J Mol Model (2006) 12: 769–780
DOI 10.1007/s00894-005-0088-5

ORIGINAL PAPER

Aliuska Helguera Morales · Miguel Ángel Cabrera Pérez · Maykel Pérez González

A radial-distribution-function approach for predicting rodent carcinogenicity

Received: 14 July 2005 / Accepted: 22 November 2005 / Published online: 19 January 2006

Abstract Carcinogenic activity has been investigated using the Radial-Distribution-Function (RDF) approach. A discriminant model was developed to predict the carcinogenic and non-carcinogenic activity on a data set of 188 compounds. The percentage of overall classification was 76.4% for the carcinogenic chemicals and 72.5% for the non-carcinogenic chemicals. The predictive power of the model was validated by two tests: a cross-validation by the resubstitution technique and a test set (compounds not used in the development of the model) with 79.3 and 72.5% good classification, respectively. The RDF descriptors were compared with eight other methodologies; Constitutional, Molecular walks counts, Galvez topological charge indices, 2D autocorrelations, Randić molecular profiles, Geometrical, 3D-MORSE, and WHIM, demonstrating that the RDF descriptors are better to the rest of the approaches used.

Keywords Dragon · Carcinogenesis · QSAR · RDF descriptors

A. H. Morales
Department of Chemistry,
Central University of Las Villas,
Santa Clara,
Villa Clara, 54830, Cuba

A. H. Morales · M. Á. Cabrera Pérez · M. P. González
Department of Drug Design, Chemical Bioactives Center,
Central University of Las Villas,
Santa Clara,
Villa Clara, 54830, Cuba

M. P. González (✉)
Unit of Services, Department of Drug Design,
Experimental Sugar Cane Station "Villa Clara-Cienfuegos",
Ranchuelo,
Villa Clara, 53100, Cuba
e-mail: mpgonzalez76@yahoo.es
Tel.: +53-42-281473
Fax: +53-42-281130

Introduction

Success in the drug design field is to identify a new compound with an appropriate balance of potency, safety and favorable pharmacokinetics (PK). Several studies in the late 1990s suggested that poor PK and toxicity were among the most important causes of late-stage failures of compounds in drug development. [1] Today, advances in combinatorial chemistry and ultra-high-throughput-screening have made it possible to obtain a dramatic increase in the size of compound collections in pharmaceutical companies, increasing the rate at which biological activity data can be obtained [2, 3].

For that reason, in recent years pharmaceutical companies have brought toxicity testing, as well as ADME (absorption, distribution, metabolism, excretion) studies, into the drug development process earlier. The ultimate here would be to use computer-based (in silico) methods to predict toxicity even before a drug candidate is synthesized. [4] Among toxic endpoints, the chemical carcinogenicity is of primary interest, because it drives much of the current regulatory actions, and its experimental determination involves time-consuming and expensive animal testing. [5] If we consider that long-term carcinogenicity studies in rodents are done late in clinical trials, where this kind of study can take up to two years and typically cost US $2.5 million, [6] in silico studies are widely justified.

Some research institutes and pharmaceutical companies are actively involved in the development of Structure-Activity Relationship (SAR) and Quantitative Structure-Activity Relationship (QSAR) models for the rodent carcinogenicity, which is the main source of experimental data and is an essential tool in risk assessment. [6–8] A large number of systems and models dedicated to the prediction of carcinogenicity have been developed and these have demonstrated that carcinogenicity is generally only poorly predicted [9, 10].

In the reported studies, only few families of descriptors have been used to model the carcinogenic activity and the predictive results were not the best. [11] At this point, the necessity of use other kinds of molecular descriptors is

evident. Among the large amount of molecular descriptors reported, [12] the 3D-Radial-Distribution-Function descriptors (RDF) have demonstrated their potential as useful tools for modeling different physicochemical and biological properties. [13–17] This has inspired us to test and/or validate RDF descriptors in the study of carcinogenesis.

The aims of this work are to use the RDF descriptors in the generation of discriminant functions by a Linear Discriminant Analysis (LDA) that permits the classification of chemicals in carcinogenic and non-carcinogenic and to demonstrate the validity of the "in silico" models by the use of different validation tests. Also, we show a comparison between RDF descriptors and other methodologies.

The radial distribution function approach

These descriptors are based on the distance distribution in the geometrical representation of a molecule and constitute a radial distribution function code [16, 18].

Formally, the radial distribution function of an ensemble of N atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of radius *r*. [17] The general form of the radial distribution function code is represented by:

$$g(r) = f \cdot \sum_{i=1}^{N-1} \sum_{j>i}^{N} A_i A_j e^{-B\left(r-r_{ij}\right)^2}$$

where *f* is a scaling factor and *N* is the number of atoms. By including characteristic atomic properties *A* of the atoms *i* and *j*, the RDF codes can be used in different tasks to fit the requirements of the information to be represented. These atomic properties enable the discrimination of the atoms of a molecule for almost any property that can be attributed to an atom. The exponential term contains the distance r_{ij} between the atoms *i* and *j* and the smoothing parameter *B*, which defines the probability distribution of the individual distances; *B* can be interpreted as a temperature factor that defines the movement of the atoms. $g(r)$ was calculated for a number of discrete points with defined intervals.

The radial distribution function in this form meets all the requirements for 3D-structure descriptors: it is independent of the number of atoms, i.e., the size of a molecule, it is unique regarding the three-dimensional arrangement of the atoms, and it is invariant against translation and rotation of the entire molecule. Additionally, the RDF descriptors can be restricted to specific atom types or distance ranges to represent specific information in a certain three-dimensional structure space, e.g., to describe steric hindrance or structure/activity properties of a molecule.

Finally, the RDF descriptors are interpretable using simple rules sets, and thus provide a possibility for conversion of the code back into the corresponding 3D structure. Besides information about interatomic distances in the entire molecule, RDF descriptors provide further valuable information, e.g., about bond distances, ring types, planar and non-planar systems and atom types.

Data sets and computational strategies

A data set of 188 compounds (119 carcinogenic and 69 non-carcinogenic) was collected from the Carcinogenic Potency Data Base (CPDB) established by Gold et al. [19] available at (http://potency.berkeley.edu/cpdb.html). The CPDB is a single standardized resource of many years of chronic, long-term carcinogenesis bioassays. It contains a large diversity of chemical structures (more than 1,300 tested substance), and includes tumor data reproduced from all of the NCI/NTP rodent bioassay *Technical Reports* as well as additional data extracted from over 1,200 literature sources subjected to extensive review [19].

The criterion followed for the selection of the compounds was; a chemical is categorized as a carcinogen only if it causes tumors at multiple organ sites or in multiple rodent species. [20] Non-carcinogenic compounds were selected when the reported experiments were negative or negative/zero, in multiple rodent species.

The data set was divided in two subsets, training and a test set. The training set contains the 75% of the chemicals of the data set and the test set contains the remaining 25%. The procedure of selection of these two groups was by a cluster analysis technique. [21, 22] The principal idea of cluster analysis consists of carrying out a partition of series of compounds in several statistically representative classes of chemicals. Thence, one may select from the member of all these classes of training and predicting series. This procedure ensures that any chemical classes (as determined by the clusters derived from k-Means Cluster Analysis, k-MCA) will be represented in both compounds series (training and prediction series). It permits the design of both, training and prediction series, which are representative of the entire "experimental universe". Fig. 1 graphically illustrates the above-described procedure, where two independent cluster analyses (one for the carcinogenic compounds and other for the non-carcinogenic compounds) were carried out to select a representative sample for the prediction and training sets.

The first k-MCA (k-MCA1) splits the carcinogenic compounds into four clusters with 30, 34, 27, 28 members, respectively. On the other hand, the series of non-carcinogenic compounds was partitioned into five clusters (k-MCA2) with 6, 8, 19, 18, 18 members. Selection of the training and prediction sets was carried out taking, in a random way and proportionally to size of the cluster, the compounds belonging to each cluster.

To ensure a statistically acceptable data partition into several clusters, the number of members in each cluster and the standard deviation of the variables in the cluster (as low as possible), was taken into consideration. We also made an inspection of the standard deviation between and within clusters, the respective Fisher ratio and their *p* level of significance considered to be lower than 0.05. [23, 24] The RDF descriptors used in both analyses show *p* levels< 0.05 for the Fisher test, and the results are shown in Table 1.

Using this technique, the dataset was split in 148 compounds for the training set (93 carcinogenic and 55

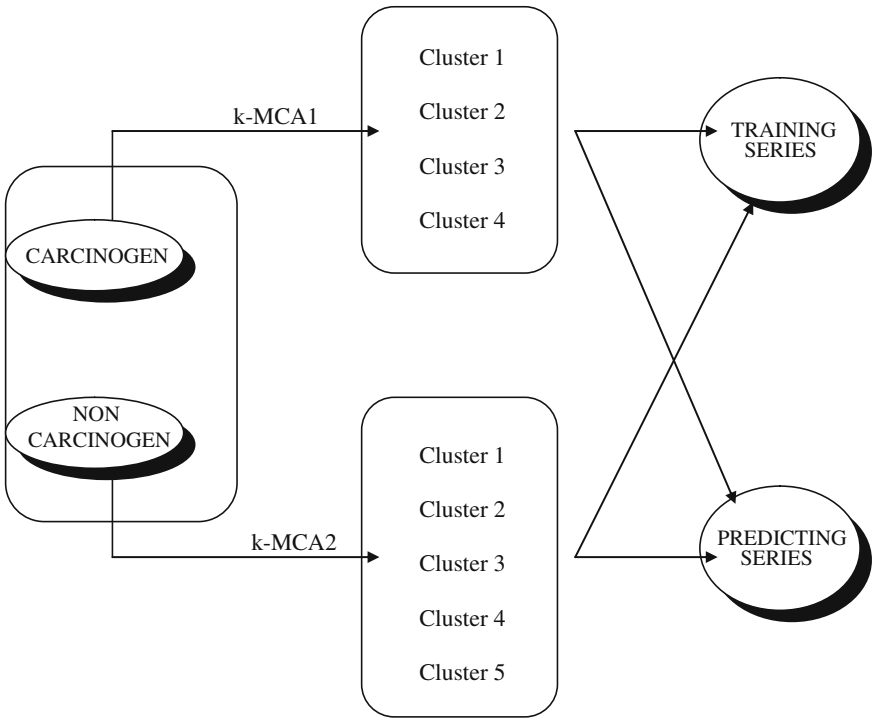

Fig. 1 Graphic representation of selection of the training and predicting series by cluster analysis derived from k-Means Cluster Analysis, k-MCA

non-carcinogenic) and 40 compounds for the test set (26 carcinogenic and 14 non-carcinogenic). The compounds belonging to the test set were never used in the development of the discriminant function and were reserved to validate the discriminant model.

Before the calculations of the RDF descriptors, we carried out geometry optimization of each compound, using the quantum chemical semi-empirical method Austin Model 1 (AM1) [25] included in MOPAC 6.0. [26] The DRAGON [27] computer software was used to calculate the RDF molecular descriptors. The atomic masses, atomic van der Waals volumes, atomic Sanderson electronegativities and atomic polarizabilities were used as bond weightings. Finally 150 RDF descriptors were calculated for each compound of the data set.

Table 1 Results of the k-Means Cluster Analysis for carcinogens and non-carcinogenic compounds

Analysis of variance				
Variables	Between SS[a]	Within SS[b]	Fisher ratio (F)	p[c]<
Statistics for k-MCA 1				
RDF 035e	3207.037	988.405	124.379	10^{-5}
RDF 040e	3006.314	1096.988	105.053	10^{-5}
RDF 045e	3063.796	1018.862	115.271	10^{-5}
RDF 050e	4751.044	1393.056	130.737	10^{-5}
Statistics for k-MCA 2				
RDF 035e	12739.76	2164.685	94.164	10^{-5}
RDF 040e	16403.48	1792.431	146.424	10^{-5}
RDF 045e	17409.25	1846.302	150.868	10^{-5}
RDF 050e	31420.16	2249.088	223.523	10^{-5}

[a]Variability between groups
[b]Variability within groups
[c]Level of significance

The discriminant function was obtained by using the stepwise Linear Discriminant Analysis (LDA) as implemented in STATISTICA version 6.0. [28] The variables to be included in the equation were selected using the forward stepwise procedure as a variable-selection strategy.

The quality of the models was determined examining the Wilks λ. This parameter indicates a perfect discrimination for λ=0 and no discrimination when λ=1. The squared Mahalanobis distance (D^2) was also determined. It indicates the separation of the respective groups, showing if the model possesses appropriate discriminatory power for differentiating between the two groups. The Fisher ratio (F) the corresponding p level (p), the percentage of good classification and the proportion between the cases and variables in the equations were also taken into account.

The compounds were considered unclassified (NC) by the model when the differences in the percentage of classification between groups did not differ by more than 5%. However, we have used the *a posteriori* probabilities in order to classify the compounds as carcinogenic/non-carcinogenic. This is the probability that the respective case belongs to a particular group (carcinogenic or non-carcinogenic). It is proportional to the Mahalanobis distance from the group centroid.

The Randić orthogonalization procedure was applied to each variable of the RDF model obtaining the corresponding orthogonal variables in order to avoid collinearity among variables, model over-fitting as well as to give the model a real interpretation of the influence of the different variables.

The Randić method of orthogonalization has been described in detail in several publications. [29–32] Thus,

we will give a general overview here. The first step to orthogonalize the molecular descriptors in the model (Eq. 1) is to select the appropriate order of orthogonalization, which in this case is the order in which the variables were selected in the forward stepwise search procedure of the LDA. The first variable (RDF080u) is taken as the first orthogonal descriptors $^{1}\Omega$(RDF080u), and the second one is orthogonalized with respect to it by taking the residual of its correlation with $^{1}\Omega$(RDF080u). The process is repeated until all variables are completely orthogonalized, and the orthogonal variables are then used to obtain the new model.

A cross-validation procedure was subsequently performed in order to assess model predictability. This cross-validation was carried out by means of a resubstitution technique composed by some main stages. First, we selected randomly 25% of the compounds out of the training series and constituted the predicting series (cv1). Afterwards, compounds in the predicting series were interactively interchanged with those in training ones (cv2, cv3, cv4).

Finally, the predictive power of discriminant model was assessed with an external dataset reported in CPDB. This test set is formed by 40 chemicals; 26 carcinogens and 14 non-carcinogens chemicals.

The other eight models were developed using the computer software DRAGON, [25] calculating the Constitutional, Molecular walk counts, Galvez topological charge indices, 2D autocorrelations, Randić molecular profiles, Geometrical, 3D-MORSE, and WHIM descriptors. [12] All statistical analyses and data exploration were carried out using Statistic 6.0. [28] The most significance parameters were identified from the data set using forward stepwise regression methods, [33] where the independent variables are individually added or deleted from the model at each step of the regression depending on the Fisher ratio values selected to enter and to remove until the 'best' model is obtained. Taking into account the λ, D^2, F, p, the percentage of good classification of training set and test set (%Class1 and %Class2) values, determined the quality of the models.

Results and discussion

In this work, the model selection was subjected to the principle of parsimony. [34] Then, we choose a function with high statistical signification but with as few parameters (descriptors) as possible.

The best QSAR model obtained with the RDF descriptors is given below together with the statistical parameters of the linear discriminant analysis.

$$\begin{aligned} C_{act} &= -0.207 \cdot RDF045u + 0.526 \cdot RDF080u - 0.717 \cdot RDF120u - 4.791 \cdot RDF010m \\ &\quad +0.091 \cdot RDF025e + 0.234 \cdot RDF065e - 0.485 \cdot RDF085e + 0.776 \cdot RDF125e \\ &\quad +0.755 \cdot RDF130e - 0.794 \cdot RDF140e + 0.209 \\ N &= 148, \lambda = 0.614, D^2 = 2.660, F = 8.621, p < 10^{-5}, \%\text{Class1} = 73.7, \%\text{Class2} = 70.0 \end{aligned} \quad (1)$$

where C_{act} is the studied property, in our case, the carcinogenic activity according to the previous criterion, N is the number of compounds included in the model, λ is the Wilk's statistic, D^2 is the squared Mahalanobis distance, F the Fisher ratio, p is the significance of the variables in the model and %Class1 and %Class2 are the percentage of overall good classification in the training and test set, respectively.

As a result of the analysis of collinearity, we detected high regression coefficients among the values for the descriptors of model 1 (Eq. 1). The Table 2 shows that most of regression coefficients among; RDF045u, RDF080u, RDF120u, RDF010m, RDF025e, RDF065e, RDF085e, RDF152e, RDF130e, RDF140e were higher than 0.70. Therefore, orthogonalization of the molecular descriptors was carried out before interpreting the model.

Table 2 Correlation matrix of the ten variables of model 1

	RDF045u	RDF080u	RDF120u	RDF010m	RDF025e	RDF065e	RDF085e	RDF125e	RDF130e	RDF140e
RDF045u	1.00	0.88	0.64	0.74	0.92	0.87	0.80	0.63	0.61	0.54
RDF080u		1.00	0.80	0.69	0.87	0.94	0.94	0.78	0.75	0.68
RDF120u			1.00	0.54	0.72	0.85	0.90	0.96	0.97	0.92
RDF010m				1.00	0.81	0.72	0.66	0.54	0.52	0.45
RDF025e					1.00	0.91	0.85	0.71	0.69	0.60
RDF065e						1.00	0.95	0.83	0.81	0.75
RDF085e							1.00	0.88	0.88	0.79
RDF125e								1.00	0.94	0.92
RDF130e									1.00	0.94
RDF140e										1.00

The orthogonalization process of molecular descriptors was introduced by Randić ten years ago as a way of improving the statistical interpretation of the model built by using interrelated indices. [35–39] The main philosophy of this approach is to avoid the exclusion of descriptors on the basis of their collinearity with other variables previously included in the model. The acceptable level of collinearity to avoid is a more subjective issue. In our view, the collinearity of the variables should be as low as possible because the interrelatedness among the different descriptors can result in highly unstable model. As can be seen in Table 2, the correlation coefficients are in general high, this demonstrates that the variables are highly correlated. It is possible to eliminate collinearity of molecular descriptors by orthogonalization process, previous explained.

The QSAR model obtained with the RDF descriptors (model 2) after the orthogonalization process is given below together with the statistical parameters.

$$\begin{aligned} C_{act} &= 0.172 + 1.569 \cdot \Omega^1 RDF080u - 0.673 \cdot \Omega^2 RDF010m - 2.044 \cdot \Omega^4 RDF085e \\ &\quad -1.993 \cdot \Omega^5 RDF045u + 2.146 \cdot \Omega^6 RDF065e - 2.710 \cdot \Omega^8 RDF120u \\ &\quad -1.451 \cdot \Omega^9 RDF140e + 2.609 \cdot \Omega^{10} RDF130e \\ N &= 148, \lambda = 0.628, D = 2.503, F_{exp} = 10.293, p < 10^{-5}, \%Class1 = 76.4, \%Class2 = 72.5 \end{aligned} \quad (2)$$

As result of the orthogonalization process of variables of the model 1, there were two variables that proved not to be statistically significant, and these were excluded in the final model (see Eqs. 1 and 2). These variables were, RDF025e and RDF125e. RDF025e is strongly correlated with RDF045u and RDF065e, containing 91.9 and 90.5% duplication. Also, this variable is collinear with RDF80u, having a coefficient of correlation of 87.4. RDF080u was the variable taken as the first orthogonal descriptor, and the rest of the descriptors were orthogonalized with respect to it. Therefore those variables that have a high correlation with it will have a high probability of leaving of the model. Similarly to the previous case, the RDF125e descriptor is highly correlated with RDF120u, RDF130e, and RDF140e, with coefficients of correlation of 95.9, 94.0, and 92.0, respectively. This is probably the main cause of the lack of RDF125e in the description of the carcinogenic activity.

The variables in the model (Eq. 2) encoded specific structure information. As can be seen, the variables in this model are related to the electronegativity ($\Omega^4 RDF085e$, $\Omega^6 RDF065e$, $\Omega^9 RDF140e$, $\Omega^{10} RDF130e$) and the atomic mass (Ω^2RDF010m) because these atomic properties were used for obtaining these descriptors. In a general way, the electronegativity is the power of an atom in a molecule to attract electrons. These electronegativity values are useful in determining the bond polarity for molecules. The bond polarity is a vector, pointing from the atom with lower electronegativity to that with a larger value. The separation of positive and negative charges causes an electric dipole moment. If the vector sum of bond polarities of a molecule is not zero, the molecule is said to have a dipole moment. The carcinogenic activity depends on the bond dipole moment. This result is explained by the fact that when the total polarity of molecules is increased, the hydrophobicity is lower, affecting the capacity of the molecule to permeate across biological membranes to reach intracellular targets like DNA in different tissues and organs in the body to elicit carcinogenicity.

The selection of this model as the best one obtained with the RDF descriptors was because it gave the lowest Wilk's statistic (λ), with the lowest number of parameters (variables), the biggest Fischer ratio (F), and best global classification for training set and test set (%Class1 and % Class2, respectively) of the models obtained with this family of descriptors. The classification results for the training set, using model 2 (Eq. 2), are illustrated in Table 3.

The model classified 77.4% (72/93) of chemicals with carcinogenic activity in the training set correctly and the 74.6% (41/55) of non-carcinogenic compounds, for a good global classification of 76.4% (113/148) (see also Table 3). The percentages of compounds unclassified in the training set are 7.5% (7/93) and 3.6% (2/55) for carcinogens and non-carcinogens, respectively. If these unclassified compounds are considered, the percentages of good classification are 79.6% (74/93) and 76.4% (42/55) for carcinogenic and non-carcinogenic chemicals, respectively, for an overall classification of 78.4% (116/148).

The percentages of *false positives* and *false negatives* in the training set were 21.8% (12/55) and 15.1% (14/93), respectively. *False positives* are those compounds without carcinogenic activity that are classified as active, and the *false negatives* are those compounds with carcinogenic activity that the model classified as inactive (see Table 3). From a practical point of view, in the development of the classification model, it is considered more important to avoid *false positives* because those are compounds that will be rejected for their wrongly predicted property and therefore they will never be evaluated experimentally, and their true carcinogenic activity would never be discovered. On the contrary, the *false negatives* compounds eventually will be detected. A classical problem in the modeling of rodent carcinogenicity is finding a major number of false positives with respect to false negatives, this being one of sources of error in the QSAR models. Contrera et al. [40] in the prediction of rodent carcinogenicity of 108 chemicals found 31% false positives and a

Table 3 Results of the classification of the compounds in the training set, according to the RDF model (Eq. 2)

No.	Compounds	Prob	Class	No.	Compounds	Prob	Class
Carcinogenic compounds							
1	[4-Chloro-6-(2,3-xylidino)-2-pyrimidinylthio]acetic acid	0.60	+	48	Diethylstilbestrol	0.48	NC
2	1-(2-Hydroxyethyl)-1-nitrosourea	0.86	+	49	Dihydrosafrole	0.59	+
3	1,1-Dimethylhydrazine	0.68	+	50	dl-Ethionine	0.59	+
4	1,2-Dibromoethane	0.41	−	51	Ethylene thiourea	0.75	+
5	1'-Hydroxysafrole	0.56	+	52	Formaldehyde	0.47	−
6	1-Nitroso-3,4,5-trimethylpiperazine	0.95	+	53	Formic acid 2-[4-(5-nitro-2-furyl)-2-thiazolyl] hydrazide	0.83	+
7	2,3,7,8-Tetrachlorodibenzo-*p*_dioxin	0.31	−	54	Glu-P-2	0.69	+
8	2,4,5-Trimethylaniline.HCl	0.79	+	55	Hydrazine	0.78	+
9	2,4,6-Trimethylaniline.HCl	0.81	+	56	Hydrazine sulfate	0.77	+
10	2,4-Dinitrotoluene, practical grade	0.57	+	57	IQ	0.76	+
11	2,5-Xylidine.HCl	0.79	+	58	Isoniazid	0.79	+
12	2-Acetylaminofluorene	0.89	+	59	Lead acetate, basic	0.49	NC
13	2-Amino-4-(5-nitro-2-furyl)thiazole	0.74	+	60	Melphalan	0.88	+
14	2-Hydrazino-4-(*p*-aminophenyl) thiazole	0.72	+	61	Methyl *tert*-butyl ether	0.67	+
15	2-Hydrazino-4-(*p*-nitrophenyl) thiazole	0.72	+	62	Metronidazole	0.83	+
16	2-Naphthylamine	0.62	+	63	N-[4-(5-Nitro-2-furyl)-2-thiazolyl]formamide	0.59	+
17	3-(5-Nitro-2-furyl)-imidazo(1,2-alpha) pyridine	0.86	+	64	*N*-[5-(5-Nitro-2-furyl)-1,3,4-thiadiazol-2-yl] acetamide	0.72	+
18	3,3',4,4'-Tetraaminobiphenyl.4HCl	0.72	+	65	Nitroso-2,3-dihydroxypropyl-2-oxopropylamine	0.91	+
19	3-Aminotriazole	0.75	+	66	Nitroso-2-oxopropylethanolamine	0.57	+
20	3-Nitro-3-hexene	0.76	+	67	Nitrosodibutylamine	0.69	+
21	4-Chloro-4'-aminodiphenylether	0.74	+	68	N-Nitrosodimethylamine	0.55	+
22	5-Azacytidine	0.72	+	69	*N*-Nitrosomethyl-2,3-dihydroxypropylamine	0.77	+
23	Acetaldehyde	0.47	-	70	*N*-Nitrosomorpholine	0.63	+
24	Acetamide	0.62	+	71	*N*-Nitroso-*N*-methylurea	0.75	+
25	Acetaminophen	0.52	NC	72	*N*-Nitrosopiperidine	0.54	+
26	AF-2	0.61	+	73	*N*-Nitrosopyrrolidine	0.69	+
27	Aldrin	0.41	−	74	*o*-Toluidine.HCl	0.79	+
28	alpha-1,2,3,4,5,6-Hexachlorocyclohexane	0.66	+	75	p,p'-DDE	0.75	+
29	Aramite	0.92	+	76	Phenacetin	0.58	+
30	Auramine-O	0.20	−	77	Phenobarbital, sodium	0.62	+
31	Benzidine	0.75	+	78	Propylthiouracil	0.60	+
32	beta-Propiolactone	0.50	NC	79	Safrole	0.66	+
33	Bis-(chloromethyl)ether	0.43	−	80	Sesamol	0.57	+
34	Bis-2-hydroxyethyldithiocarbamic acid,potassium	0.72	+	81	Sterigmatocystin	0.35	−
35	Caffeic acid	0.65	+	82	Streptozotocin	0.96	+
36	Captafol	0.22	−	83	Styrene oxide	0.53	+
37	Captan	0.49	NC	84	Tamoxifen citrate	0.85	+
38	Carbon tetrachloride	0.41	−	85	Thioacetamide	0.65	+
39	Chlorambucil	0.23	−	86	Thiouracil	0.64	+
40	Chloroform	0.43	−	87	Trichloroethylene	0.44	−
41	Chloromethyl methyl ether	0.50	NC	88	Trp-P-1 acetate	0.75	+
42	Ciprofibrate	0.87	+	89	Trp-P-2 acetate	0.74	+
43	Cyclophosphamide	0.88	+	90	Uracil	0.62	+
44	Dibromodulcitol	0.82	+	91	Urethane	0.77	+
45	Dibromomannitol	0.62	+	92	Vinyl acetate	0.69	+
46	Dichloroacetylene	0.44	−	93	Vinyl chloride	0.49	NC
47	Dieldrin	0.78	+				

Table 3 (continued)

No.	Compounds	Prob	Class	No.	Compounds	Prob	Class
Non-carcinogenic Compounds							
1	1,1,1-Trichloroethane(grado tec.)	0.47	−	29	FD & C blue no. 2	0.01	−
2	1,4-Dichlorobenzene	0.57	+	30	FD & C green no. 3	0.01	−
3	1-[(5-Nitrofurfurylidene)amino]hydantoin	0.31	−	31	FD & C yellow no. 5	0.05	−
4	1-Chloro-2,4-dinitrobenzene	0.44	−	32	FD & C yellow no. 6	0.09	−
5	2,3,4,5,6-Pentachlorophenol (Dowicide EC-7)	0.68	+	33	Flecainide acetate	0.00	−
6	5,5-Diphenylhydantoin	0.70	+	34	Fluoxetine.HCl	0.00	−
7	6-Dimethylamino-4,4-diphenyl-3-heptanone.HCl	0.02	−	35	Hexamethylenetetramine	0.51	NC
8	Aspirin	0.86	+	36	Isopropyl-*N*-(3-chlorophenyl) carbamate	0.23	−
9	Barium acetate	0.44	−	37	Maleic hydrazide	0.63	+
10	Benzoate, sodium	0.43	−	38	Methotrexate	0.04	−
11	Benzoguanamine	0.07	−	39	Misoprostol	0.10	−
12	Black PN	0.24	−	40	Monochloroacetic acid	0.55	+
13	C.I. pigment yellow 12	0.01	−	41	Nefiracetam	0.77	+
14	C.I. pigment yellow 16	0.01	−	42	*N* −Nitrosocimetidine	0.46	−
15	C.I. pigment yellow 83	0.05	−	43	Oxamyl	0.15	−
16	Cadmium acetate	0.44	−	44	Oxprenolol.HCl	0.18	−
17	Caffeine	0.40	−	45	Phenyl isothiocyanate	0.44	−
18	Chlorodifluoromethane	0.42	−	46	Prazepam	0.07	−
19	Chromium (iii) acetate	0.47	−	47	Propranolol.HCl	0.01	−
20	Compound 50-892	0.03	−	48	Quercetin dihydrate	0.28	−
21	Cyclohexylamine.HCl	0.60	+	49	Rutin trihydrate	0.00	−
22	Deltamethrin	0.00	−	50	Saccharin	0.63	+
23	Diazepam	0.05	−	51	Sotalol.HCl	0.26	−
24	Dimethylformamide	0.49	NC	52	Tetramethylthiuram disulfide	0.02	−
25	Diphenyl-*p*-phenylenediamine	0.07	−	53	Tilidine fumarate	0.70	+
26	Erythorbate, sodium	0.89	+	54	Trifluralin, technical grade	0.01	−
27	Estazolam	0.02	−	55	Urea	0.74	+
28	Ethynodiol diacetate	0.04	−				

+ Positive values are for compounds with carcinogenic activity
− Negative values are for compounds with non-carcinogenic activity
NC non-classified chemicals

24% false negative. Our model shows similar behavior, in general it classifies the carcinogen chemicals better, and this implies that there is minor amount of false negatives if these are compared with false positives. In our opinion, this is due to the use of an unbalanced dataset for building the training set. The model was made with a dataset that contain 93 carcinogenic and 55 non-carcinogenic compounds. Therefore, this approach contains more structural information for evaluation of carcinogenic chemicals than for evaluation of non-carcinogenic chemicals. In other words, the models are more trained in the prediction of carcinogenic compounds. Similar results for an unbalance dataset were found by other authors, for instance; Franke et al. [41] in the modeling of rodent overall carcinogenicity of aromatic amines, used 66 chemicals for the training set, 53 carcinogenic amines and 13 non-carcinogens. The discriminant model classifies 84.6% of non-carcinogenic compounds (11/13) correctly and 88.7% of carcinogenic amines (47/53). As can be seen, the best percent of classification is for carcinogenic chemicals. Recently, in a study for the prediction of chemicals that induces agranulocytosis, Gonzalez-Díaz et al. [36] used a training set of 151 chemicals belonging to a non-congeneric series. The model classifies 87.7% agranulocytosis-causing chemicals (50/57) correctly and 96.8% of non-toxic compounds (91/94). In this case the best prediction was for the non-toxic compounds (which have the major amount of structural information in the training set). Similar results were found with the other models developed in this paper (see Table 4). These results will be discussed later.

A cross-validation by the resubstitution technique (removing 25% of the training set) was carried out. The range of good classification was between 76.6 and 82.9%. The average of global classification of the model was 79.3% (see Table 5). The compounds unclassified were not considered in the previous calculations. As can be seen in the Table 5, the average of the global classification of the predicting series (CV-average) is higher than the global classification of the training set. Nevertheless QSAR models would usually be expected to perform better with the training set. However, different results might be due to the inappropriate selection of training and predicting series

Table 4 The statistical parameters of the linear regression models obtained for the nine kinds of descriptors

Descriptors	λ	D^2	F	$p<$	Class. training set			Class. test set		
					% $C1_{car}$	% $C1_{n\text{-}car}$	% Class1	% $C2_{car}$	% $C2_{n\text{-}car}$	% Class2
Constitutional	0.695	1.853	7.622	0.001	68.8	67.3	68.2	65.4	64.3	65.0
Molecular walk count	0.770	1.261	5.185	0.001	72.0	67.3	70.3	57.1	78.6	64.3
Galvez topological charge indices	0.776	1.220	5.016	0.001	72.0	58.2	66.9	73.1	57.1	67.5
2D autocorrelations	0.678	2.010	8.266	0.001	74.2	70.9	73.0	61.5	42.9	55.0
Randić molecular profiles	0.751	1.404	5.774	0.001	69.9	69.1	69.6	73.1	64.3	70.0
Geometrical	0.665	2.132	8.769	0.001	77.4	72.7	75.7	73.1	42.9	62.5
RDF	0.628	2.503	10.293	0.001	77.4	74.8	76.4	73.1	71.4	72.5
3D-MORSE	0.642	2.364	9.722	0.001	74.2	69.1	72.3	50.0	50.0	50.0
WHIM	0.697	1.832	7.536	0.001	77.4	72.7	75.7	70.4	35.7	58.5

Note: % $C1_{car}$, % $C1_{n\text{-}car}$ and % class 1 are per cent of classification of carcinogenic chemicals, non-carcinogenic chemicals and global classification for training set, respectively
% $C2_{car}$, % $C2_{n\text{-}car}$ and % class 2 are per cent of classification of carcinogenic chemicals, non-carcinogenic chemicals and global classification for test set, respectively

(CVs) of chemicals. This problem could be solved by using a better way for selecting both training and CVs of chemicals, for example by using Cluster analysis.

The most important criterion for the quality of the discriminant model is based on the statistics for the test set. Eq. (2) classified the 73.1% (19/26) and 71.4% (10/14) of carcinogenic and non-carcinogenic compounds, respectively, correctly. The global classification was 72.5%. The percentage of *false negative* and *false positive* compounds was 23.1% (6/26) and 28.6% (4/14), respectively. In Table 6 the classification of compounds in the external test set is shown.

Comparison with other approaches

As we pointed out previously, one of the objectives of the current work is to compare the reliability of the RDF descriptors for describing the property under study with other descriptors and methods. Consequently, we have developed eight other models using the same dataset that was included in the RDF QSAR model. The results obtained with Constitutional, Molecular walk counts, Galvez topological charge indices, 2D autocorrelations, Randić molecular profiles, Geometrical, RDF, 3D-MORSE, and WHIM descriptors are given in Table 4. In addition, the descriptors used and their meaning are given in Tables 7 and 8. With the objective of compare the statistic parameters of these models, all have the same number of variables.

As can be seen, there are differences concerning the discrimination of the groups (carcinogens and non-carcinogens) given by the Wilk's statistic for these models compared with that obtained with the RDF descriptors. The value of Wilk's lambda indicates the unexplained variance. For this reason, the best model obtained is with the RDF

Table 5 Classification matrices and accuracy for training and re-substitution cross-validation

CV1				CV2			
Class.	Percent	Carc.	Non-carc.	Class.	Percent	Carc.	Non-carc.
Carc.	78.6	55	15	Carc.	91.3	63	6
Non-carc.	75.6	10	31	Non-carc.	66.7	14	28
Total	77.5	65	46	Total	82.0	77	34
CV3				CV4			
Class.	Percent	Carc.	Non-carc.	Class.	Percent	Carc.	Non-carc.
Carc.	84.3	59	11	Carc.	74.3	52	18
Non-carc.	80.5	8	33	Non-carc.	80.5	8	33
Total	82.9	67	44	Total	76.6	60	51
CV-Average				Training set			
Class.	Percent	Carc.	Non-carc.	Class.	Percent	Carc.	Non-carc.
Carc.	85.1	57	13	Carc.	79.6	74	19
Non-carc.	70.5	10	31	Non-carc.	76.4	13	42
Total	79.3	67	44	Total	78.4	87	61

Carc: Number of carcinogenic compounds;
Non-carc: Number of non-carcinogenic compounds

Table 6 Results of the classification of the compounds in the external prediction set, according to model 2 (Eq. 2)

No.	Compounds	Prob	Class	No.	Compounds	Prob	Class
Carcinogenic compounds							
1	1,2-Dimethylhydrazine.2HCl	0.75	+	14	Doxylamine succinate	0.69	+
2	2,2,2-Trifluoro-N-[4-(5-nitro-2- furyl)_2_thiazolyl] acetamide	0.24	−	15	Glu-P-1	0.82	+
3	2,4-Diaminotoluene.2HCl	0.87	+	16	Hexachlorobenzene	0.65	+
4	2,5-Dimethoxy-4'-aminostilbene	0.17	−	17	MeA-alpha-C acetate	0.70	+
5	2-Hydrazino-4-(5-nitro-2-furyl)thiazole	0.93	+	18	MeIQx	0.43	−
6	Azobenzene	0.57	+	19	*N*-[4-(5-Nitro-2-furyl)-2-thiazolyl]acetamide	0.62	+
7	Benzo(a)pyrene	0.74	+	20	*N*-Hydroxy-2-acetylaminofluorene	0.09	−
8	Butylated hydroxyanisole	0.25	−	21	Nitrobenzene	0.48	NC
9	Catechol	0.83	+	22	*N*-Nitroso-2,3-dihydroxypropyl-2-hydroxypropylamine	0.95	+
10	Chrysazin	0.59	+	23	*N*-Nitrosoallyl-2-oxopropylamine	0.60	+
11	D & C red no. 5	0.72	+	24	*o*-Phenylenediamine.2HCl	0.87	+
12	DDT	0.83	+	25	PhIP.HCl	0.35	−
13	Diallylnitrosamine	0.68	+	26	Piperonyl butoxide	1.00	+
Non-carcinogenic compounds							
27	1,2-dichloroethane	0.46	−	34	FD & C red no. 3	0.04	−
28	2,4,5-Trichlorophenoxyacetic acid	0.39	−	35	Fenvalerate	0.00	−
29	Adipamide	0.84	+	36	Praziquantel	0.33	−
30	Benzalazine	0.77	+	37	Sorbic acid	0.59	+
31	Dichlorodifluoromethane	0.40	−	38	Toluene diisocyanate	0.54	+
32	Etodolac	0.02	−	39	Trichlorofluoromethane	0.39	−
33	FD & C blue no. 1	0.00	−	40	Triprolidine HCl monohydrate	0.10	−

+ Positive values are for compounds with carcinogenic activity
− Negative values are for compounds with non-carcinogenic activity
NC non-classified chemicals

descriptors. These are also related with the lower λ (λ = 0.628) and therefore with the best discriminatory power. On the other hand, the squared Mahalanobis distance of the RDF QSAR model is higher than the rest of the models. This parameter is associated with the distance between the centroids of the respective groups, showing that the RDF model possesses the best percent of classification. Consequently, the RDF model has the higher Fisher ratio, showing its statistical significance. We have until now compared the methods, evaluating their fitting quality, but what is the behavior of the model for prediction?

The RDF model presents a percentage of good classification for training set of 76.4%, although apparently this classification seems modest if it is compared with other biological activity, for instance anti-inflammatory activity and herbicide properties, [14, 15] it can be considered a good classification for the carcinogenic activity of the heterogeneous series of compounds. [40, 42] In this sense, other families of descriptors such as WHIM, Geometrical, and 2D autocorrelations present similar percent of global classification of the training set (75.7, 75.7 and 73.0%, respectively) while they have worse statistical parameters, as discussed above. A superficial analysis of the problem would conclude that these models present a similar prediction power for screening of new molecules for their

Table 7 The descriptors used for building each model

Descriptors	Variables
Constitutional	Ss, nHM, nR03, nF, Me, nO, nH, nR04
Molecular walk count	SRW06, SRW02, MWC09, MWC03, MWC02, MWC01, MWC04, SRW09
Galvez topological charge indices	JGI8, GGI8, JGI1, JGI2, GGI5, JGI7, GGI1, JGI10
2D autocorrelations	MATS2e, ATS8v, GATS8e, ATS8m, ATS5m, ATS4p, ATS6v, ATS7v
Randić molecular profiles	DP20, DP01, DP04, DP10, DP02, SHP2, DP12, SP01
Geometrical	DDI, MAXDN, SPAN, J3D, G(N..I), G(N..S), G1, G2
RDF	RDF045u, RDF080u, RDF120u, RDF010m, RDF065e, RDF085e, RDF130e, RDF140e
3D-MORSE	Mor05u, Mor09m, Mor23u, Mor08u, Mor15u, Mor20u, Mor17u, Mor11u
WHIM	L3e, E3s, E1e, P2e, P1u, G2v, G2p, E3u

Table 8 Symbols of the descriptors used in the models and their definitions

Symbols	Descriptor definition
Ss	Sum of Kier–Hall electrotopological states
Me	Mean atomic Sanderson electronegativity (scaled on carbon atom)
nH	Number of hydrogen atoms
nO	Number of oxygen atoms
nHM	Number of heavy atoms
nR03	Number of 3-membered rings
nR04	Number of 4-membered rings
MAXDN	Maximal electrotopological negative variation
MWC01	Molecular walk count of order 01 (number of non-H bonds, nBO)
MWC02	Molecular walk count of order 02
MWC03	Molecular walk count of order 03
MWC04	Molecular walk count of order 04
MWC09	Molecular walk count of order 09
SRW02	Self-returning walk count of order 02 (twice the number of non-H bonds)
SRW06	Self-returning walk count of order 06
SRW09	Self-returning walk count of order 09
ATS5m	Broto–Moreau autocorrelation of a topological structure—lag 5/weighted by atomic masses
ATS8m	Broto–Moreau autocorrelation of a topological structure—lag 8/weighted by atomic masses
ATS6v	Broto–Moreau autocorrelation of a topological structure—lag 6/weighted by atomic van der Waals volumes
ATS7v	Broto–Moreau autocorrelation of a topological structure—lag 7/weighted by atomic van der Waals volumes
ATS8v	Broto–Moreau autocorrelation of a topological structure—lag 8/weighted by atomic van der Waals volumes
ATS4p	Broto–Moreau autocorrelation of a topological structure—lag 4/weighted by atomic polarizabilities
MATS2e	Moran autocorrelation–lag 2/weighted by atomic Sanderson electronegativities
GATS8e	Geary autocorrelation—lag 8/weighted by atomic Sanderson electronegativities
GGI1	Topological charge index of order 1
GGI5	Topological charge index of order 5
GGI8	Topological charge index of order 8
JGI1	Mean topological charge index of order1
JGI2	Mean topological charge index of order2
JGI7	Mean topological charge index of order7
JGI8	Mean topological charge index of order8
JGI10	Mean topological charge index of order10
DP01	Molecular profile no. 01
DP02	Molecular profile no. 02
DP04	Molecular profile no. 04
DP10	Molecular profile no. 10
DP12	Molecular profile no. 12
DP20	Molecular profile no. 20
SP01	Shape profile no. 01
SHP2	Average shape profile index of order 2
J3D	3D-Balaban index
DDI	D/D index
G1	gravitational index G1
G2	Gravitational index G2 (bond-restricted)
SPAN	Span R
G(N..S)	Sum of geometrical distances between N..S
G(N..I)	Sum of geometrical distances between N..I
Mor05u	3D-MoRSE—signal 05/unweighted
Mor08u	3D-MoRSE—signal 08/unweighted
Mor11u	3D-MoRSE—signal 11/unweighted
Mor15u	3D-MoRSE—signal 15/unweighted
Mor17u	3D-MoRSE—signal 17/unweighted
Mor20u	3D-MoRSE—signal 20/unweighted
Mor23u	3D-MoRSE—signal 23/unweighted

Table 8 (continued)

Symbols	Descriptor definition
Mor09m	3D-MoRSE—signal 09/weighted by atomic masses
P1u	1st component shape directional WHIM index/unweighted
E3u	3rd component accessibility directional WHIM index/unweighted
G2v	2st component symmetry directional WHIM index/weighted by atomic van der Waals volumes
L3e	3rd component size directional WHIM index/weighted by atomic Sanderson electronegativities
P2e	2nd component shape directional WHIM index/weighted by atomic Sanderson electronegativities
E1e	1st component accessibility directional WHIM index/weighted by atomic Sanderson electronegativities
G2p	2st component symmetry directional WHIM index/weighted by atomic polarizabilities
E3s	3rd component accessibility directional WHIM index/weighted by atomic electrotopological states
RDF045u	Radial Distribution Function-4.5/unweighted
RDF080u	Radial Distribution Function-8.0/unweighted
RDF120u	Radial Distribution Function-12.0/unweighted
RDF010m	Radial Distribution Function-1.0/weighted by atomic masses
RDF025e	Radial Distribution Function-2.5/weighted by atomic Sanderson electronegativities
RDF065e	Radial Distribution Function-6.5/weighted by atomic Sanderson electronegativities
RDF085e	Radial Distribution Function-8.5/weighted by atomic Sanderson electronegativities
RDF125e	Radial Distribution Function-12.5/weighted by atomic Sanderson electronegativities
RDF130e	Radial distribution function-13.0/weighted by atomic Sanderson electronegativities
RDF140e	Radial distribution function-14.0/weighted by atomic Sanderson electronegativities

carcinogenic activity. In this case a deep analysis is necessary for determining which model is the best.

The prime aim in developing a QSAR is that it can be used for predictive purposes. It is therefore important that the statistics given with the QSAR provide an indication of its predictivity. This was achieved by the use of external validation (the same test set was used for all models). Table 4 shows the percent of good classification for the different methodologies. Surprisingly, the three models referred above (WHIM, Geometrical, and 2D-autocorrelations, descriptors) fail when they have to predict the external set. The percentages of good classification are not as good as for the RDF QSAR model (see Table 4). The results obtained with the model using the RDF descriptors are better than the rest of the methodologies used, which are unable to make good predictions, apart from the important statistic parameters of quality, such as Wilk's lambda (ë), squared Mahalanobis distance (D^2) the Fisher ratio (F) and the good classification of the training set (Class1).

For the reasons explained, we feel the RDF descriptors can be used for predicting the carcinogencity activity of new chemicals, thus contributing to the design and development of safe drugs, saving substantial amounts of money, time and animals.

Concluding remarks

The prediction of carcinogenicity has been a goal of pharmaceutical companies due to the importance of this property during the drug-development process. For this reason, several in silico methods have been used in order to predict this property in the early stage of drug development and some of them have become important tools for selecting new drug candidates. In this study, RDF descriptors were used to predict the carcinogenic activity for a non-congeneric series of compounds. The procedure has showed that a good discriminant model can be obtained using the radial distribution function unweighted and weighted by the atomic Sanderson electronegativity and atomic masses. The linear model developed in the current work is easily calculated and suitable for the rapid prediction of carcinogenicity, and the external validation and cross-validation of the final model support this claim. This suggests that the present method should be regarded as the one of choice for lead-optimization programs in the drug discovery process.

Acknowledgements The authors would like to express gratitude to Dr. Romualdo Benigni and Dr. Johann Gasteiger for sending us useful information for the development of this work.

References

1. Kennedy T (1997) Drug Disc Today (DDT) 2:436–444
2. Modi S (2003) Drug Discov Today (DDT) 8:621–623
3. van de Waterbeemd H, Gifford E (2003) Nat Rev Drug Discov 2:192–204
4. Dearden JC (2003) J Comput-Aided Mol Des 17:119–127
5. Benigni R, Giuliani A (2003) Bioinformatics 19:1194–1200
6. Haseman J, Melnick R, Tomatis L, Huff J (2001) J Food Chem Toxicol 39:739–744
7. Fung VA, Barrett JC, Huff J (1995) Environ Health Perspect 103:680–683
8. Huff J (1993) Environ Health Perspect 100:201–210
9. Richard AM (1998) Toxicol Lett 102–103:611–616
10. Richard AM, Beingni R (2002) SAR QSAR Envirom Res 13:1–19

11. Morales AH, Cabrera MA, Combes RD, Gonzalez MP (2005) Current Comp Aid Drug Des 3:237–255
12. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim, Germany
13. González MP, Terán C, Fall Y, Teijeira M, Besada P (2005) Bioorg Med Chem 13:601–608
14. Yan A, Gasteiger J (2003) J Chem Inf Comput Sci 43:429–434
15. Wegner JK, Frohlich H, Zell A (2004) J Chem Inf Comput Sci 44:931–939
16. Gasteiger J, Schuur J, Selzer P, Steinhauer L, Steinhauer V, Fresenius J (1997) Anal Chem 359:50–59
17. Hemmer MC, Steinhauer V, Gasteiger J (1999) J Vibra Spectros 19:151–164
18. Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V (1996) J Chem Inf Comput Sci 36:1030–1037
19. Gold LS, Manley NB, Slone TH, Rohrbach L (1999) Environ Health Perspect Suppl 107:527–532
20. Enslein K (1999) HDI Comput Toxicol NEWS 22
21. González MP, González HD, Molina RR, Cabrera MA, Ramos de Armas R (2003) J Chem Inf Comput Sci 43:1192–1199
22. González MP, Días LC, Morales AH, Rodríguez YM, Gonzaga de Oliveira L, Gómez LT, González HD (2004) Bioorg Med Chem 12:4467–4475
23. McFarland JW, Gans DJ (1995) Cluster Significance Analysis. In: Manhnhold R, Krogsgaard-Larsen P, Timmerman H (eds) Method and Principles in Medicinal Chemistry, vol 2. In: van Waterbeemd H (ed) Chemometric methods in molecular design. VCH, Weinheim, pp 295–307
24. Johnson RA, Wichern DW (1988) Applied MultiVariate Statistical Analysis. Prentice-Hall, New York
25. Michael JS, Dewar E, Zoebisch G, Eamonn F, Stewart JP (1985) J Am Chem Soc 107:3902–3909
26. Stewart JJP (1990) MOPAC manual, 6th edn. Frank J Seiler Research Laboratory, US Air Force academy, Colorado Springs, CO, p 189
27. Todeschini R, Consonni V, Pavan M (2002) Dragon Software, version 2.1
28. StatSoft Inc (2002) STATISTICA 6.0, version 6.0
29. Randić M (1991) J Chem Inf Comput Sci 31:311–320
30. Randić M (1991) New J Chem 15:517–525
31. Randić M (1991) J Mol Struct (Theochem) 233:45–59
32. Lučić B, Nikolić S, Trinajstić N (1995) J Chem Inf Comput Sci 35:532–538
33. Kowalski RB, Wold S (1982) Pattern recognition in chemistry. In: Krishnaiah PR, Kanal LN (eds) Handbook of statistics. North Holland Publishing Company, Amsterdam 673–697
34. Hawkins DM (2004) J Chem Inf Comput Sci 44:1–12
35. González-Díaz H, Ramos R, Molina RR (2003) Bioinformatics 19:2079–2087
36. González-Díaz H, Marrero Y, Hernández I, Bastida I, Tenorio E, Nasco O, Uriarte E, Castañedo NC, Cabrera-Pérez MA, Aguila E, Marrero O, Morales A, González MP (2003) Chem Res Toxicol 16:1318–1327
37. González-Díaz H, Bastida I, Castañedo N, Nasco O, Olazabal E, Morales A, Serrano HS, Ramos R (2004) Bull Math Biol 66:1285–1311
38. González-Díaz H, Uriarte E, Ramos R (2005) Bioorg Med Chem 13:323–331
39. Klein DJ, Randić M, Babic D, Lucic B, Nikolic S, Trinajstic N (1997) Int J Quantum Chem 63:215–221
40. Contrera JF, Matthews EJ, Benz RD (2003) Regul Toxicol Pharmacol 38:243–259
41. Franke R, Gruska A, Giuliani A, Benigni R (2001) Carcinogenesis 22:1561–1571
42. Benigni R, Richard AM (1996) Mutat Res 371:29–46