–9.
65. Vighi, M.; Gramatica, P.; Consolaro, F.; Todeschini, R. *Ecotoxicol. Environ. Saf.* **2001**, *49*, 206–220.
66. Munter, T.; Kronberg, L.; Sjöholm, R. *Chem. Res. Toxicol.* **1996**, *9*, 703708.
67. Cemelli, E.; Wagner, E. D.; Anderson, D.; Richardson, S. D.; Plewa, M. J. *Environ. Sci. Technol.* **2006**, *40*, 1878–1883.
68. Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *J. Vib. Spectrosc.* **1999**, *19*, 151–164.
69. Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517–523.
70. Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; RSP-Wiley: Chichetser (UK), 1983.
71. Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. *J. Comput. Chem.* **1984**, *5*, 581–588.
72. Balaban, A. T.; Balaban, T.-S. *J. Math. Chem.* **1991**, *8*, 383–397.
73. Magnuson, V. R.; Harriss, D. K.; Basak, S. C. Chemical applications of topology and graph theory. In King, R. B., Ed.; Elsevier: Amsterdam (The Netherlands), 1983.
74. Moran, P. A. P. *Biometrika* **1950**, *37*, 1723.
75. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359–360.
76. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 757–764.
77. González, M. P.; Helguera, A. M.; González-Díaz, H. *Polymer* **2004**, *45*, 2073.
78. LaLonde, R. T.; Leo, H.; Perakyla, H.; Dence, C. W.; Farrell, R. P. *Chem. Res. Toxicol.* **1992**, *5*, 392.
79. Benigni, R.; Giuliani, A. *Bioinformatics* **2003**, *19*, 1194–1200.
80. González, M. P.; Dias, L. C.; Helguera, A. M. *Polymer* **2004**, *45*, 5353–5359.

81. Guengerich, F. P. *Jpn. J. Toxicol. Environ. Health* **1997**, *43*, 69–82.
82. Woo, Y.-T.; Lai, D.; Arcos, J. C.; Argus, M. F. *Chemical Induction of Cancer, Structural Bases and Biological Mechanism*. In *Aliphatic and Polyhalogenated Carcinogens*; Academic Press: Orlando, Florida, 1985; vol. IIIB.
83. Woo, Y. T.; Lai, D. Y.; McLain, J. L.; Ko Manibusan, M.; Dellarco, V. *Environ. Health Perspect.* **2002**, *110*, 75.
84. Simon, P.; Epe, B.; Mtzel, P.; Schiffmann, D.; Wild, D.; Ottenwlder, H.; Fedtke, N.; Bolt, H. M.; Henschler, D. *J. Biochem. Toxicol.* **1997**, *1*, 43–55.
85. Castelain, P. H.; Criado, B.; Cornet, M.; Laib, R.; Rogiers, V.; Kirsch-Volders, M. *Mutagenesis* **1993**, *8*, 387–393.
86. Eder, E.; Henschler, D.; Neudecker, T. *Xenobiotica* **1982**, *12*, 831–848.
87. Eder, E.; Weinfurtner, E. *Chemosphere* **1994**, *29*, 2455–2466.
88. Van Beerendonk, G. J. M.; Nivard, M. J. M.; Vogel, E. W.; Nelson, S. D.; Meerman, J. H. N. *Mutagenesis* **1992**, *7*, 19–24.
89. McGregor, D. B.; Cruzan, G.; Callander, R. D.; May, K.; Banton, M. *Mutat. Res.* **2005**, *565*, 181.
90. Stolzenberg, S. J.; Hine, C. H. *J. Toxicol. Environ. Health* **1979**, *5*, 1149–1158.
91. Simmon, V. F.; Kauhanen, K.; Tardiff, R. G. Progress in Genetic Toxicology. In *Chapter Mutagenic Activity of Chemicals Identified in Drinking Water*; Elsevier: North Holland Press, Amsterdam, 1977; pp 249–268.
92. Heck, J. D.; Vollmuth, T. A.; Cifone, M. A.; Jagannath, D. R.; Myhr, B.; Curren, R. D. *The Toxicologist* **1989**, *9*, 257.
93. Philipose, B.; Singh, R.; Khan, K. A.; Giri, A. K. *Mutat. Res.* **1997**, *393*, 123–131.
94. Krasner, S. W.; Weinberg, H. S.; Richardson, S.; Pastor, S. J.; Chinn, R.; Sclimenti, M. J.; Onstad, G. D.; Thruston, A. D. *Environ. Sci. Technol.* **2006**, *40*, 7175–7185.
95. Franzen, R.; Goto, S.; Tanabe, K.; Morita, M. *Mutat. Res.* **1998**, *417*, 31–37.
96. Gordon, W. P.; Soederlund, E. J.; Holme, J. A.; Nelson, S. D.; Iyer, L.; Rivedal, E.; Dybing, E. *Carcinogenesis* **1985**, *6*, 705–709.
97. Omichinski, J. P.; Soederlund, E. J.; Bausano, J. A.; Dybing, E.; Nelson, S. D. *Mutagenesis* **1987**, *2*, 287–292.