11th International Electronic Conference on Synthetic Organic Chemistry (ECSOC-11) 1-30 November 2007
http://www.usc.es/congresos/ecsoc/11/ECSOC11.htm & http://www.mdpi.org/ecsoc-11/

[g003]

# QSAR modeling for predicting carcinogenic potency of nitroso-compounds using 0D-2D molecular descriptors.

Aliuska Morales Helguera [a, b, c], M. Natália D.S. Cordeiro [a], Maykel Pérez González [c, *], Miguel A. Cabrera Pérez [c], Reinaldo Molina Ruiz [a, c], Yunierkis Pérez Castillo [c].

[a]REQUIMTE, Department of Chemistry, Univeristy of Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal.

[b]Department of Chemistry; [c]Molecular Simulation and Drug Design, Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.

[*]To whom correspondence should be addressed
Email: mpgonzalez76@yahoo.es
Phone: (53)-42-281473, (53)-42-281192
Fax: (53)-42-281130

**Keyword**: nitroso-compounds, molecular descriptors, carcinogenicity, QSAR, Genetic Algorithm.

## Abstract

This paper reports a QSAR study for predicting carcinogenic potency of nitroso-compounds bioassayed in female rats administrated by gavage as oral route. Several different theoretical molecular descriptors, - 0D, 1D and 2D - calculated only on the basis of knowledge of the molecular structure and an efficient variable selection procedure, such as Genetic Algorithm, led to models with satisfactory predictive ability. But the best-final QSAR model is based on the combination between; 0D, 1D and 2D-DRAGON descriptors capturing a reasonable interpretation. This QSAR model is able to explain around 86% of the variance in the experimental activity and manifest good predictive ability as indicated by the higher $q^2$s of cross validations, which demonstrate the practical value of the final QSAR model for screening and priority testing. This model can be applied to nitroso-compounds different from the studied nitroso-compounds (even those not yet synthesized) as it is based on theoretical molecular descriptors that might be easily and rapidly calculated.

## Introduction

Carcinogenesis is a problem known to affect population all over the world and a major international health issue. Almost every sphere of human activity in society faces exposure to potential chemical hazards of some sort. Prevention of environmentally-induced cancers is a major health problem whose solutions do depend on the rapid and accurate screening of potential chemical hazards. Lately, theoretical approaches such as; Quantitative Structure–Activity Relationship (QSAR) are increasingly used for accessing the risks of environmental chemicals, since they can markedly reduce costs, avoid animal testing, and speed up policy decisions.

Amongst other chemicals, the nitroso-compounds are most likely the more important carcinogens. Of the 300 nitroso-compounds evaluated so far, more than 90 % have demonstrated to be carcinogenic in a wide variety of animal species [1]. These compounds are known to induce tumors in several vital organs causing pancreatic cancer, gastrointestinal cancer, renal or childhood brain tumors, etc. [2, 3]. This has stimulated several experimental and theoretical investigations about cancer induction by this family of compounds [3-6].

Sources of human exposure to such compounds range from occupational settings (*e.g.*: in the rubber industry) to the proper life style (diet, tobacco habits, use of cosmetics) or resort to pharmaceuticals and agricultural chemicals [7]. Furthermore, they can be generated in the body by nitrosation of amines or by reaction with products of nitric oxide generated during inflammation or infection.

The present work aims at developing a validated QSAR model for predicting the toxicity of environmental nitroso-compounds from molecular structure alone. The toxicological endpoint is carcinogenic potency, $TD_{50}$, of a set of 26 nitroso-compounds, divided into *N*-nitrosoureas (12 chemicals), *N*-nitrosamines (13 chemicals) and *C*-nitroso-compounds (1 chemical), which have been bioassayed in female rat using gavage as route of administration. We examined the use of regression models along with feature selection algorithms derived from a variety of molecular representations. For this training set, the combined descriptors provided the best model and exhibited good quality and predictive power, as judged by extensive cross-validation. Our final model shall aid in the future as an oriented tool toward preliminary ranking and prioritization of chemicals for toxicological assessment or the synthesis of nitroso-compounds with lower carcinogenicity.

## Results and Discussion

For QSAR modelling, several combinations of DRAGON descriptors - 0, 1 and 2 dimension - were considered in our study (see Table 1). Following the principle of parsimony [8] we choose the five-variable models as the "best" models.

**Table 1**. Brief description of types of descriptors used in the study.

| **Descriptors** | **Dimensionality** | **Molecular descriptors** |
|---|---|---|
| constitutional | 0D | molecular weight, no. of atoms, no. of non-H atoms, no. of bonds, no. of heteroatoms, no. of multiple bonds, no. of aromatic bonds, no. of functional groups (hydroxy, amine, aldehyde, carbonyl, nitro, nitroso, ...), no. of rings, no. of circuits, no. of H-bond donors, no. of H-bond acceptors, chemical composition |
| topological indices | 2D | molecular size index, molecular connectivity indices, information contents, Kier shape indices, path/walk-Randic shape indices, Zagreb indices, Schultz indices, Balaban J index, Wiener indices, information contents |
| molecular walk counts | 2D | molecular walk counts of order 1-10, self-re-turning of order 1-10 |
| Burden eigenvalues | 2D | positive and negative Burden eigenvalues weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume |
| topological charge indices | 2D | order 1-10 of Galvez charge topological indices, mean topological charge indices order 1-10, global topological charge index, maximum, minimum, average and total charges, local dipole index |
| autocorrelation descriptors | 2D | Broto-Moreau autocorrelation of a topological structure, Moran autocorrelation, Geary autocorrelation, H-autocorrelation weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume, leverage autocorrelation weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume, R-autocorrelation weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume |
| connectivity | 2D | Calculated from the vertex degree of the atoms in the H-depleted molecular graph |
| eigenvalue based indices | 2D | Calculated by the eigenvalues of square (usually symmetric) matrix representing a molecular graph |

**Table 2**. The statistical parameters of the linear regression models obtained for the 9 combinations of descriptors involved in the comparison

| **Block*** | **Dimensionality** | **Models** | **$R^2$** | **$q^2_{LOO}$** | **$q^2_{boot}$** | **LOF** | **AIC** |
|---|---|---|---|---|---|---|---|
| Cons | 0D | Sv Ss nAT nBO nCIR | 51.82 | 28.34 | 0.52 | 0.577 | 0.454 |
| WPC | 2D | MWC09 SRW08 MPC04 MPC05 piPC02 | 61.79 | 37.61 | 9.31 | 0.457 | 0.36 |
| TopChar | 2D | GGI2 GGI6 GGI10 JGI3 JGI4 | 63.08 | 42.61 | 32.12 | 0.442 | 0.348 |
| Eig | 2D | LP1 VRA2 VEv2 VRe2 VEp2 | 66.12 | 43.26 | 25.7 | 0.406 | 0.32 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Connec) | 2D | X2A X4A X0v X3Av X4Av | 69.00 | 54.05 | 28.56 | 0.371 | 0.292 |
| 2D-A | 2D | ATS1v ATS3v ATS4v ATS3e ATS5e | 77.96 | 64.85 | 52.15 | 0.264 | 0.208 |
| Burden | 2D | BEHm7 BELm6 BELv7 BEHe3 BELp3 | 78.37 | 62.64 | 56.51 | 0.259 | 0.204 |
| Top | 2D | ZM2V Dz Whete MAXDP T(N..N) | 81.14 | 66.27 | 42.25 | 0.226 | 0.178 |
| **0D-2D** | **0D, 1D and 2D** | **BEHm7 JGI9 VEA2 C-001 C-026** | **88.56** | **80.35** | **74.08** | **0.137** | **0.108** |

*Cons: Constitutionals; WPC: Walk Path Counts, TopChar: Topological charge indices; Eig: eigenvalue based indices; Connec: Connectivity indices; 2D-A: 2D-Autocorrelations; Burden: Burden eigenvalues; Top: Topological descriptors; 0D-2D: combination of several descriptors from 0D dimension to 2D dimension.

As can be seen in Table 2, the value of determination coefficients; $R^2$ is lower than 82 for all methodologies, except the 0D-2D combination, which has an $R^2$ equal to 88.56. This model also yielded the best values for other statistical parameters like the Akaike's information criterion (AIC) and the Friedman´s lack-of-fit function (LOF) [9] which has the lowest values in comparison with the rest of the methodologies. In the same way, the Fisher ratio is the highest. Moreover, the validation parameters confirm the before. All methodologies had statistical results inferior to the results yielded by 0D-2D combination. The results of the cross-validated determination coefficient for the leave-one-out ($q^2_{\mathrm{LOO}}$) and bootstrapping ($q^2_{\mathrm{boot}}$) procedures have values lower than 67.00 and 57.00 respectively. For all these reasons, we considered that the model resulting of combination of 0D, 1D and 2D DRAGON descriptors can be useful tools for the prediction of carcinogenic potency of nitroso-compounds.

This model is given below, together with detailed statistics of the MLR analysis.

**Model 1**

$$\begin{aligned} -\log TD_{50} &= -1.942 \cdot \mathrm{BEHm7} + 26.372 \cdot \mathrm{JGI9} - 28.197 \cdot \mathrm{VEA2} + 0.590 \cdot \mathrm{C\text{-}001} \\ &\quad -1.219 \cdot \mathrm{C\text{-}026} + 9.528 \end{aligned} \qquad (1)$$

$$N = 26 \quad R^2 = 88.56 \quad S = 0.228 \quad F = 30.959 \quad p < 10^{-5} \quad \rho = 4.333$$

$$LOF = 0.137 \quad AIC = 0.108$$

$$q^2_{\mathrm{LOO\text{-}CV}} = 80.35 \quad S_{\mathrm{CV}} = 0.298 \quad q^2_{\mathrm{Boot}} = 74.08 \quad R^2_{\mathrm{Scram}} = 0.148$$

An aspect deserving special attention is the degree of colinearity between the variables of the model, which can readily be diagnosed by analyzing the cross-correlation matrix (**Table 3**). As seen in Table 3, the pair of descriptors, (VEA2; BEHm7) is correlated each with other. For that reason, it is of interest to examine the performance of orthogonal complements.

**Table 3**. Intercorrelation among the four descriptors selected as statistically significant by the MLR-GA technique

| | **C-026** | **C-001** | **VEA2** | **JGI9** | **BEHm7** |
|---|---|---|---|---|---|
| **C-026** | 1.00 | -0.26 | 0.00 | 0.00 | -0.10 |
| **C-001** | | 1.00 | 0.07 | -0.21 | 0.13 |
| **VEA2** | | | 1.00 | -0.49 | -0.95 |
| **JGI9** | | | | 1.00 | 0.39 |
| **BEHm7** | | | | | 1.00 |

Following the Randić´s orthogonalization technique, we determined orthogonal complements for all variables in Model 1 (**eq. 1**), which in turn were further standardized, to enable derivation of the following best equation (Model 2, **eq. 2**). Predicted, observed values, simple residuals and deleted residuals are given in Table 5.

**Model 2.**

$$\begin{aligned} -\log TD_{50} = {} & \text{-0.322}\cdot{}^{5}\Omega\text{BEHm7} + 0.204\cdot{}^{4}\Omega\text{JGI9} + 0.244\cdot{}^{2}\Omega(\text{C - 001}) \\ & \text{- 0.453}\cdot{}^{1}\Omega(\text{C - 026}) \text{ - 0.609} \end{aligned} \tag{2}$$

$$N = 26 \quad R^2 = 86.82 \quad S = 0.272 \quad F = 34.590 \quad p < 10^{-5} \quad \rho = 5.200$$

$$LOF = 0.125 \quad AIC = 0.109$$

$$q^2_{\text{LOO-CV}} = 76.29 \quad S_{\text{CV}} = 0.328 \quad q^2_{\text{Boot}} = 70.64 \quad R^2_{\text{Scram}} = 0.069$$

where the symbol ${}^{i}\Omega\, X$ means the orthogonal complement of variable *X*, while the subscript refers to the order selected for orthogonalizing the variables.

Descriptor ${}^{3}\Omega$VEA2 has been excluded as it was found to be statistically non-significant. Its omission, however, had little effect on the overall fitness of the model as the statistics are as robust as before. Yet there are significant differences between Model 1 and Model 2 as regards the interpretation of the results. By comparing **eq. 1** with **eq. 2**, one can see that there are no changes in either the sign of the regression coefficients. Nevertheless, the relative contributions of the variables in the orthogonal-descriptor model are different to those in the non-orthogonalized model. Therefore, for purposes of QSAR interpretability, we shall use the orthogonal-descriptor model defined in **eq. 2**.

According to Model 2 (**eq. 2**), the $^{2}\Omega$C-001 variable – number of $CH_3R/CH_4$ fragments (where R represents any group linked through carbon) – has a positive influence on carcinogenic potency, expressed as $-\log TD_{50}$. A positive regression coefficient indicates that an increased of number of $CH_3R/CH_4$ fragments decrease the $TD_{50}$ value and increase the carcinogenic activity. In contrast, $^{1}\Omega$C-026 variable – number of R—CX—R fragments, where X represents any electronegative atom (O, N, halogens) – has a negative one. This means that the carcinogenic activity of this set of nitroso-compounds is favored by the absence of R—CX—R substructures.

Finally, two topochemicals descriptors are also inside the QSAR model, $^{5}\Omega$BEHm7 – highest eigenvalue no.7 of Burden matrix/weighted by atomic masses – and $^{4}\Omega$JGI9 – mean topological charge index of order9 – the first descriptor has negative contribution on carcinogenic activity, while the second one has positive contribution. The physical interpretation of these complex topological indices (Burden eigenvalues and Topological Charge indices) is difficult because they essentially condense a large amount of structural and property information into a single number; even so these descriptors have been extensively used in Medicinal Chemistry [10-12].

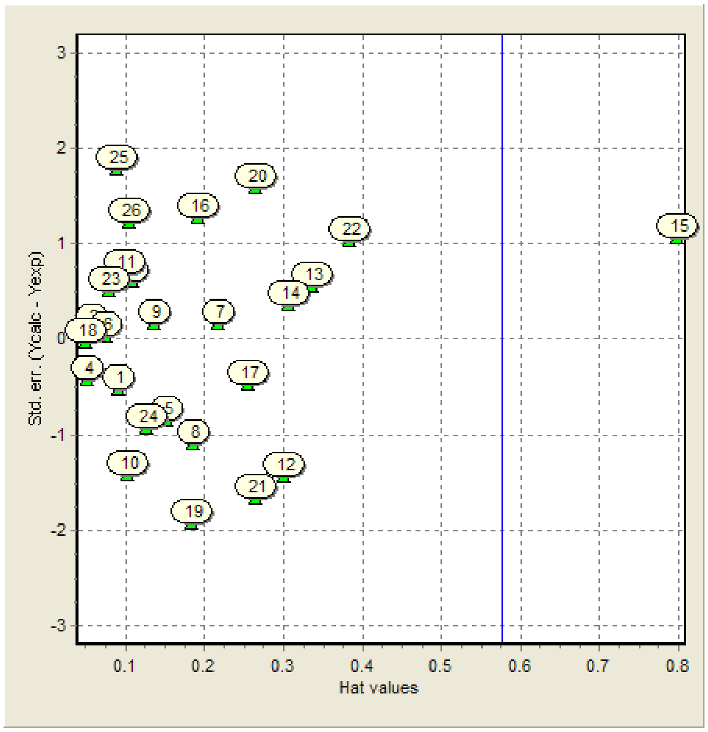

**Figure 1.** Williams plot: plot of standardized residuals ($y$-axis) *versus* leverages (hat values; $x$-axis) for each compound of the training set.

Finally, the applicability domain was established for Model 2, determining the leverage values for each compound. Figure 1 shows the Williams plot; i.e. plot of standardized residuals ($y$-axis) *versus* leverages ($x$-axis) for each compound of the training set. From this plot, the applicability domain is established inside a squared area within ±2 standard deviations and a leverage threshold $h^*$ ($h^* = 3p'/n$, being $p'$ the number of model parameters and $n$ the number of compounds). As seen in figure 1, the majority of compounds of the training set are inside of this area, however one nitroso-compound (chemical 15) has a leverage greater than $h^*$, but show standard deviation values within the limit, which implies that they are not to be consider outliers but influential chemicals [13]. For future predictions, predicted carcinogenicity data must be considered reliable only for those chemicals that fall within the applicability domain on which the model was constructed [14]

## 3. Conclusions

The relationship between the chemical structure of nitroso-compounds and their carcinogenicity in female rats administrated by gavage oral route has been investigated with the principal objective of developing QSAR models for setting testing priorities, and for screening of putative new chemical molecules before their synthesis. The use of several different theoretical molecular descriptors, calculated only on the basis of knowledge of the molecular structure, and an efficient variable selection procedure, such as Genetic Algorithm, led to models with satisfactory predictive ability for carcinogenicity.

The most accurate QSAR model was based on a combination between 0D, 1D and 2D DRAGON descriptor capturing a reasonable interpretation. This model can be applied to novel nitroso-chemicals as it is based on theoretical molecular descriptors that might be easily and rapidly calculated. Finally, it must be underlined that the predicted data must be considered reliable only for those chemicals that fall within the applicability domain on which the model was obtained.

## Materials and Methods

***Data set.*** A set of 26 nitroso-compounds (*N*-nitroso and *C*-nitroso) was used as the training set of chemicals. These had been experimentally assayed for carcinogenic potency ($TD_{50}$) in female rats and using gavage as oral administration route. For a given target site(s), and in the absence of tumors in control animals, $TD_{50}$ is taken to be the chronic dose (in mg/kg of body weight per day) that induces tumors in half of the test animals at the end of a standard lifespan for the species [15]. Thus, a low value of $TD_{50}$ indicates a potent carcinogen, whereas a high value reflects a weak carcinogen. The lowest $TD_{50}$ values reported for each chemical, expressed in µmol/kg of body weight per day and log-transformed ($-\log TD_{50}$), was used in the following QSAR modelling. This training set has been collected from Carcinogenic Potency Database (CPDB) published in the CRC Handbook of Carcinogenic Potency and Genotoxicity Databases [16] and in internet site (http://potency.berkeley.edu/cpdb.html). Table 4 and 5 give a complete list of the chemicals along with the Simplified Molecular Input Line Entry Specification (SMILES) code, the Chemical Abstract Service (CAS) Registry Number and experimental data for each chemical.

**Tabla 4.** Names, CAS numbers and SMILES of nitroso-compounds used in this QSAR study.

| Comp. | Name | CAS |
|---|---|---|

| No. | | Number | SMILES |
|---|---|---|---|
| **1** | N-Nitrosomethyl(2-oxopropyl)amine | 55984-51-5 | CN(CC(C)=O)N=O |
| **2** | N-Nitrosodiethylamine | 55-18-5 | CCN(CC)N=O |
| **3** | N-Nitrosobis(2-oxopropyl)amine | 60599-38-4 | O=NN(CC(=O)C)CC(=O)C |
| **4** | N-Nitroso-bis-(4,4,4-trifluoro-N-butyl)amine | 83335-32-4 | N(CCCC(F)(F)F)(CCCC(F)(F)F)N=O |
| **5** | N-Nitrosodipropylamine | 621-64-7 | O=NN(CCC)CCC |
| **6** | 2-Nitrosomethylaminopyridine | 16219-98-0 | C1=CC=CC(=N1)N(N=O)C |
| **7** | N-nitrosothialdine | 81795-07-5 | CC1SC(C)SC(C)N1N=O |
| **8** | N-Nitroso-N-methyl-N-dodecylamine | 55090-44-3 | O=NN(C)CCCCCCCCCCCC |
| **9** | N-Nitrosomethyl-(2-tosyloxyethyl) amine | --- | NC(CN=O)COS(=O)(C1=CC=C(C)C=C1)=O |
| **10** | N-Nitrosomethyl-(3-hydroxypropyl)amine | 70415-59-7 | N(N(CCCO)C)=O |
| **11** | N-Nitrosodithiazine | 114282-83-6 | N1(CSCSC1)N=O |
| **12** | Nitrosododecamethyleneimine | 40580-89-0 | O=NN(CCCCCC1)CCCCCC1 |
| **13** | 3-Nitrosomethylaminopyridine | 69658-91-9 | C1=CC=C(C=N1)N(N=O)C |
| **14** | 4-Nitrosomethylaminopyridine | 16219-99-1 | C1=CC(=CC=N1)N(N=O)C |
| **15** | 1-Ethylnitroso-3-(2-oxopropyl)-urea | --- | O=C(N(CC)N=O)NCC(=O)C |
| **16** | N-n-Butyl-N-nitrosourea | 869-01-2 | O=C(N(CCCC)N=O)N |
| **17** | 2-Oxopropylnitrosourea | --- | N(C(=O)N)(N=O)CC(C)=O |
| **18** | 1-Nitroso-1-hydroxyethyl-3-chloroethylurea | 96806-34-7 | O=C(N(CCO)N=O)NCCCl |
| **19** | 1-Amyl-1-nitrosourea | 10589-74-9 | O=C(N(CCCCC)N=O)N |
| **20** | N-Hexylnitrosourea | 18774-85-1 | O=C(N(CCCCCC)N=O)N |
| **21** | 1-(2-Hydroxyethyl)-1-nitrosourea | 13743-07-2 | O=C(N(CCO)N=O)N |
| **22** | 1-Allyl-1-nitrosourea | 760-56-5 | C(C(N(C(N([H])[H])=O)N=O)([H])[H])(=C([H])[H])[H] |
| **23** | 1-Nitroso-1-(2-hydroxypropyl)-3-chloroethylurea | 96806-35-8 | O=C(N(CC(C)O)N=O)NCCCl |
| **24** | N-Nitrosobenzthiazuron | 51542-33-7 | O=C(N(C)N=O)NC1=NC2=C(S1)C=CC=C2 |
| **25** | 1-(2-oxopropyl)nitroso-3-(2-chloroethyl)urea | 110559-85-8 | O=C(NCCCl)N(N=O)CC(C)=O |
| **26** | 1-(3-Hydroxypropyl)-1-nitrosourea | 71752-70-0 | O=C(N(CCCO)N=O)N |

**Table 5.** Observed, predicted and residual values of 26 nitroso-compounds used for derived the final QSAR model 2 (**eq. 2**).

| Comp. | | Carcinogenic potency[a] | | | RES[b] | $RES_{del}$[c] |
|---|---|---|---|---|---|---|
| No. | Name | $TD_{50}$ | $P_{obs}$ | $P_{pred}$ | | |
| **1** | N-Nitrosomethyl(2-oxopropyl)amine | 0.144 | -0.567 | -0.702 | 0.135 | 0.148 |
| **2** | N-Nitrosodiethylamine | 0.348 | -0.478 | -0.446 | -0.032 | -0.034 |
| **3** | N-Nitrosobis(2-oxopropyl)amine | 1.081 | -0.680 | -0.524 | -0.155 | -0.174 |
| **4** | N-Nitroso-bis-(4,4,4-trifluoro-N-butyl)amine | 1.093 | -0.265 | -0.375 | 0.111 | 0.117 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **5** | N-Nitrosodipropylamine | 1.429 | -0.397 | -0.610 | 0.213 | 0.251 |
| **6** | 2-Nitrosomethylaminopyridine | 1.56 | -0.625 | -0.614 | -0.011 | -0.012 |
| **7** | N-nitrosothialdine | 2.512 | 0.459 | 0.499 | -0.041 | -0.052 |
| **8** | N-Nitroso-N-methyl-N-dodecylamine | 2.872 | -0.155 | -0.422 | 0.267 | 0.327 |
| **9** | N-Nitrosomethyl-(2-tosyloxyethyl) amine | 13.434 | -0.045 | 0.001 | -0.046 | -0.053 |
| **10** | N-Nitrosomethyl-(3-hydroxypropyl)amine | 29.458 | -0.531 | -0.894 | 0.363 | 0.404 |
| **11** | N-Nitrosodithiazine | 34.016 | -0.936 | -0.759 | -0.177 | -0.196 |
| **12** | Nitrosododecamethyleneimine | 37.489 | 0.842 | 0.515 | 0.328 | 0.468 |
| **13** | 3-Nitrosomethylaminopyridine | 44.407 | -1.128 | -1.002 | -0.126 | -0.189 |
| **14** | 4-Nitrosomethylaminopyridine | 195.422 | -1.469 | -1.385 | -0.084 | -0.121 |
| **15** | 1-Ethylnitroso-3-(2-oxopropyl)-urea | 1.109 | -0.039 | 0.093 | -0.131 | -0.650 |
| **16** | N-n-Butyl-N-nitrosourea | 1.839 | -1.574 | -1.261 | -0.312 | -0.386 |
| **17** | 2-Oxopropylnitrosourea | 2.205 | -1.647 | -1.757 | 0.109 | 0.146 |
| **18** | 1-Nitroso-1-hydroxyethyl-3-chloroethylurea | 2.495 | -0.458 | -0.464 | 0.006 | 0.007 |
| **19** | 1-Amyl-1-nitrosourea | 3.003 | -0.034 | -0.509 | 0.475 | 0.582 |
| **20** | N-Hexylnitrosourea | 3.343 | -2.291 | -1.920 | -0.371 | -0.505 |
| **21** | 1-(2-Hydroxyethyl)-1-nitrosourea | 3.396 | -1.532 | -1.920 | 0.388 | 0.528 |
| **22** | 1-Allyl-1-nitrosourea | 3.687 | -0.400 | -0.178 | -0.222 | -0.360 |
| **23** | 1-Nitroso-1-(2-hydroxypropyl)-3-chloroethylurea | 4.222 | -0.524 | -0.389 | -0.135 | -0.146 |
| **24** | N-Nitrosobenzthiazuron | 4.783 | -0.193 | -0.428 | 0.235 | 0.269 |
| **25** | 1-(2-oxopropyl)nitroso-3-(2-chloroethyl)urea | 6.551 | -0.816 | -0.348 | -0.468 | -0.513 |
| **26** | 1-(3-Hydroxypropyl)-1-nitrosourea | 8.632 | -0.343 | -0.025 | -0.319 | -0.356 |

[a] Carcinogenic activity estimated as $TD_{50}$ (chronic dose in μmol/kg of body weight per day inducing tumors in 50% of the test animals at the end of a lifetime) and then log-transformed to $P = -\log TD_{50}$.

[b]RES = $P_{\text{obs}}-P_{\text{pred}}$

[c]Deleted residuals.

***Molecular Descriptors.*** Our models are based on nine different sets of descriptors with a long history of usage in structure–activity and structure–property correlation [17-21], which are available in the DRAGON software package (version 5.4 - 2006) [22]. These sets of molecular descriptors can be grouped into according to their dimensionality in: 0D, 1D and 2D, which are conformationally independent. The type of the descriptors used in this study is given in Table 1.

***Modelling technique.*** The objective was to obtain a mathematical function (**eq. 3**) that best describes carcinogenic potency, *P* (= −log$TD_{50}$), as a linear combination of the predictor *X*-variables (descriptors), with the coefficients $a_k$. Such coefficients were optimized by means of Multi-Linear Regression (MLR) analysis, implemented in software MobyDigs (version 1.0) [23], using Genetic Algorithm-Variable Subset Selection (GA-VSS).

$$P = a_1X_1 + a_2X_2 + \mathrm{K} + a_kX_k + a_0 \qquad (1)$$

***Feature selection.*** The Genetic Algorithm (GA) approach was used as the variable selection method [24, 25]. Starting from a population of 100 random models with a number of variables equal to or less than a user-defined maximum value, the algorithm explores new combinations of variables, selecting them by a mechanism of population evolution involving processes analogous to biological reproduction/mutation. The models based on the selected subsets of variables were tested and evaluated by the cross-validated explained variance ($q^2$), and only the best quality models were retained in the population undergoing the evolution procedure. The variables for the obtained models were found to be highly significant, within a 95% confidence level.

**Orthogonalization procedure**. One very useful and informative approach of avoiding multi-colinearity is the *orthogonal descriptors* technique suggested by Randić [26-28] some years ago. In the Randić's approach, after choosing a starting descriptor, subsequent descriptors are added only as their orthogonal complements to the descriptors already present. This approach has the advantages that: a) the regression coefficients are stable (i.e., they do not change as new descriptors are added); and b) the new information supplied by each additional descriptor is clearly distinguishable in the final equation statistics. In order to address the problem of multi-colinearity, we have applied Randić's approach by inserting the variables in descending order based on their relative contributions to $q^2$, and then pursuing to their orthogonalization. The resulting orthogonal-descriptor model was standardized afterwards.

***Model evaluation.*** Several diagnostic statistical tools were used for evaluating our model equations, in terms of the criteria *goodness-of-fit* and *goodness-of-prediction*. Measures of *goodness-of-fit* have been estimated by standard statistics such as determination coefficient, $R^2$; the standard deviation, *S*; the Fisher's statistic, *F*; as well as the ratio between the number of compounds and the number of adjustable parameters in the model, known as ρ statistics. On the other hand, goodness of the prediction was evaluated by means of cross validation (CV), basically leave-one-out (LOO-CV), bootstrapping and scrambling validation techniques [9].

Apart from the classical regression parameters listed above, we analyzed other important statistics, the Akaike's information criterion (*AIC*) and the *Friedman´s* lack-of-fit function (*LOF*) [9]. These gave us enough criteria for comparing models with different parameters, numbers of variables and chemicals.

In summary, good overall quality of the models is indicated by a large *F* (significance of the models), *FIT* and $\rho$ values; small *AIC* and *LOF* (overfitting) values; $R^2$ (goodness of fit) and $q^2$ (predictability) values close to one. In the case of $R^2_{Scram}$, this should have a value close to zero, as it checks random correlations.

**Acknowledgement.** The authors acknowledge the Portuguese *Fundação para a Ciência e a Tecnologia* (FCT) (**SFRH/BD/22692/2005**) for financial support.

## References

1. Gonzalez-Mancebo S; Gaspar J; Calle E; Pereira S; Mariano A; Rueff J; J., C., Stereochemical effects in the metabolic activation of nitrosopiperidines: correlations with genotoxicity. *Mutat Res.* **2004,** 558, (1-2), 45-51.
2. Luan, F.; Zhang, R.; Zhao, C.; Yao, X.; Liu, M.; Hu, Z.; Fan, B., Classification of the carcinogenicity of N-nitroso compounds based on support vector machines and linear discriminant analysis. *Chem Res Toxicol* **2005,** 18, (2), 198-203.
3. Minami, T.; Sakita, Y.; Okazaki, Y.; Tsutsumi, M.; Konishi, Y., Lack of involvement of metallothionein expression in pancreatic carcinogenesis by N-nitrosobis (2-oxopropyl) amine in Syrian hamsters. *Cell Mol Biol (Noisy-le-grand)* **2000,** 46, (2), 445-50.
4. Gonzalez-Mancebo, S.; Gaspar, J.; Calle, E.; Pereira, S.; Mariano, A.; Rueff, J.; Casado, J., Stereochemical effects in the metabolic activation of nitrosopiperidines: correlations with genotoxicity. *Mutat Res.* **2004** 558, ((1-2)), 45-51.
5. Singer, S. S., Decomposition reactions of (hydroxyalkyl) nitrosoureas and related compounds: possible relationship to carcinogenicity. *J Med Chem* **1985,** 28, (8), 1088-93.
6. Lijinsky, W., Structure-activity relations in carcinogenesis by N-nitroso compounds. *Cancer Metastasis Rev* **1987,** 6, (3), 301-56.
7. Benigni, R., Structure-activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem Rev* **2005,** 105, (5), 1767-800.
8. Hawkins, D. M., The problem of overfitting. *J Chem Inf Comput Sci* **2004,** 44, (1), 1-12.
9. Todeschini, R.; Consonni, V., *Handbook of Molecular Descriptors.* Wiley VCH: Weinheim, Germany, 2000.
10. González, M. P.; Teran, C.; Teijeira, M.; Besada, P.; Gonzalez-Moa, M. J., BCUT descriptors to predicting affinity toward A3 adenosine receptors. *Bioorg Med Chem Lett* **2005,** 15, (15), 3491-5.

11. Jain, H. K.; Agrawal, R. K., QSAR Analysis of Indomethacin Derivatives as Selective COX–2 Inhibitors. *Internet Electronic Journal of Molecular Design* **2006,** 5, 224–236.

12. Fernandez, M.; Caballero, J.; Helguera, A. M.; Castro, E. A.; González, M. P., Quantitative structure-activity relationship to predict differential inhibition of aldose reductase by flavonoid compounds. *Bioorg Med Chem* **2005,** 13, (9), 3269 - 3277.

13. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P., Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental health perspectives* **2003,** 111, (10), 1361-1375.

14. Vighi, M.; Gramatica, P.; Consolaro, F.; Todeschini, R., QSAR and Chemometric Approaches for Setting Water Quality Objectives for Dangerous Chemicals. *Ecotoxicology and Environmental Safety* **2001,** 49, 206-220.

15. Gold, L. S.; Manley, N. B.; Slone, T. H.; Rohrbach, L., Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996. *Environmental Health Perspectives* **1999,** 107, Suppl. 4, 527–600

16. Gold, L. S.; Manley, N. B.; Slone, T. H., *Handbook of Carcinogenic Potency and Genotoxicity Databases* CRC Press: Boca Raton, FL, 1997.

17. Helguera, A. M.; Rodríguez-Borges, J. E.; García-Mera, X.; Fernández, F.; Cordeiro, M. N. D. S., Probing the anticancer activity of nucleoside analogues: A QSAR model approach using an internally consistent training set. *J Med Chem* **2007,** 50, 1537-1545.

18. Morales, A. H.; Cabrera Perez, M. A.; González, M. P., A radial-distribution-function approach for predicting rodent carcinogenicity. *J Mol Model* **2006,** 12, (6), 769-80.

19. González, M. P.; Morales, A. H., TOPS-MODE versus DRAGON descriptors to predict permeability coefficients through low-density polyethylene. *J Comput Aided Mol Des* **2003,** 17, (10), 665-72.

20. González, M. P.; Helguera, A. M.; Medina, R.; Ruiz, R. M., QSAR with Constitutional Descriptors for the Herbicidal Properties of Fluorovinyloxyacetamides. *Internet Electron J Mol Des* **2004,** 3, (4), 200-208.

21. Helguera, A. M.; Gonzalez, M. P.; Cordeiro, M. N. D. S.; Perez, M. A. C., Quantitative Structure Carcinogenicity Relationship for detecting structural alerts in nitroso-compounds. *Toxicol Appl Pharmacol* **2007,** 221, (2), 189-202.

22. Todeschini, R.; Consonni, V.; Pavan, M. *Dragon Software*
version 2.1; 2002.

23. Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A.; Pavan, M. *Mobydigs Computer Software*, 1.0; TALETE srl: Milano, 2004.

24. Kubinyi, H., Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant Struct Act Relat* **1994,** 13, 285 - 294.

25. Kubinyi, H., Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant Struct Act Relat* **1994,** 13, 393 - 401.

26. Randić, M., Resolution of Ambiguities in Structure-Property Studies by Use of Orthogonal Descriptors. *J Chem Inf Comput Sci* **1991,** 31 311-320.

27. Randić, M., Orthogonal Molecular Descriptors. *New J Chem* **1991,** 15, (7), 517-525.

28. Randić, M., Correlation of enthalphy of octanes with orthogonal connectivity indices. *J Mol Struct (Teochem)* **1991,** 233, 45-59.