Available online at www.sciencedirect.com
ScienceDirect

Journal of Molecular Graphics and Modelling 25 (2006) 275–288

Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Automatic extraction of structural alerts for predicting chromosome aberrations of organic compounds

Ernesto Estrada *, Enrique Molina

Complex Systems Research Group, RIAIDT, Edificio CACTUS, University of Santiago de Compostela, Santiago de Compostela 15782, Spain

Received 1 September 2005; accepted 8 January 2006
Available online 17 February 2006

Abstract

We use the topological sub-structural molecular design (TOPS-MODE) approach to formulate structural alert rules for chromosome aberration (CA) of organic compounds. First, a classification model was developed to group chemicals as active/inactive respect to CA. A procedure for extracting structural information from orthogonalized TOPS-MODE descriptors was then implemented. The contributions of bonds to CA in all the molecules studied were then generated using the orthogonalized classification model. Using this information we propose 22 structural alert rules which are ready to be implemented in expert systems for the automatic prediction of CA. They include, among others, structural alerts for *N*-nitroso compounds (ureas, urethanes, guanidines, triazines), nitro compounds (aromatic and heteroaromatic), alkyl esters or phosphoric acids, alkyl methanesulfonates, sulphonic acids and sulphonamides, epoxides, aromatic amines, azaphenanthrene hydrocarbons, etc. The chemico-biological analysis of some of the structural alerts found is also carried out showing the potential of TOPS-MODE as a knowledge generator.

Keywords: TOPS-MODE; Topological descriptors; Knowledge-generation; Clastogenicity; Chromosome aberration; Structure–toxicity relationships; QSAR

1. Introduction

In QSAR analysis a quantitative model is used to predict the biological response of a chemical based on a series of molecular descriptors or physicochemical properties [1]. However, the structural information contained in such descriptors or properties is encrypted [2] in a way that does not allow the extraction of structural rules to form a *knowledge base* similar to that provided by human expertise [3]. In the case of toxicological assessment of chemicals these knowledge bases are the heart of expert systems, such as DEREK [4] and TOPKAT [5], used to evaluate the toxicological profile of chemicals [6]. One of these toxicological endpoints which is of relevant importance is the chromosome aberration or clastogenicity produced by chemicals. Chromosome aberrations (CA) are DNA changes generated by different repair mechanisms of DNA double strand breaks, which are microscopically visible [7]. They are consequences of human exposure to ionising radiation or to mutagenic chemicals [8–11]. The frequencies of CA in peripheral lymphocytes show a positive correlation with the later onset of cancer in humans [7].

The necessity for the automatic generation of structural alerts for predicting CA and other toxicological endpoints is evident. On one hand, classical QSARs permit the classification of chemicals as clastogenic/nonclastogenic but their information cannot be easily incorporated on the existing expert systems due to the cryptic nature of the variables included in such models [2]. On the other hand, the traditional method for extracting knowledge from human expertise requires a great amount of (available) information about a set of chemicals permitting the expert their generalization. However, the rate of producing new chemical entities overtakes the rate of their toxicological profile evaluation. Thus a method that permits to extract knowledge from the minimum information available about a series of chemicals is necessary to keep expert systems updated. In this sense, an expert system can be considered as *knowledge archive* where a collection of knowledge is expressed using some formal representation language. An *automatic knowledge generator* is a methodology that will provide new structural alerts to the knowledge archive in a cyclic way keeping it updated. In previous works [12,13] we have shown that the so-called topological sub-structural molecular design (TOPS-MODE) approach [14–19] represents a useful platform for the automatic generation of toxicological structural alerts. In these works a general strategy for

* Corresponding author.
E-mail address: estrada66@yahoo.com (E. Estrada).

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.01.002

knowledge flow concerning skin sensitization based on the combined use of TOPS-MODE and DEREK expert system was proposed [12,13].

The main purpose of the current work is to generate structural alert rules that permit the identification of CA in chemical compounds using information coded in their molecular structure. Thus, we develop a classification model using the TOPS-MODE approach, which allows to calculate the contribution of each part of a molecule to the activity under study. Using this information we identify structural regions responsible for the clastogenic activity of chemicals and transform this information into structural alert rules which are ready to be implemented in expert systems such as DEREK.

2. Data set

A data set of 383 organic compounds compiled by Serra et al. was used for the purposes of the current study [20]. These compounds were selected among those reported on the *Compilation of Chromosomal Mutation Test Data* containing tests carried out by the *National Drug and Food safety Laboratory and the First Laboratory of the mutational Genetics Department of the Safety and Biotesting Research Center in Japan* [21]. These compounds were tested at two different times of exposure, mainly 24 and 48 h, in cultured Chinese hamster lung cells. After exposition, cells were processed by standard methods and chromosomal aberrations were identified. Compounds were classified as positive if there were 10% or greater aberrant cells and negative if there were 5% or less aberrant cells. Compounds classified as "equivocal" due to their percentage of aberrant cells (5–10% aberrant cells) were not included in this study as well as they were not considered in Serra et al.'s work. From this original data set three compounds could not be included in the current study as they have macromolecular structures, such as polymeric one (compounds 161, 185, 267 in Serra et al.'s work [20]). Three compounds in the original data set were salts of other compounds in the data set. For instance, compound 40 (in Serra et al. 's paper [20]) is aniline–HCl and compound 141 is aniline. Compound 11 is the sodium salt of the L-glutamic acid and compound 200 is L-glutamic acid. Finally, compound 324 is the salt of 115. In all cases salts were excluded from our data set. There are other five pairs of compounds which were geometric isomers distinguished neither by our approach nor by descriptors used by Serra et al. [20]. They are: 163/355, 42/136, 90/146, 348/361 and 166/175. In all cases one of the compounds in each pair was eliminated from our data set. Consequently, our data set is formed by 372 organic compounds including known carcinogens, drugs, food additives, agrochemicals, cosmetic materials, medicinal products, and household materials.

This data set was divided into two subsets, one containing 216 compounds (100 clastogenics and 116 nonclastogenic) was used as a training set for developing the classification model. The other formed by 156 compounds (11 clastogenic and 145 nonclastogenic) was used as a prediction set. Our main objective is to extract as much structural information as possible from this data set in order to formulate structural alerts for clastogenicity. Consequently, we keep the minimum number of clastogenic compounds out of the training set. In fact, we selected only those compounds used by Serra et al. [20] as the prediction set for the *k*-nearest neighbour model, i.e., we do not use any cross-validation set. In that work, however, the number of nonclastogenic compounds in the training sets is very much higher than the number of clastogenic ones. For instance, for the *k*-NN model development they used 245 nonclastogenic compounds and 101 clastogenic compounds and for SVM model development the training set consisted on 218 nonclastogenic and only 90 clastogenic compounds. Here we preferred to have a more compensated training set having approximately the same number of clastogenic and nonclastogenic compounds. Consequently, we selected at random several nonclastogenic compounds originally in the training set to be in the prediction set. This produced a training set having 116 nonclastogenic and 100 clastogenic compounds and the prediction set was finally conformed by these compounds plus those originally in the prediction set [20].

3. Methodology

3.1. The TOPS-MODE approach

In the last 10 years we have developed an approach to QSAR/QSPR based on the use of spectral moments of the bond matrix as molecular descriptors. It is known as TOPS-MODE approach, which is the acronym of topological substructural molecular descriptors/design [14–19]. TOPS-MODE approach is based on the calculation of spectral moments of molecular bond matrices appropriately weighted to account for hydrophobic, electronic and steric molecular features. Spectral moments are the trace of the *k*th power of a matrix, i.e., the sum of all the main diagonal entries of such matrices [14–16].

A bond matrix is a square symmetric matrix in which non-diagonal entries are ones or zeroes if the corresponding bonds have a common atom or not, respectively [22]. These matrices represent the molecular skeleton without taking into account hydrogen atoms. Bonds weights are placed as diagonal entries of such matrices and represent quantitative contributions to different physicochemical properties. Among bond weights currently in use in our approach we have standard bond distance (SD), standard bond dipole moments (DM), hydrophobicity (H) [23], polar surface area (PS) [24], polarizability (Pol) [25], molar refractivity (MR) [25], van der Waals radii (vdW) [26], and Gasteiger–Marsilli charges (Ch) [27].

The starting point for our approach is to calculate TOPS-MODE descriptors of the different types, e.g., H, PS, Pol, MR, vdW, and Ch, for the series of molecules under study. Then, we develop a quantitative model describing the property under study in term of the spectral moments. In general this model can be of the following form:

$$P = b_0 + \sum_{j=1}^{L} b_j \mu_j \qquad (1)$$

where P is the property under study, b_j the coefficients of the quantitative model (linear regression or discriminant analysis) and b_0 is the error.

The jth spectral moment of the bond matrix can be expressed as a sum of bond moments, which are simply the corresponding entries of the jth power of the bond matrix:

$$\mu_j = \sum_i^m \mu_j(i) \tag{2}$$

where $\mu_j(i)$ is the bond moment of the ith bond in a molecule with m bonds. Then, model (2) can be written as

$$P = b_0 + \sum_{j=1}^{L} b_j \sum_{i=1}^{m} \mu_j(i) = b_0 + \sum_{i=1}^{m} \sum_{j=1}^{L} b_j \mu_j(i) \tag{3}$$

where the right-hand side in (3) represents the contribution of bond i to the property P and will be called ‘‘bond contribution’’ and represented by $P(i)$:

$$P(i) = \sum_{j=1}^{L} b_j \mu_j(i) \tag{4}$$

and the property P is expressed as an additive function of bond contributions:

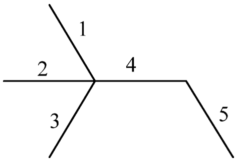
$$P = \sum_{i=1}^{m} P(i) \tag{5}$$

3.2. Calculation of bond contributions

Bond contributions are numeric characterization of such bonds which permit to identify some groups or regions of a molecular framework which can be responsible for a property/activity [28]. By carefully analyzing similar regions in different molecules we can obtain general rules about the contributions of molecular fragments to a particular property/activity. They are based on the sub-structural nature of TOPS-MODE. This procedure consists in transforming a QSPR or QSAR model into a bond additive scheme in which you can calculate the property under study as the sum of bond contributions for a molecule. Here we give a simple example of how to calculate bond contributions from a QSPR model obtained by using the TOPS-MODE approach. This model describing molar refraction of alkanes is given below [14]:

$$\text{MR (cm}^3) = 5.703 + 5.506\mu_0 - 0.329\mu_2 + 0.193\mu_3 - 0.033\mu_4,$$
$$N = 69, \quad r = 0.9999, \quad s = 0.05, \quad F = 168\,569 \tag{6}$$

Using model (6) for the molar refraction of alkanes we will calculate bond contributions for the molecule of 2,2-dimethylbutane with the bond numbering given below:



The total spectral moments can be expressed as sum of bond spectral moments $\mu_k(i)$ of the form: $\mu_k = \sum_i \mu_k(i)$, where i are the different bonds in the molecule. The bond spectral moments for the bonds of this molecule are as follows:

Bond (i)	$k = 0$	$k = 2$	$k = 3$	$k = 4$
1	1	3	6	22
2	1	3	6	22
3	1	3	6	22
4	1	4	6	24
5	1	1	0	4

Now we will substitute these expressions into the QSPR model obtaining bond molar refractions as exemplified for bond 1:

$$\text{MR}(1) = 5.506 \times 1 - 0.329 \times 3 + 0.193 \times 6 - 0.033 \times 22$$
$$= 4.95\ \text{cm}^3$$

In a similar way the contributions of the other bonds are obtained giving the values of MR(4) = 4.42 cm^3 and MR(5) = 5.04 cm^3. It is clear that the sum of these bond contributions plus the intercept of the QSPR model (6) gives the value of the molar refraction of the molecule: 30.026 cm^3.

3.3. Orthogonalization of TOPS-MODE descriptors

One of the inherent characteristic of the TOPS-MODE approach is that spectral moments are collinear among them. This means that there is redundancy in the information contained in any pair of collinear descriptors. Two descriptors are called collinear if they have a significant linear correlation between them as measured by the linear correlation coefficient. The main drawback of collinearity from the point of view of a QSAR model is that of the stability of the coefficients in the linear regression model. This introduces a difficulty in interpreting the linear models obtained with collinear variables because the sign and magnitude of the coefficients in the regression model can be affected by the removal or introduction of a new variable to the model. In the case of the TOPS-MODE approach this can be traduced into false interpretation of bond contributions because the magnitude and sign of them can be falsified by the effect produced by the existence of collinear variables in the model. Consequently, we have implemented by first time the Randić method of orthogonalization [29–31] to eliminate the collinearities among the TOPS-MODE variables. In doing so, we have developed a new approach to extract the information contained in these variables after orthogonalization.

The Randić method of orthogonalization has been described in details in several publications [29–33]. Thus, we will give only a general overview here. The first step in orthogonalizing the molecular descriptors 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 linear discriminant analysis. The first variable ($v1$) is taken as the first orthogonal descriptors ${}^1\Omega(v1)$ and the second one is orthogonalized respect to it by taking the residual of its correlation with ${}^1\Omega(v1)$. The process is repeated until all variables are completely orthogonalized and the orthogonal variables are then used to obtain the new model. Let consider the following QSAR/QSPR model: $P = a_0 + a_1X_1 + a_2X_2 + a_3X_3$, then the orthogonalization of the independent variables is carried out as follows:

(i) orthogonalizes X_1: $\Omega(X_1) = X_1$,
(ii) orthogonalizes X_2 respect to $X_1 : \Omega(X_2) = X2 - \hat{X}_2$, where $\hat{X}_2 = b_0 + b_1X_1$,
(iii) orthogonalizes X_3 respect to $\Omega(X_2) : \Omega(X_3) = X_3 - \hat{X}_3$, where $\hat{X}_3 = b'_0 + b'_1\Omega(X_2)$,
(iv) orthogonalizes $\Omega(X_3)$ respect to $X_1 : {}^2\Omega(X_3) = \Omega(X_3) - \hat{\Omega}(X_3)$, where $\hat{\Omega}(X_3) = b''_0 + b''_1X_1$.

The orthogonalized variables are: X_1, $\Omega(X_2)$ and ${}^2\Omega(X_3)$ and the coefficients in steps (ii)–(iv) are obtained by linear regression analysis.

In order to extract the information contained in the orthogonalized descriptors, i.e., bond contributions, we implemented the following iterative procedure:

(i) calculate bond contributions to $\Omega(X_1)$: $C[\Omega(X_1)] = C(X_1)$,
(ii) calculate bond contributions to $\Omega(X_2)$: $C[\Omega(X_2)] = C(X_2) - b_1C(X_1)$,
(iii) calculate bond contributions to $\Omega(X_2) : C[\Omega(X_3)] = C(X_3) - b'_1C[\Omega(X_2)]$,
(iv) calculate bond contributions to ${}^2\Omega(X_3) : C[{}^2\Omega(X_3)] = C[\Omega(X_3)] - b''_1C(X_1)$.

The final bond contributions are $C(X_1)$, $C[\Omega(X_2)]$ and $C[{}^2\Omega(X_3)]$. This procedure represents the extraction of the information contained into a bond contribution of a variable which is duplicated by the other variables in the model.

4. Classification model

A linear discriminant model was developed using our training data set of 216 compounds. The model contains 14 TOPS-MODE descriptors accounting for hydrophobic, electronic and steric features of molecules. The model classifies correctly 86% of the total number of compounds in the training set (86.9% of good classification for nonclastogenic and 84.9% for clastogenic compounds). In the test set the percentage of good classification is 82.8% (82.9% for nonclastogenic and 81.8% for clastogenic).

The classification model obtained is given below together with the statistical parameters of the linear discriminant analysis, where λ is the Wilks' statistics, D^2 is the squared Mahalanobis distance and F is the Fisher ratio:

$$\begin{aligned}\text{Class}(1) = {} & 0.58435\mu_1^{\text{PS}} - 0.00121\mu_5^{\text{vdW}} + 0.11879\mu_2^{\text{Ch}} \\ & - 0.1696\mu_2^{\text{PS}} + 0.01414\mu_3^{\text{PS}} - 4.668 \times 10^{-4}\mu_4^{\text{PS}} \\ & + 0.00215\mu_4^{\text{MR}} + 5.397 \times 10^{-6}\mu_5^{\text{PS}} - 1.633 \\ & \times 10^{-4}\mu_5^{\text{MR}} - 2.403 \times 10^{-5}\mu_8^{\text{H}} - 0.74912\mu_2^{\text{Pol}} \\ & + 1.81602\mu_1^{\text{Pol}} - 0.01532\mu_5^{\text{Ch}} + 0.11233\mu_5^{\text{Ch}} \\ & + 0.48018,\end{aligned}$$

$$\text{Wilks} - \lambda = 0.629;\ F(14, 194) = 8.142;\ D^2 = 2.353;\ p < 0.0000$$

This model shows the best performance in predicting clastogenicity in both the training and test sets among all the models generated using TOPS-MODE descriptors and linear discriminant analysis. In Table 1 we give the classification of all compounds used in the training and test sets using this model.

Then, we proceed to orthogonalize the variables in this model in order to eliminate any collinearity present among the variables included in the model. Following the Randić orthogonalization procedure previously described we generate the following orthogonal classification model:

$$\begin{aligned}\text{Class}(2) = {} & 0.00906[\Omega(\mu_1^{\text{PS}})] - 1.552 \times 10^{-4}[\Omega(\mu_5^{\text{vdW}})] \\ & + 0.01485[\Omega(\mu_4^{\text{Ch}})] - 0.00209[\Omega(\mu_2^{\text{PS}})] + 2.626 \\ & \times 10^{-4}[\Omega(\mu_3^{\text{PS}})] - 3.842 \times 10^{-5}[\Omega(\mu_4^{\text{PS}})] + 1.152 \\ & \times 10^{-4}[\Omega(\mu_4^{\text{MR}})] + 1.201 \times 10^{-6}[\Omega(\mu_5^{\text{PS}})] - 9.82 \\ & \times 10^{-5}[\Omega(\mu_5^{\text{MR}})] - 3.826 \times 10^{-5}[\Omega(\mu_8^{\text{H}})] \\ & - 0.06262[\Omega(\mu_2^{\text{Pol}})] + 1.66893[\Omega(\mu_1^{\text{Pol}})] \\ & - 0.00785[\Omega(\mu_5^{\text{Ch}})] + 0.11233[\Omega(\mu_3^{\text{Ch}})] \\ & - 0.65174\end{aligned}$$

Here the letter Ω is used to indicate that the corresponding variable in brackets is orthogonalized respect to the rest of the variables included in the model. We remark that the classification of compounds using the orthogonalized model are exactly the same as with the non-orthogonalized one and that the main differences are in the interpretation of the results. As can be seen there are not changes in the sign of coefficients except for the intercept. However, the relative contribution of the variables in the orthogonalized model is significantly different compared to those in the non-orthogonalized one. For instance, the variables μ_2^{PS} and μ_4^{Ch} have similar contributions (in absolute terms) in the non-orthogonalized model. However, in the orthogonalized model the contribution of μ_4^{Ch} is 10 times larger than that of μ_2^{PS}. These differences in the relative importance of the variables in both models can influence the contributions of the different bonds to the clastogenicity of the compounds under study, which are the main purpose of our

Table 1
Predictions made by using TOPS-MODE classification model for clastogenic (1) and nonclastogenic (−1) compounds in the training and test sets

No.	CAS	Class	Pred.	Prob.
1	684-93-5	1	1	97.01
2	760-56-5	1	1	85.05
3	869-01-2	1	1	63.24
4	10589-74-9	1	1	56.94
5	60391-92-6	1	1	59.09
6	49606-40-8	1	1	89.42
7	28895-91-2	1	1	61.22
8	154-93-8	1	1	81.53
9	615-53-2	1	1	87.04
10	19935-86-5	1	1	81.75
11	6558-78-7	1	1	77.44
12	24423-85-6	1	1	57.32
13	64005-62-5	1	1	72.52
14	58139-35-8	1	1	68.15
15	58139-33-6	1	1	64.80
16	70-25-7	1	1	89.76
17	63885-23-4	1	1	88.57
18	13010-08-7	1	1	81.88
19	6494-81-1	1	1	97.33
20	56525-09-8	1	1	92.31
21	1116-54-7	1	1	81.94
22	59665-11-1	1	1	64.76
23	56986-35-7	1	1	56.17
24	121-88-0	1	1	83.10
25	67-20-9	1	1	69.52
26	3688-53-7	1	−1	25.79
27	75321-20-9	1	1	91.43
28	42397-65-9	1	1	91.77
29	42397-64-8	1	1	91.32
30	680-31-9	1	1	98.97
31	52-24-4	1	1	99.88
32	62-73-7	1	1	96.47
33	60-51-5	1	1	88.77
34	121-75-5	1	1	75.85
35	50-18-0	1	1	89.60
36	66-27-3	1	1	86.62
37	62-50-0	1	1	81.36
38	55-98-1	1	1	91.46
39	128-44-9	1	1	89.51
40	133-67-5	1	1	95.73
41	54-31-9	1	U	47.80
42	339-44-6	1	1	98.70
43	2783-94-0	1	1	83.67
44	1934-21-0	1	1	83.95
45	133-06-2	1	−1	37.89
46	106-89-8	1	1	78.14
47	96-09-3	1	1	54.34
48	122-60-1	1	1	69.16
49	1024-57-3	1	1	58.15
50	106-50-3	1	1	63.60
51	100-22-1	1	1	85.60
52	156-43-4	1	1	67.30
53	90-04-0	1	1	78.68
54	1129-41-5	1	1	71.45
55	63-25-2	1	1	80.56
56	51-21-8	1	1	54.34
57	17902-23-7	1	1	74.57
58	147-94-4	1	1	93.41
59	58-55-9	1	1	99.42
60	58-08-2	1	1	99.82
61	598-72-1	1	−1	37.97
62	2052-01-9	1	−1	30.99
63	79-10-7	1	1	77.97
64	79-06-1	1	1	65.85
65	140-88-5	1	1	71.43
66	260-94-6	1	−1	36.76
67	230-27-3	1	1	64.83
68	67977-01-9	1	1	95.80
69	74-31-7	1	−1	11.78
70	93-46-9	1	−1	11.97
71	118-71-8	1	−1	29.74
72	501-30-4	1	−1	45.01
73	154-23-4	1	1	62.86
74	57-50-1	1	1	88.35
75	59-92-7	1	−1	24.77
76	96-13-9	1	1	80.76
77	616-23-9	1	U	48.54
78	57-13-6	1	−1	34.28
79	67-64-1	1	−1	19.14
80	50-00-0	1	[a]	
81	75-07-0	1	[a]	
82	104-55-2	1	[a]	
83	2111-75-3	1	[a]	
84	57-14-7	1	1	99.15
85	306-37-6	1	1	99.80
86	38604-70-5	1	1	72.89
87	98-92-0	1	U	50.71
88	121-79-9	1	1	71.58
89	1401-55-4	1	1	81.39
90	15972-60-8	1	1	75.89
91	494-03-1	1	−1	38.71
92	62450-07-1	1	1	72.21
93	62450-06-0	1	1	71.85
94	396-01-0	1	1	99.27
95	61-25-6	1	1	99.41
96	50-07-7	1	1	97.78
97	458-37-7	1	1	88.03
98	83-88-5	1	1	71.11
99	81-88-9	1	1	93.18
100	55726-47-1	1	−1	4.42
101	614-95-9	1[b]	1	85.49
102	7090-25-7	1[b]	1	74.30
103	56986-37-9	1[b]	1	66.97
104	122-14-5	1[b]	1	92.84
105	968-81-0	1[b]	−1	24.67
106	25956-17-6	1[b]	1	98.23
107	95-54-5	1[b]	1	67.06
108	2451-62-9	1[b]	1	99.98
109	60-27-5	1[b]	1	80.47
110	54-85-3	1[b]	1	73.69
111	59-98-3	1[b]	−1	33.89
112	924-16-3	−1	U	49.30
113	59665-03-1	−1	−1	72.76
114	64005-60-3	−1	−1	61.58
115	59665-06-4	−1	−1	71.94
116	61347-09-9	−1	1	21.93
117	98-95-3	−1	U	50.90
118	88-72-2	−1	−1	53.19
119	99-09-2	−1	1	27.06
120	35089-69-1	−1	−1	53.03
121	121-14-2	−1	1	20.75
122	81-15-2	−1	1	16.69
123	5324-12-9	−1	1	22.77
124	26087-47-8	−1	−1	96.20
125	2921-88-2	−1	−1	66.89
126	83-86-3	−1	−1	68.85
127	631-27-6	−1	−1	60.59
128	1156-19-0	−1	−1	57.16

Table 1 (*Continued*)

No.	CAS	Class	Pred.	Prob.
129	62-53-3	−1	−1	61.11
130	95-51-2	−1	−1	73.53
131	548-93-6	−1	1	43.50
132	591-62-8	−1	−1	54.19
133	539-89-9	−1	−1	75.20
134	107-92-6	−1	−1	99.74
135	144-62-7	−1	−1	98.04
136	110-15-6	−1	−1	99.38
137	110-17-8	−1	−1	70.47
138	328-50-7	−1	−1	98.52
139	87-69-4	−1	−1	99.24
140	224-42-0	−1	−1	77.45
141	226-36-8	−1	−1	85.30
142	61-68-7	−1	−1	87.44
143	50-60-2	−1	−1	55.45
144	90-30-2	−1	−1	84.30
145	135-88-6	−1	−1	85.24
146	10236-47-2	−1	1	35.08
147	58-86-6	−1	−1	58.78
148	50-81-7	−1	1	33.95
149	50-70-4	−1	−1	93.29
150	72-18-4	−1	−1	86.62
151	56-84-8	−1	−1	83.95
152	25104-18-1	−1	−1	80.18
153	3081-61-6	−1	−1	82.85
154	54-12-6	−1	−1	68.13
155	75-34-3	−1	−1	97.57
156	156-59-2	−1	−1	76.68
157	75-35-4	−1	−1	98.21
158	56-23-5	−1	−1	98.63
159	422-05-9	−1	−1	87.17
160	58-89-9	−1	−1	76.79
161	108-90-7	−1	−1	89.75
162	120-83-2	−1	−1	90.82
163	120-82−1	−1	−1	97.00
164	609-19-8	−1	−1	94.94
165	4901-51-3	−1	−1	96.75
166	87-86-5	−1	−1	97.93
167	62-56-6	−1	−1	98.15
168	598-50-5	−1	−1	84.12
169	625-52-5	−1	−1	60.26
170	38869-91-9	−1	−1	76.97
171	464-49-3	−1	−1	96.50
172	484-78-6	−1	−1	62.74
173	112-31-2	−1	[a]	
174	123-11-5	−1	[a]	
175	121-33-5	−1	[a]	
176	121-32-4	−1	[a]	
177	105-54-4	−1	−1	57.65
178	123-86-4	−1	−1	66.26
179	105-46-4	−1	−1	76.37
180	140-11-4	−1	−1	77.47
181	94-26-8	−1	−1	64.28
182	4247-02-3	−1	−1	81.73
183	87-29-6	−1	U	47.67
184	101-14-4	−1	−1	85.02
185	80-05-7	−1	−1	98.61
186	101-68-8	−1	−1	76.02
187	77-73-6	−1	−1	78.13
188	2371-42-8	−1	−1	98.25
189	611-32-5	−1	U	50.84
190	612-60-2	−1	−1	53.17
191	1107-26-2	−1	−1	72.84
192	7235-40-7	−1	−1	94.61
193	465-42-9	−1	−1	98.29

Table 1 (*Continued*)

No.	CAS	Class	Pred.	Prob.
194	108-88-3	−1	−1	98.21
195	100-41-4	−1	−1	98.41
196	100-42-5	−1	−1	81.31
197	1014-70-6	−1	1	12.72
198	834-12-8	−1	1	30.68
199	110-54-3	−1	−1	98.20
200	110-82-7	−1	−1	98.99
201	110-86-1	−1	−1	58.50
202	9003-39-8	−1	1	35.73
203	105-60-2	−1	−1	73.48
204	115-77-5	−1	1	14.67
205	4543-95-7	−1	−1	68.49
206	78-70-6	−1	−1	86.95
207	2216-51-5	−1	−1	98.28
208	54−11-5	−1	1	46.06
209	92-52-4	−1	−1	98.10
210	104-67-6	−1	−1	89.16
211	94-74-6	−1	−1	86.12
212	128-37-0	−1	−1	98.27
213	85-01-8	−1	−1	80.64
214	129-00-0	−1	−1	68.25
215	87084-52-4	−1	−1	95.76
216	56-53-1	−1	−1	89.17
217	19666-30-9	−1	−1	87.04
218	140-03-4	−1	−1	94.89
219	57-63-6	−1	−1	86.53
220	67-42-5	−1	−1	69.48
221	117-81-7	−1	−1	95.41
222	52-86-8	−1	1	46.42
223	17673-25-5	−1	−1	79.35
224	50-14-6	−1	−1	99.84
225	14929-11-4	−1	−1	88.54
226	299-88-7	−1	U	50.80
227	52423-28-6	−1	−1	96.21
228	62-75-9	−1[b]	1	0.50
229	54897-62-0	−1[b]	−1	56.55
230	54897-63-1	−1[b]	−1	66.49
231	3817-11-6	−1[b]	−1	60.81
232	64005-58-9	−1[b]	U	50.73
233	68061-82-5	−1[b]	1	9.26
234	6494-88-8	−1[b]	1	3.22
235	64005-59-0	−1[b]	−1	78.72
236	99-08-1	−1[b]	−1	56.00
237	99-99-0	−1[b]	−1	56.21
238	35089-66-8	−1[b]	−1	69.01
239	5458-83-3	−1[b]	−1	69.01
240	59-87-0	−1[b]	−1	66.71
241	1582-09-8	−1[b]	1	10.47
242	126-73-8	−1[b]	−1	58.04
243	333-41-5	−1[b]	−1	66.89
244	107-35-7	−1[b]	1	14.19
245	58-94-6	−1[b]	1	1.64
246	64-77-7	−1[b]	−1	57.16
247	76824-35-6	−1[b]	−1	72.07
248	860-22-0	−1[b]	1	2.63
249	554-00-7	−1[b]	−1	84.85
250	535-87-5	−1[b]	1	36.78
251	105-40-8	−1[b]	1	35.13
252	623-78-9	−1[b]	1	41.11
253	7451-46-9	−1[b]	−1	60.61
254	7450-62-6	−1[b]	1	37.51
255	50-06-6	−1[b]	−1	68.66
256	64-19-7	−1[b]	−1	98.83
257	56-40-6	−1[b]	−1	84.15
258	50-21-5	−1[b]	−1	98.77

Table 1 (*Continued*)

No.	CAS	Class	Pred.	Prob.
259	110-44-1	−1[b]	−1	70.47
260	6915-15-7	−1[b]	−1	98.41
261	77-92-9	−1[b]	−1	86.31
262	122-39-4	−1[b]	−1	85.30
263	101-67-7	−1[b]	−1	99.22
264	153-18-4	−1[b]	1	27.54
265	90-80-2	−1[b]	−1	53.26
266	7512-17-6	−1[b]	−1	73.88
267	73-32-5	−1[b]	−1	85.31
268	142-47-2	−1[b]	−1	87.40
269	67-21-0	−1[b]	−1	88.20
270	71-00-1	−1[b]	1	37.63
271	75-09-2	−1[b]	1	1.03
272	107-06-2	−1[b]	1	20.77
273	71-55-6	−1[b]	−1	97.90
274	79-01-6	−1[b]	−1	98.43
275	127-18-4	−1[b]	−1	98.16
276	542-18-7	−1[b]	−1	99.17
277	3322-93-8	−1[b]	1	32.38
278	106-46-7	−1[b]	−1	98.67
279	95-50-1	−1[b]	−1	99.12
280	108-70-3	−1[b]	−1	97.25
281	15950-66-0	−1[b]	−1	99.47
282	88-06-2	−1[b]	−1	94.47
283	95-95-4	−1[b]	−1	94.70
284	935-95-5	−1[b]	−1	96.58
285	118-74-1	−1[b]	−1	99.70
286	627-06-5	−1[b]	−1	66.33
287	592-17-6	−1[b]	−1	86.82
288	1792-17-2	−1[b]	−1	78.85
289	520-45-6	−1[b]	−1	67.88
290	5392-40-5	−1[b]	[a]	
291	5103-71-9	−1[b]	−1	86.84
292	5566-34-7	−1[b]	−1	81.54
293	37415-56-8	−1[b]	−1	98.10
294	16561-29-8	−1[b]	−1	98.28
295	70497-14-2	−1[b]	−1	68.59
296	7491-76-1	−1[b]	−1	81.87
297	105-37-3	−1[b]	U	50.68
298	110-45-2	−1[b]	−1	73.66
299	123-92-2	−1[b]	−1	87.37
300	106-27-4	−1[b]	−1	89.04
301	123-68-2	−1[b]	−1	54.52
302	659-70-1	−1[b]	−1	96.53
303	93-58-3	−1[b]	−1	53.22
304	119-36-8	−1[b]	−1	61.93
305	101-97-3	−1[b]	−1	77.13
306	4191-73-5	−1[b]	−1	76.73
307	120-61-6	−1[b]	U	50.00
308	84-66-2	−1[b]	−1	62.27
309	94-36-0	−1[b]	−1	90.91
310	84-74-2	−1[b]	−1	62.27
311	3648-21-3	−1[b]	−1	88.87
312	127-47-9	−1[b]	−1	87.70
313	59-02-9	−1[b]	−1	99.75
314	514-78-3	−1[b]	−1	94.78
315	115-32-2	−1[b]	−1	94.25
316	72-43-5	−1[b]	−1	96.83
317	10097-16-2	−1[b]	−1	62.91
318	61-50-7	−1[b]	−1	68.84
319	53-86-1	−1[b]	−1	63.19
320	97-56-3	−1[b]	−1	79.26
321	54-88-6	−1[b]	−1	65.30
322	132-27-4	−1[b]	−1	82.99
323	1079-21-6	−1[b]	−1	75.65
324	108-30-5	−1[b]	−1	61.14
325	352-97-6	−1[b]	−1	89.38
326	100-51-6	−1[b]	−1	89.90
327	141-97-9	−1[b]	−1	73.43
328	59-67-6	−1[b]	−1	56.99
329	57912-86-4	−1[b]	−1	90.93
330	103-84-4	−1[b]	−1	82.60
331	91-22-5	−1[b]	−1	53.17
332	94-75-7	−1[b]	−1	90.80
333	25013-16-5	−1[b]	−1	69.27
334	69-93-2	−1[b]	1	6.07
335	23333-91-7	−1[b]	−1	91.76
336	363-03-1	−1[b]	−1	56.87
337	1912-24-9	−1[b]	−1	52.81
338	139-40-2	−1[b]	−1	54.61
339	148-79-8	−1[b]	−1	71.52
340	59-00-7	−1[b]	1	26.49
341	58-15-1	−1[b]	1	6.11
342	60-00-4	−1[b]	−1	87.67
343	70699-77-3	−1[b]	−1	82.84
344	439-14-5	−1[b]	−1	56.79
345	584-79-2	−1[b]	−1	93.73
346	57-83-0	−1[b]	−1	99.58
347	50-37-3	−1[b]	1	19.33
348	50-55-5	−1[b]	1	0.52
349	101-25-7	−1[b]	1	0.12
350	86-30-6	−1[b]	−1	62.66
351	78-43-3	−1[b]	1	41.55
352	24019-05-4	−1[b]	−1	89.21
353	108-64-5	−1[b]	−1	82.33
354	123-66-0	−1[b]	−1	69.76
355	105-68-0	−1[b]	−1	85.98
356	106-32-1	−1[b]	−1	79.59
357	103-36-6	−1[b]	−1	56.38
358	526-95-4	−1[b]	−1	94.20
359	80-68-2	−1[b]	−1	78.07
360	87-61-6	−1[b]	−1	99.71
361	106-47-8	−1[b]	−1	76.12
362	62-55-5	−1[b]	−1	98.74
363	56-81-5	−1[b]	U	50.63
364	106-24-1	−1[b]	−1	85.35
365	90-43-7	−1[b]	−1	88.67
366	1897-45-6	−1[b]	−1	69.40
367	120-12-7	−1[b]	−1	91.18
368	7287-19-6	−1[b]	−1	57.34
369	34522-32-2	−1[b]	−1	85.29
370	5522-43-0	−1[b]	1	10.53
371	471-80-7	−1[b]	−1	97.56
372	59-30-3	−1[b]	−1	58.25

[a] Outlier.

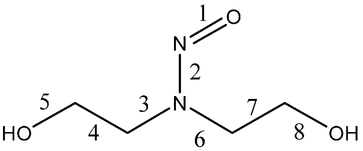
[b] Test set.

research. For example, in Table 2 we illustrate the contributions for all bonds of *N*-nitrosodiethanolamine before and after orthogonalization. As can be seen the contributions of the C–OH bonds are negative before orthogonalization and positive after the model is orthogonalized. The importance of this bond contribution for understanding the chromosome aberration produced by this chemical is further analyzed here.

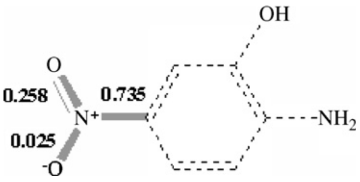
5. Generation of structural alerts

In order to generate the structural alert rules we start by calculating the bond contributions of all bonds in all the

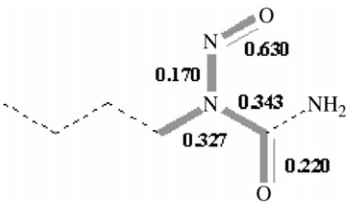
Table 2
Bond contributions for *N*-nitrosodiethanolamine before (NO) and after (O) orthogonalization of variables in the TOPS-MODE classification model

Structure numbering	Bond	C(NO)	C(O)
	1	0.623	0.725
	2	0.134	−0.294
	3	0.234	0.372
	4	−0.077	−0.059
	5	−0.066	0.051
	6	0.234	0.372
	7	−0.077	−0.059
	8	−0.066	0.051

molecules studied as explained previously. Then, we group all these molecules into chemical classes following a criterion of functional groups or similar molecular structural regions. For instance, we group together all compounds having a nitro group directly bonded to a benzene ring and we check that most of them have the same (qualitative) contribution for the certain similar regions, such as the nitro group:



When the pattern of clastogenicity is extended beyond a functional group to include a molecular region which is repeated in several compounds with the same fragment we select this region as the structural alert instead of the functional groups that compose it:



Only those structural patterns for which existed a sufficiently large number of compounds or enough chemico-biological information about its possible role in CA were proposed as structural alerts. Then, these structural alerts were tested for robustness using our prediction sets of compounds. In Table 3 we give details for all the rules generated in this work, which include the prototype structure defining the structural alert, the compounds in the dataset that respond to this classification and some examples of them identifying the bond contributions of the region responsible for the clastogenic activity. The 22 structural alert rules generated here account for the following structural classes of compounds: *N*-nitrosoureas, *N*-nitrosourethanes, *N*-nitrosoguanidines and *N*-nitrosotriazines, nitro compounds (aromatic and heteroaromatic), alkyl esters or phosphoric acids, alkyl methanesulfonates, sulphonic acids and sulphonamides, epoxides, aromatic amines, phenols or precursors, urethanes, uracil and xanthines, aromatic imines and azo compounds, α,β-unsaturated carboxylic acids, amides, esters and ketones, hydrazines or precursors, azaphenanthrene hydrocarbons, resorcinol or precursors, quinones, other *N*-nitroso compounds, and compounds with nitrogen in an heterocyclic ring. Note that one compound can have more than one structural alert.

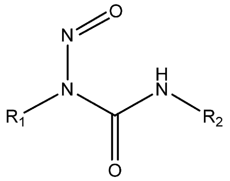
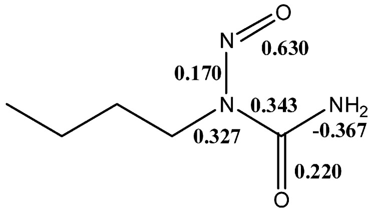
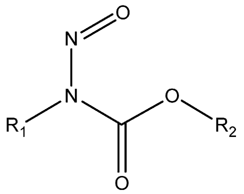
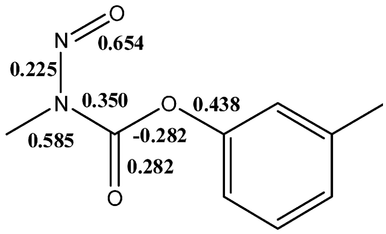
6. Chemico-biological observations

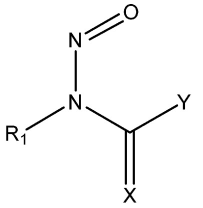
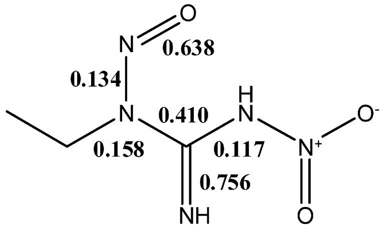
Clastogens, which are chromosome breaking chemicals, can induce CA by different mechanisms. These include DNA alkylation, inhibition of deoxyribonucleotide synthesis, denaturation or degradation of DNA, production of labile DNA by chemical reaction and/or incorporation of abnormal precursors as well as removal of DNA bound metals [7]. Depending on the mechanism the pattern of aberrations induced by a clastogen can vary. Even inside one generic mode of actions different chemicals can follow different routes to produce CA [8–11]. As a consequence the efficiency of inducing CA by different chemicals depends very much on which mechanisms and which route they use to induce the aberrations.

Most of the chemicals represented in the structural alerts extracted in this work can be considered as alkylating agents. The chromosome aberrations induced by these chemicals are distributed non-randomly among the chromosomes with the heterochromatic regions being more often involved than euchromatic regions [7]. DNA alkylation can occurs at two different positions, such as N-7 and O-6 positions of guanine and phosphate groups. The N-7 guanine is more often involved because it is more nucleophilic that O-6 guanine and phosphate group [7,8]. For instance, epoxides are well-known alkylating agents and are represented here by a specific structural alert. In the case of styrene oxide (see Table 3) it is know to react with DNA bases forming (O^6)-guanosine, (N^2)-guanosine and (N-7) guanine adducts [34]. Other alkylating agents, such as aromatic amines, need to be activated previous to their interaction with DNA bases. The initial step in the activation for aromatic and heteroaromatic amines appears to be the enzymatic N-oxidation by cytochrome P450 monooxygenases to yield a N-hydroxylamine product. The N-hydroxy species are then transformed into reactive nitrogen esters or nitrenium ions, which can attack DNA forming adducts [35,36]. Similarly, nitro aromatic compounds, represented by another structural alert, are mainly activated by means of nitroreduction and oxidative pathways, which involve several enzymes in different organisms [37]. In these cases the bond contributions illustrated in Table 3 indicate these groups (epoxide, amine and nitro) as responsible for the clastogenic action of these classes of compounds.

A close inspection of Table 3 reveals that bond contributions also identify most of the well-known groups which alkylate DNA bases. In general, the information provided by these structural alerts rules generated here together with other theoretical and experimental evidences can help to understand the mechanisms of clastogenic action of the chemicals involved or to propose new routes for their activation as clastogen. We

Table 3
Structural alerts selected by means of the bond contributions obtained from the TOPS-MODE classification model

Name	Description	Examples
N-Nitrosourea	 R_1 = aliphatic C R_2 = H, C	 O N 0.630 0.170 N 0.343 NH2 0.327 -0.367 0.220 O 1, 2, 3, 4, 5, 6, 7, 8
N-Nitrosourethane (*N*-nitrocarbamate)	 R_1 = Alkyl R_2 = C	 O N 0.654 0.225 N 0.350 O 0.438 0.585 -0.282 0.282 O 9, 10, 11, 12, 13, 14, 15, 101, 102

N-Nitrosoguanidine and *N*-nitrosotriazine	
	X = NH, N in aromatic heterocycle
	Y = NH-NO_2, N in aromatic heterocycle
	R_1 = Alkyl
		 O
		N
		0.638
		0.134
		H
		N
		0.410
		N
		O-
		0.158
		0.117
		N+
		0.756
		NH
		O
		16, 17, 18, 19, 20

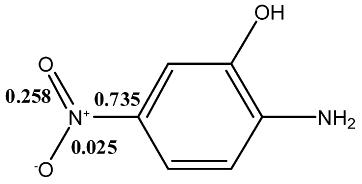
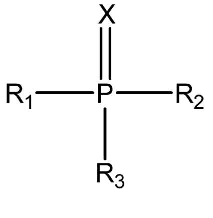
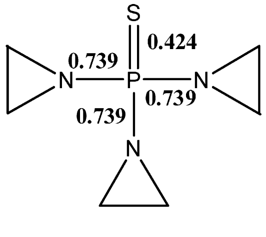
Aromatic nitro compounds	R_1——NO_2 R_1 = Aryl, heteroaromatic, PAH	 24, 25, 27, 28, 29, 104
Alkyl ester of phosphoric or phosphonic acid	 X = O, S R_1, R_2 = O, N R_3 = O, S, N	 30, 31, 32, 33, 34, 35, 104

Table 3 (*Continued*)

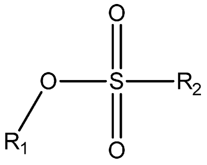
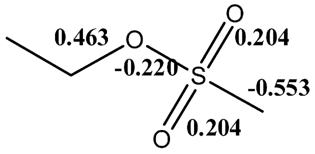
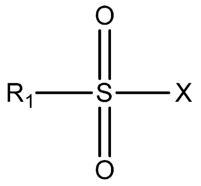
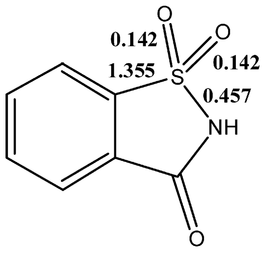
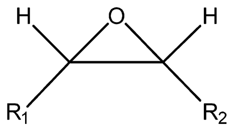
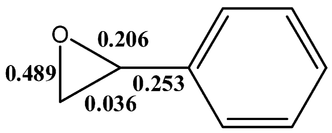
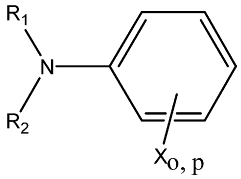
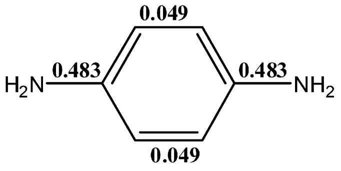
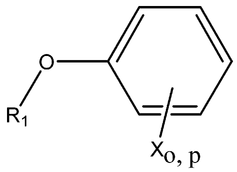
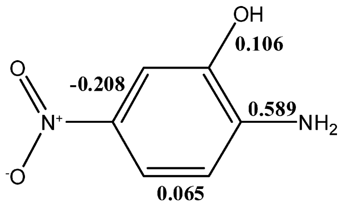
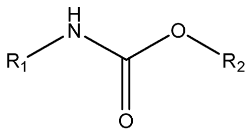
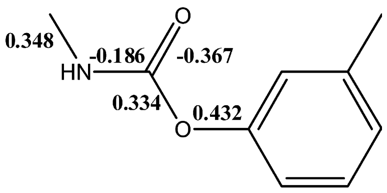
Name	Description	Examples
Alkyl methanesulfonate	 R_1 = Saturated C R_2 = Me	 O 0.463 O 0.204 -0.220 S -0.553 O 0.204 36, 37, 38
Sulphonic acid and sulphonamide	 X = OH, NH-R_2 R_1 = C aromatic R_2 = H, C in aromatic and aliphatic heterocycle, unsaturated C (C=O) in aliphatic heterocycle	 39, 40, 42, 43, 44, 106
Epoxide	 R_1 = H, C R_2 = C	 46, 47, 48, 49, 68, 108
Aromatic amine	 X = N-R_3R_4, O-R_5 R_1-R_4 = H, Alkyl R_5 = Alkyl	 50, 51, 52, 53, 107
Phenol or precursor	 X = OH, NH_2, O-R_2 R_1 = H, Alkyl R_2 = Alkyl	 OH 0.106 O -0.208 N+ 0.589 NH2 -O 0.065 24, 52, 53, 73, 88, 89, 95, 97
Urethane (carbamate)	 R_1 = H, Me R_2 = C	 54, 55, 96

Table 3 (*Continued*)

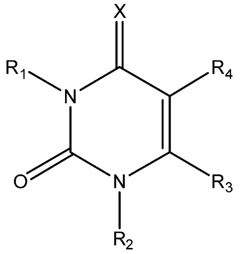
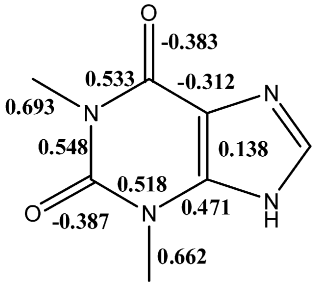
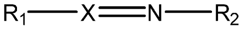
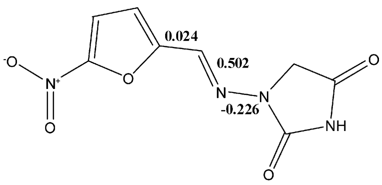
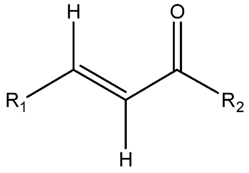
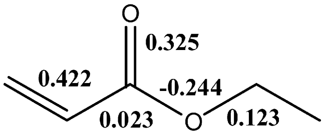
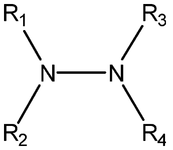
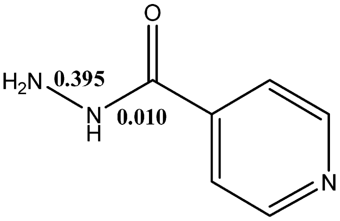
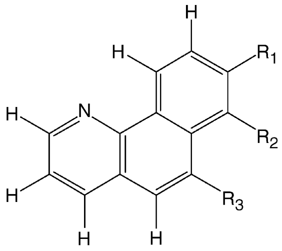
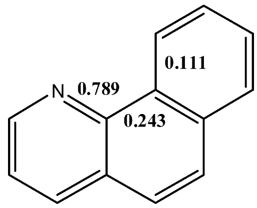
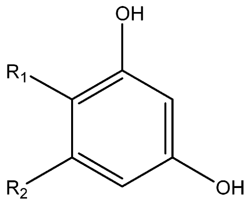
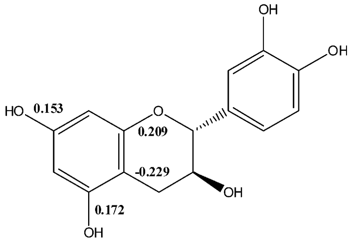
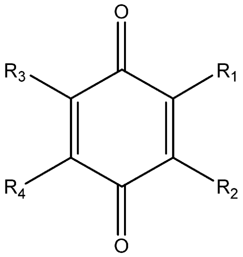
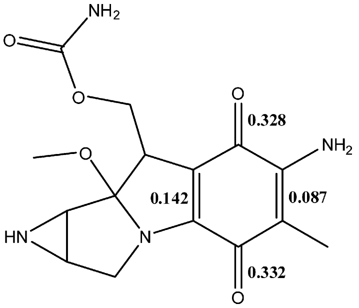
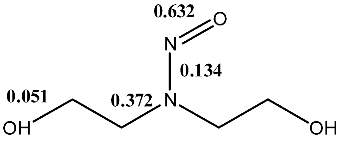
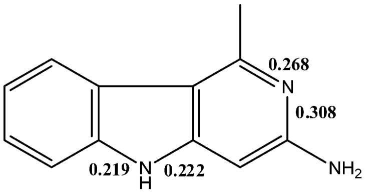
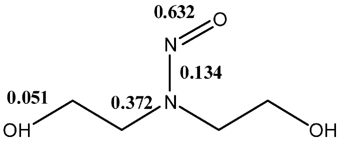
Name	Description	Examples
Uracil and xanthine (purine)	 X = O, NH R_1 = H, Alkyl R_2 = H, Alkyl, C in aromatic heterocycle R_2 = H, N in aromatic heterocycle R_2 = H, F, N in aromatic heterocycle	 O -0.383 0.533 -0.312 N 0.693 N 0.548 0.138 0.518 O -0.387 N 0.471 N H 0.662 56, 57, 58, 59, 60
Aromatic imine and azo compounds	 X = C, N R_1 = C R_2 = aromatic C, N in aliphatic heterocycle	 0.024 0.502 -O N+ O O N N -0.226 O NH O 25, 43, 44, 106
α,β-Unsaturated carboxylic acid, amide, ester and ketone	 R_1 = H, aromatic C R_2 = OH, NH_2, O-R_3, aliphatic C R_3 = Alkyl	 64, 65, 97
Hydrazine or precursor	 R_1-R_4 = H, C	 84, 85, 86, 110
Azaphenanthrene polycyclic aromatic hydrocarbon (bay region N)	 R_1, R_2 = H, aromatic C R_3 = H, aliphatic C	 67, 68

Table 3 (*Continued*)

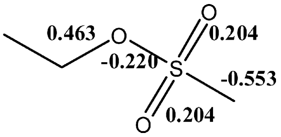
Name	Description	Examples
Resorcinol or precursor	 R_1 = C in aliphatic ring R_2 = O in aliphatic heterocycle	 OH OH HO 0.153 O 0.209 -0.229 OH 0.172 OH
Quinone	 R_1, R_2, R_3, R_4 = Any	 NH2 O O O 0.328 NH2 O 0.142 0.087 HN N 0.332 O
N-*N*itroso compounds (others)	Miscellaneous *N*-nitroso compounds	 21, 22, 23, 103
Compounds with N in heterocycle ring	Miscellaneous compounds	 92, 93, 94, 98, 109

will illustrate some representative examples here. The first example is provided by *N*-nitrosodiethanolamine (NDELA), which appears under the rule "*N*-nitroso compounds (others)". This compound has positive contribution to clastogenicity not only for bonds involved in the nitroso group but also for the C–N bonds as well as for the C–O bond of the hydroxyl groups:

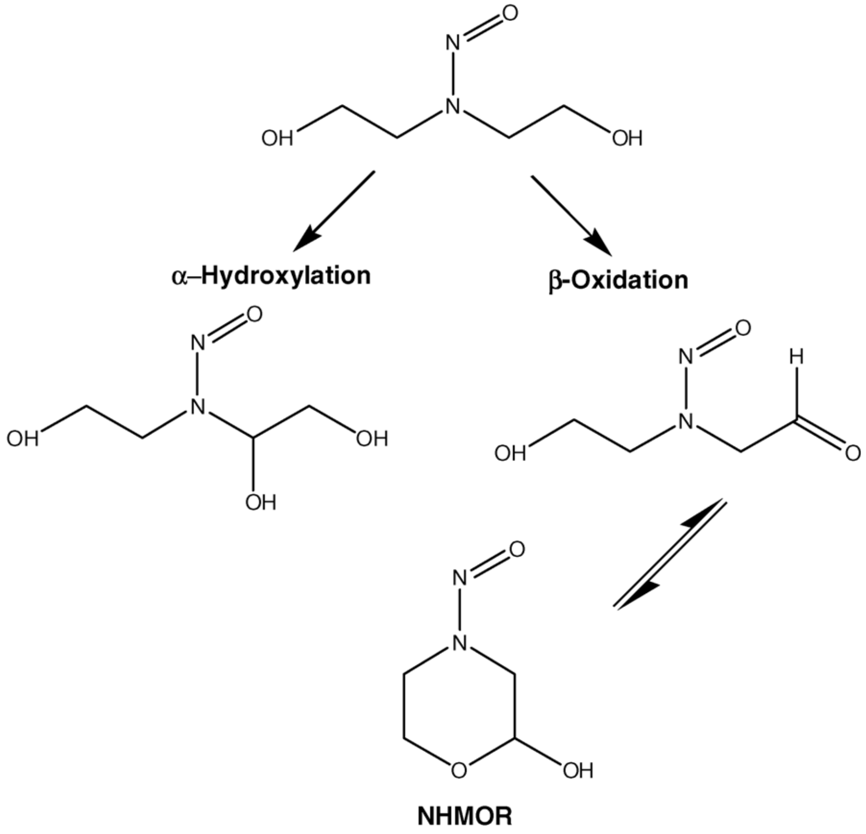


NDELA is a bident carcinogen that undergoes competitive rat liver microsome-mediated oxidation at both the α (adjacent to N)- and β-positions of the 2-hydroxyl chains as illustrated in Scheme 1 [38,39]. These are the two positions "alerted" by the bond contributions given below. NHMOR (*N*-nitrosoethylethanalamine) (see Scheme 1) has been observed to generate glyoxal-deoxyguanosine in DNA both in vitro and in vivo [39].

Another example is that of alkyl methanesulphonates, which are also alkylating agents. Methyl methanesulphonate is more efficient in inducing chromosome aberrations than ethyl or isopropyl methanesulphonate in plants, mammalian cells and *Drosophila* [40]. The reason is that methyl methanesulphonate alkylates N-7 (the most nucleophilic centre) more than O-6, while ethyl or isopropyl methanesulphonate alkylate preferentially O-6 guanine [40]. As can be seen from the bond contributions of ethyl methanesulphonate the S–C(methyl) bond has a negative contribution while the highest positive contribution is that of the O–C(ethyl) bond:



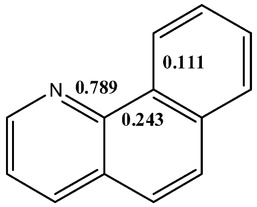
These results agree with the well-known chemical fact that the methylsulphonate ion is a good "leaving" group indicating



Scheme 1.

that the reaction that takes place is initiated by the breaking of the O–C(R) bond: $ROSO_2Met + Nu \rightarrow R - Nu + {}^{-}OSO_2Met$ [41].

Another example of the potentiality of the current approach is provided by the structural alert concerning azaphenanthrene analogues or polycyclic aromatic hydrocarbons (PAHs) with nitrogen in the bay region. As an example we illustrate the bond contributions for benzo[*h*]quinoline (BhQ):



This structure shows positive contributions for bonds in the bay region and negative contributions for the bonds on the other side of such region. BhQ has been observed experimentally to be a potent ligand for aryl hydrocarbon receptor (AhR) [42]. AhR is a ligand-activated transcription factor that mediates cellular responses through which dioxins and PAHs cause altered gene expression and toxicity [43]. Despite phenanthrene was observed to be very weak as AhR ligand this activity was significantly potentiated (about 10-fold) by nitrogen-substitution in position 4 (BhQ) and positions 1,5 (1,7-phenanthroline), moderately enhanced by nitrogen substitution in position-1 (benzo[*f*]quinoline) and not affected at all by nitrogen substitution in positions 1,8 (4,7-phenanthroline) [42]. According to Saeki et al. [42] ‘‘the result suggest that the N atom in the bay-region is more effective in enhancing the ligand activity than the non-bay-region nitrogen atom’’. These experimental results plenty agree with the theoretical ones obtained here by using TOPS-MODE approach as a knowledge-generator.

All these examples indicate, as we also previously showed for the skin sensitization of organic compounds, that the current approach constitutes a real example of knowledge-generation in which the combination of TOPS-MODE classification model, bond contributions as well as chemico-biological analysis permit the clarification of complex processes such as chromosome damage and the molecular interactions involved in it.

7. Conclusion

We have developed a classification model that permits the identification of clastogenic compounds with great variability of molecular structures. This model does not only allows the classification of chemicals as clastogenic or nonclastogenic but also permits the identification of the molecular regions responsible for the clastogenic activity. Using this information we have generated 22 rules containing structural alerts for this genotoxic activity, which permit a clear identification of certain classes of compounds as potential clastogens. These rules can be easily implemented in expert systems such as DEREK for the identification of clastogenic compounds using a procedure based on the identification of the structural patterns identified in this work in the target compounds to be evaluated by the expert system.

Acknowledgement

This work was partially financed by SEAC, Unilever, Colworth. EE manks the ‘‘Ramón y Cajal’’ program, Spain for partial financial support. EM thanks Unilever and the University of Santiago de Compostela for a post-doctoral fellowship.

References

[1] G.D. Veith, On the nature, evolution and future of quantitative structure–activity relationships (QSAR) in toxicology, SAR QSAR Environ. Res. 5/6 (2005) 323–330.

[2] N. Trinajstić, M. Randić, D.J. Klein, On the quantitative structure–activity relationships in rug research, Acta Pharm. Jugosl. 36 (1986) 267–279.

[3] E. Benfenati, G. Gini, Computational predictive programs (expert systems) in toxicology, Toxicology 119 (1997) 213–225.

[4] D.M. Sanderson, C.G. Earnshaw, Computer prediction of possible toxic action from chemical structure; the DEREK system, Human Exp. Toxicol. 10 (1991) 261–273.

[5] K. Enslein, V.K. Gombar, B.W. Blake, Use of SAR in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the TOPKAT program, Mutat. Res. 305 (1994) 47–61.

[6] T. Tunkel, K. Mayo, C. Austin, A. Hickerson, P. Howard, Practical considerations on the use of predictive models for regulatory purposes, Environ. Sci. Technol. 39 (2005) 2188–2199.

[7] A.T. Natarajan, Chromosome aberrations: past, present and future, Mutat. Res. 504 (2002) 3–16.

[8] M.A. Bender, H.G. Griggs, J.S. Bedford, Mechanisms of chromosomal aberration production. III. Chemicals and ionising radiation, Mutat. Res. 23 (1974) 197–212.

[9] M. Ishidate Jr., M.C. Harnois, T. Sofuni, A comparative analysis of data on the clastogenicity of 951 chemical substances tested in mammalian cell cultures, Mutat. Res. 195 (1988) 151–213.

[10] S.M. Galloway, Chromosome aberrations induced in vitro: mechanisms, delayed expression, and intriguing questions, Environ. Mol. Mutagen. 23 (1994) 44–53.

[11] G. Obe, P. Pfeiffer, J.R. Savage, C. Johannes, W. Goedecke, P. Jeppesen, A.T. Natarajan, W. Martinez-Lopez, G.A. Folle, M.E. Drets, Chromosomal aberrations: formation, identification and distribution, Mutat. Res. 504 (2002) 17–36.

[12] E. Estrada, G. Patlewicz, M. Chamberlain, D. Basketter, S. Larbey, Computer-aided knowledge generation for understanding skin sensitization mechanisms: the TOPS-MODE approach, Chem. Res. Toxicol. 16 (2003) 1226–1235.

[13] E. Estrada, G. Patlewicz, Y. Gutierrez, From knowledge generation to knowledge archive. A general strategy using TOPS-MODE with DEREK to formulate new alerts for skin sensitisation, J. Chem. Inf. Comput. Sci. 44 (2004) 688–698.

[14] E. Estrada, Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes, J. Chem. Inf. Comput. Sci. 36 (1996) 844–849.

[15] E. Estrada, Spectral moments of the edge adjacency matrix in molecular graphs. 2. Molecules containing heteroatoms and QSAR applications, J. Chem. Inf. Comput. Sci. 37 (1997) 320–328.

[16] E. Estrada, Spectral moments of the edge adjacency matrix in molecular graphs. 3. Molecules containing cycles, J. Chem. Inf. Comput. Sci. 38 (1998) 23–27.

[17] E. Estrada, A. Peña, R. García-Domenech, Designing sedative/hypnotic compounds from a novel substructural graph-theoretical approach, J. Comput.-Aided Mol. Des. 12 (1998) 583–595.

[18] E. Estrada, E. Uriarte, A. Montero, M. Teijeira, L. Santana, E. De Clercq, A novel approach to the rational selection and design of anticancer compounds, J. Med. Chem. 43 (2000) 1975–1985.

[19] E. Estrada, E. Uriarte, Recent advances on the role of topological indices in drug discovery research, Curr. Med. Chem. 8 (2001) 1573–1588.

[20] J.R. Serra, E.D. Thompson, P.C. Jurs, Development of binary classification of structural chromosome aberrations for a diverse set of organic compounds from molecular structure, Chem. Res. Toxicol. 16 (2003) 153–163.

[21] T. Ojima, S. Hayashi, A. Matsuoka (Eds.), Compilation of Chromosome Mutation Test Data, Life Science Information Center, Japan, 1998.

[22] E. Estrada, Edge adjacency relationships and a novel topological index related to molecular volume, J. Chem. Inf. Comput. Sci. 35 (1995) 31–33.

[23] R. Wang, Y. Gao, L. Lai, Calculating partition coefficient by atom-additive method, Perspect. Drug Discovery Des. 19 (2000) 47–66.

[24] P. Ertl, B. Rhode, P. Selzer, Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties, J. Med. Chem. 43 (2000) 3714–3717.

[25] A.K. Ghose, G.M. Crippen, Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure–activity relationships. 2. Modeling dispersive and hydrophobic interactions, J. Chem. Inf. Comput. Sci. 27 (1987) 21–35.

[26] A. Bondi, van der Waals volumes and radii, J. Phys. Chem. 68 (1964) 441–451.

[27] J. Gasteiger, M. Marsilli, A new model for calculating atomic charges in molecules, Tetrahedron Lett. 34 (1978) 3181–3184.

[28] E. Estrada, E. Molina, Novel local (fragment-based) topological molecular descriptors for QSPR/QSAR and molecular design, J Mol. Graphics Modell. 20 (2001) 54–64.

[29] M. Randić, Resolution of ambiguities in structure–property studies by use of orthogonal descriptors, J. Chem. Inf. Comput. Sci. 31 (1991) 311–320.

[30] M. Randić, Orthogonal molecular descriptors, N. J. Chem. 15 (1991) 517–525.

[31] M. Randić, Correlation of enthalpy of octanes with orthogonal connectivity indices, J. Mol. Struct. (Theochem.) 233 (1991) 45–59.

[32] B. Lučić, S. Nikolić, N. Trinajstić, D. Jurić, The structure–property models can be improved using the orthogonalized descriptors, J. Chem. Inf. Comput. Sci. 35 (1995) 532–538.

[33] D.J. Klein, M. Randić, D. Babić, B. Lučić, S. Nikolić, N. Trinajstić, Hierarchical orthogonalization of descriptors, Int. J. Quant. Chem. 63 (1997) 215–222.

[34] P. Vodicka, K. Hemminki, Identification of alkylation products of styrene oxide in single- and double-stranded DNA, Carcinogenesis 9 (1988) 1657–1660.

[35] I.A. Miller, E.C. Miller, Some historical aspects of N-aryl carcinogens and their metabolic activation, Environ. Health Perspect. 49 (1983) 3–12.

[36] J.C. Sasaki, R.S. Feller, M.E. Colvin, Metabolic oxidation of carcinogenic arylamines by P450 monooxygenases: theoretical support for one-electron transfer mechanism, Mutat. Res. 506/507 (2002) 79–89.

[37] V. Purohit, A.K. Basu, Mutagenicity of nitroaromatic compounds, Chem. Res. Toxicol. 13 (2000) 673–692.

[38] R.N. Loeppky, Q. Ye, P. Goelzer, Y. Chen, DNA adducts from *N*-nitrosodieihanolamine and related β-oxidized nitrosamines in vivo: ^{32}P-posilabeling methods for glyoxal- and *O*^{6}-hydroxyethyldeoxyguanosine adducts, Chem. Res. Toxicol. 15 (2002) 470–482.

[39] R.N. Loeppky, P. Goelzer, Microsome-mediated oxidation of N-nitrosodiethanolamine (NDELA), a bident carcinogen, Chem. Res. Toxicol. 15 (2002) 457–469.

[40] H.J. Rhaese, N.K. Boetker, The molecular basis of mutagenesis by methyl and ethyl methanesulfonates, Eur. J. Biochem. 32 (1973) 166–172.

[41] K.P.C. Volhart, N.E. Schore, Organic Chemistry. Structure and Function, 4th ed., Palgrave Macmillan, Hampshire, UK, 2002.

[42] K. Saeki, T. Matsuda, T. Kato, K. Yamada, T. Mizutami, S. Matsui, K. Fukuhara, N. Miyada, Activation of the human Ah receptor by aza-polycyclic aromatic hydrocarbons and their halogenated derivatives, Biol. Pharm. Bull. 26 (2003) 448–452.

[43] S. Safe, Molecular biology of the Ah receptor and its role in carcinogenesis, Toxicol. Lett. 120 (2001) 1–7.