Available online at www.sciencedirect.com

ScienceDirect

Bioorganic & Medicinal Chemistry 14 (2006) 7347–7358

Bioorganic &
Medicinal
Chemistry

QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection

Liane Saíz-Urra,[a] Maykel Pérez González[a,b,c,*] and Marta Teijeira[b]

[a]*Chemical Bioactive Center, Central University of Las Villas, Santa Clara, Villa Clara, C.P. 54830, Cuba*
[b]*Department of Organic Chemistry, Vigo University, C.P. 36200, Vigo, Spain*
[c]*Service Unit, Experimental Sugar Cane Station "Villa Clara-Cienfuegos", Ranchuelo, Villa Clara, C.P. 53100, Cuba*

Received 28 February 2006; revised 12 May 2006; accepted 31 May 2006
Available online 8 September 2006

Abstract—Deoxyribonucleic acid (DNA) topoisomerases are involved in diverse cellular processes, such as replication, transcription, recombination, and chromosome segregation. Searching new compounds that inhibit both topoisomerases I and II is very important due to the deficiency of the specific inhibitors to overcome multidrug resistance (MDR). A QSAR study was developed, employing the 3D-MoRSE descriptors and a set of 64 benzophenazines in order to model the inhibition of the topoisomerases I and II, expressed by the cytotoxicity of these compounds (IC_{50}) versus drug-resistant human small cell lung carcinoma line cell H69/LX4. A comparison with other approaches such as the Topological, BCUT, Galvez topological charge indexes, 2D autocorrelations, Randić molecular profile, Geometrical, RDF, and WHIM descriptors was carried out. The mathematical models were obtained by means of the multiple regression analysis (MRA) and the variables were selected using the genetic algorithm. The model relative to the 3D-MoRSE descriptors was considered as the best, taking into account its statistical parameters. It was able to describe more than 82.2% of the variance in the experimental activity once the outliers were extracted.

1. Introduction

Deoxyribonucleic acid (DNA) topoisomerases are involved in very diverse cellular processes such as replication, transcription, recombination, and chromosome segregation. Such enzymes solve topological problems related to DNA double helical structure by breaking and rejoining DNA strands. There are two major classes of topoisomerases, the topoisomerases I, which break and reseal one strand of DNA, and the topoisomerases II that alter DNA topology by catalyzing the passing of an intact DNA double helix through a transient double-stranded break made in a second helix and has a critical role in DNA processing required for the separation of chromosomes to complete mitosis.[1]

Topoisomerase inhibitors have been widely used in the treatment of cancer. The inhibitors of topoisomerase I, for instance, camptothecin analogs such as irinotecan and topotecan have been reported for the treatment of colon cancer.[2] Furthermore a number of antitumor drugs, including the anthracycline doxorubicin, the epipodophyllotoxins etoposide and teniposide, and amsacrine are thought to be cytotoxic by virtue of their ability to stabilize a covalent topoisomerase II–DNA intermediate (the cleavable complex).[1]

However, a deficiency of these specific inhibitors, for both types of topoisomerases, results in an inability to overcome multidrug resistance (MDR).[3,4] Research of new compounds that present a dual inhibition toward topoisomerases I and II is very important. There has been extensive research of dual inhibitor enzymes, like intoplicine,[5] XR5000 (DACA),[6] and TAS-103.[7] Moreover, due to the similarity in the shape between the acridines and phenazines have been reported the synthesis and antitumoral activity of phenazine derivative, specifically, substituted phenazine-1-carboxamides, being the most active the 8,9-benzo[α]phenazine-1-carboxamide compounds.[8,9]

Keywords: QSAR; Topoisomerase inhibitors; Benzophenazines; 3D-MoRSE descriptors.

* Corresponding author. Tel.: +53 42281473; e-mail: mpgonzalez76@yahoo.es

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.05.081

Quantitative structure–activity relationships (QSAR) have been broadly used for some years mainly in medical research.[10–14] This methodology makes use of the molecular descriptors offering valuable and simple information about the structure of the molecules which is used later in the elaboration of the predictive models. The employment of this methodology allows cost savings by reducing the laboratory resources needed, and the time required to create and investigate new drugs with certain desired biological activity. For this reason, QSAR is a useful alternative tool in the research of novel compounds with dual inhibition toward the topoisomerase enzymes.

In addition, several works including QSAR studies employing physicochemical descriptors and multiregression analyses (MRA) have been reported. For instance, Garg et al.[15] formulated a QSAR for two sets of DNA-binding topoisomerase agents (bis-acridines and bis-phenazines) showing for the acridines only a small negative hydrophobic effect and for the phenazines a strong hydrophobic effect. They suggested that, despite the structural similarity of the compounds, different modes of enzyme and/or DNA binding may be involved.[15]

Mekapati et al.[16] developed seven QSAR models for the anticancer activity (growth inhibition) of various tumor cells by bis(11-oxo-11*H*-indeno[1,2-*b*]quinoline-6-carboxamides), bis(phenazine-1-carboxamides), and bis- (naphthalimides), finding positive hydrophobic interactions in two models, consistent with other QSAR studies.

More recently, Verma[17] developed sixteen quantitative structure–activity relationships (QSAR) for different sets of compounds that are camptothecin analogs, 1,4-naphthoquinones, unsaturated acids, benzimidazoles, quinolones, and miscellaneous fused heterocycles to understand chemical–biological interactions governing their inhibitory activities toward topoisomerases I and II.

Our aim was to carry out a QSAR study relative to a set of benzophenazines as dual topoisomerase I and II inhibitors. For this purpose, we used 3D descriptors, specifically the 3D-MoRSE, owing to the flexibility of these descriptors, since they afford the possibility for choosing an appropriate atomic property and in this way we could adapt them to the specific problem under study. Besides, these descriptors present an advantage as they code with fixed-length representation of 3D molecular structure, allowing us the comparison of data sets comprising molecules of different size, and number of atoms.[18,19]

2. Results and discussion

The model selection was subjected to the principle of parsimony.[20] Then, we chose a function with higher statistical significance but having as few parameters as possible. For that reason, although several models were developed for the 3D-MoRSE, changing the number of variables in every step of the analysis, the preliminary best model that we found was described with the following equation and with the statistical parameters of the regression presented next:

$$\begin{aligned}-\log(\mathrm{IC}_{50}) = {} & -0.79(\pm 0.12)\cdot \mathrm{Mor07m}\\ & -1.19(\pm 0.25)\cdot \mathrm{Mor11m}\\ & +3.80(\pm 0.52)\cdot \mathrm{Mor16m}\\ & +10.13(\pm 2.00)\cdot \mathrm{Mor12v}\\ & -2.15(\pm 0.62)\cdot \mathrm{Mor26v}\\ & -0.28(\pm 0.06)\cdot \mathrm{Mor03e}\\ & -0.72(\pm 0.27)\cdot \mathrm{Mor24e}\\ & -9.50(\pm 1.91)\cdot \mathrm{Mor12p}\\ & +3.38(\pm 0.55)\cdot \mathrm{Mor18p}\\ & +1.25(\pm 0.52) \qquad (1)\end{aligned}$$

$N = 64$, $R^2 = 0.726$, $S = 0.321$, $F = 15.946$ $p < 10^{-5}$, $\rho = 6.4$, AIC = 0.141, FIT = 0.986, $q^2_{\text{CV-LOO}} = 0.619$, $S_{\text{CV-LOO}} = 0.379$, $q^2_{\text{CV-LGO}} = 0.594$, $S_{\text{CV-LGO}} = 0.391$, where N is the number of compounds included in the model, R^2 is the square of the correlation coefficient, S is the standard deviation of the regression, F is the Fisher ratio, p is the significance of the model, and ρ is the ratio between number of cases and adjustable parameter numbers. AIC is the Akaike's information criterion and FIT is the Kubinyi function. Furthermore, we calculated the validation parameters shown previously like cross-validated squared regression coefficient q^2 and the standard deviation S_{cv} of the LOO and LGO procedures.

The parameter q^2 is used as a criterion of both robustness and predictive ability of the model. Many authors consider high q^2 (for instance, $q^2 > 0.5$) as an indicator or even as the ultimate proof that the model is highly predictive.[21] As we can see, the model proponed by us as the best one presents appropriate values of statistical parameters. Nevertheless it would be interesting to show the analysis that we carried out to determine the best model with the 3D-MoRSE descriptors family.

As we mentioned above, we developed several models for the 3D-MoRSE descriptors changing the number of variables in every step of the analysis. In other words, once a model was developed, we calculated their statistical parameters and tested if the addition of a new variable to the model was justified. If it were the case, we compared the results with the previous models and we repeated again and again this analysis whenever we included a new variable.

In this connection it was very important to calculate the parameter ρ, as a criterion in order to know when to stop the introduction of new variables in the development of the model. Our data set contains 64 compounds and the maximum number of variables is fourteen for the minimum value of ρ (4.267).

This criterion was not the only one employed to determine the optimum number of variables to include in the development of the models. Another criterion that we used is the ratio between the number of cases and variables included in the model. It has been reported that this relation is appropriate when there are five cases per variable (5:1) as a minimum value.[22] As our data set consisted of 64 compounds, the maximum number of variables is twelve. We also applied the Akaike's information criterion and Kubinyi function to determine if a variable should be included in the model. That is to say, if the Akaike's information criterion decreases in value when adding an additional variable and the Kubinyi function increases in value, then, the introduction of this new variable is justified.

The equations that describe the models relative to eight and ten variables are shown below with their statistical parameters.

$$\begin{aligned}-\log(\mathrm{IC}_{50}) = {} & -0.69(\pm 0.12)\cdot \mathrm{Mor07m}\\ & -1.27(\pm 0.26)\cdot \mathrm{Mor11m}\\ & +3.92(\pm 0.54)\cdot \mathrm{Mor16m}\\ & +3.13(\pm 0.57)\cdot \mathrm{Mor18p}\\ & -0.26(\pm 0.06)\cdot \mathrm{Mor03e}\\ & -2.37(\pm 0.65)\cdot \mathrm{Mor26v}\\ & +9.18(\pm 2.08)\cdot \mathrm{Mor12v}\\ & -8.37(\pm 1.97)\cdot \mathrm{Mor12p}\\ & +0.73(\pm 0.51)\end{aligned} \tag{2}$$

$N = 64$, $R^2 = 0.691$, $S = 0.338$, $F = 15.350$, $p < 10^{-5}$, $\rho = 7.11$, AIC = 0.152, FIT = 0. 959, $q^2_{\text{CV-LOO}} = 0.591$, $S_{\text{CV-LOO}} = 0.389$, $q^2_{\text{CV-LGO}} = 0.579$, $S_{\text{CV-LGO}} = 0.423$.

$$\begin{aligned}-\log(\mathrm{IC}_{50}) = {} & -0.64(\pm 0.12)\cdot \mathrm{Mor07m}\\ & -1.59(\pm 0.28)\cdot \mathrm{Mor11m}\\ & +4.49(\pm 0.55)\cdot \mathrm{Mor16m}\\ & +3.10(\pm 0.56)\cdot \mathrm{Mor18p}\\ & -0.26(\pm 0.06)\cdot \mathrm{Mor03e}\\ & -1.59(\pm 0.67)\cdot \mathrm{Mor26v}\\ & +12.38(\pm 2.23)\cdot \mathrm{Mor12v}\\ & -11.61(\pm 2.13)\cdot \mathrm{Mor12p}\\ & +0.55(\pm 0.19)\cdot \mathrm{Mor06v}\\ & -0.57(\pm 0.28)\cdot \mathrm{Mor14p}\\ & -0.02(\pm 0.71)\end{aligned} \tag{3}$$

$N = 64$, $R^2 = 0.701$, $S = 0.329$, $F = 18.8$, $p < 10^{-5}$, $\rho = 8.0$, AIC = 0.114, FIT = 1.162, $q^2_{\text{CV-LOO}} = 0.620$, $S_{\text{CV-LOO}} = 0.372$, $q^2_{\text{CV-LGO}} = 0.593$, $S_{\text{CV-LGO}} = 0.391$.

The quality of the statistical parameters for the model with eight variables was adequate, although the intercept of the equation turned out not to be significant. Insofar as the ρ value was equal to 7.11 and the relationship between the number of cases and the number of variables in the model was 8:1 (>5:1) we investigated the possibility of improving the statistical results by introducing a ninth variable into the model. It was determined by genetic algorithm that the variable Mor24e would be added to the model. The introduction of this new variable conformed to the Akaike's information criterion (decrement from 0.152 to 0.141) and the Kubinyi function (increased from 0.959 to 0.986). The statistical parameters in this model improved with increases in the value of the R^2, q^2, and F, and a decrease in the value of S.

In order to find better results and a more predictive model, we introduced a tenth variable and analyzed the statistical parameters (Eq. 3). In spite of the improvement in the other parameters of the comparison, the value of the Kubinyi function got worse (decrement of 0.986–0.906) as well as the Akaike's information criterion (increase of 0.141–0.143) and the intercept of the equation was not significant. We determined that the model with nine variables was best and would be used for the analysis.

Although we determined that the 3D-MoRSE descriptors family was statistically significant we carried out a comparison of different methodologies to validate our model. The results obtained from this comparison are presented in Table 1.

We can see in Table 1 that every model was developed using 12 variables except for the model proposed by our research, due to the reasons explained above. We determined that 12 variables were required to achieve optimum statistical parameters and this was supported by the Akaike's information criterion and other tests.

The models with twelve variables relative to the Topological, BCUT, and 2D autocorrelations have comparable R^2 value (0.67, 0.602, and 0.664, respectively) but lower than the R^2 reported for the model relative to the 3D-MoRSE descriptors. The Galvez topological charge indexes, Randić molecular profile, Geometrical, and WHIM descriptors are included in another group with R^2 values lower still than those reported for the group analyzed previously (0.373, 0.502, 0.534, and 0.592, respectively, and all of them lower than 0.600).

It is important to highlight the low predictive capability of all of the models with respect to these descriptors, except the model that contains the Topological descriptors and presents a q^2 value of 0.527, the rest reported a q^2 value lower than 0.5.

In lieu of these facts and the likenesses established, comparison between the models with the RDF and 3D-MoRSE descriptors was undertaken. The R^2 parameter is very similar with values of 0.70 and 0.726, respectively. The other statistical parameters were similarly trending with the model of the 3D-MoRSE descriptors yielding superior results, except for the Akaike's information criterion that is slightly lower for the RDF descriptors. With respect to the predictive capabilities, the 3D-MoRSE approach yielded a significantly higher q^2 value (0.619 vs 0.538).

Table 1. The statistical parameters of the linear regression models obtained for the nine kinds of descriptors involved in the comparison

Kind of descriptor	Variables	R^2	S	F	p	$q^2_{\text{CV-LOO}}$	$S_{\text{CV-LOO}}$	AIC	FIT
Topological	12	0.67	0.363	8.64		0.527	0.434	0.154	0.498
BCUT	12	0.602	0.398	6.438		0.398	0.49	0.186	0.371
Galvez topological charge indexes	12	0.373	0.5	2.524		0.098	0.662	0.293	0.146
2D autocorrelations	12	0.664	0.366	8.409		0.471	0.459	0.157	0.485
Randic molecular profile	12	0.502	0.446	4.286		0.232	0.553	0.233	0.247
Geometrical	12	0.534	0.431	4.875		0.291	0.536	0.218	0.281
RDF	12	0.70	0.346	9.898		0.538	0.429	0.140	0.572
3D-MoRSE	9	0.726	0.321	15.946		0.619	0.379	0.141	0.986
WHIM	12	0.592	0.404	6.158		0.397	0.490	0.191	0.356

All models contained twelve variables, except with 3D-MoRSE descriptors.

Consequently, we conducted a comparison of the same methodologies mentioned above, using all of the models relative to the 3D-MoRSE descriptors in order to emphasize the differences among this methodology and the rest, especially the RDF descriptors, because their results are the most similar to the 3D-MoRSE results. We developed the models taking into account only nine variables with the aim of equaling the conditions under study. The results of this comparison are given in Table 2.

As we can see, the only model that has predictive capability besides the 3D-MoRSE is the RDF model. However, the values of its statistical parameters of the RDF model are significantly lower than those of the 3D-MoRSE descriptors. The proposed model contains the best descriptors in order to predict the anticancer activity. It is the simplest, with just nine variables has and yielded the best statistical results. We proceeded to test whether the superiority of the 3D-MoRSE descriptors to the other methodologies in this training set for this biological property was the result of collinearity among variables. Collinearity was avoided by making an orthogonalization of molecular descriptors because interrelatedness among different descriptors can result in highly unstable models.

The QSAR model obtained with the 3D-MoRSE (Eq. 4) after orthogonalization and standardization is given below, together with the statistical parameters of the regression analysis.

$$\begin{aligned}-\log(\mathrm{IC}_{50}) = {} & -0.21(\pm 0.04)\cdot \Omega^1\mathrm{Mor07m}\\ & -0.22(\pm 0.04)\cdot \Omega^2\mathrm{Mor11m}\\ & +0.16(\pm 0.04)\cdot \Omega^3\mathrm{Mor16m}\\ & -0.14(\pm 0.04)\cdot \Omega^4\mathrm{Mor26v}\\ & -0.17(\pm 0.04)\cdot \Omega^5\mathrm{Mor03e}\\ & -0.20(\pm 0.04)\cdot \Omega^6\mathrm{Mor12p}\\ & +0.14(\pm 0.04)\cdot \Omega^7\mathrm{Mor18p}\\ & -0.08(\pm 0.04)\cdot \Omega^8\mathrm{Mor24e}\\ & +0.04(\pm 0.04)\cdot \Omega^9\mathrm{Mor12v}\\ & -2.24(\pm 0.04) \qquad (4)\end{aligned}$$

$N = 64$, $R^2 = 0.727$, $S = 0.321$, $F = 15.946$, $p < 10^{-5}$, $\rho = 6.4$, AIC = 0.120, FIT = 0.991, $q^2_{\text{CV-LOO}} = 0.619$, $S_{\text{CV-LOO}} = 0.379$.

As a result of orthogonalization the variables Mor24e and Mor12v became insignificant. This fact might result because during the orthogonalization process, the information contained in the variable to be orthogonalized, that is common to the information contained in the variables previously orthogonalized, is finally eliminated when the variable undergoes the process. In other words, in this case the variables Mor24e and Mor12v did not provide new information and that is why they turned out not to be significant. This aspect is shown in the next table (Table 3).

As we can see, there is an increase in the R^2 values in the every step for all variables in this report. Though, specifically in the case of the step from Mor26v to Mor24e, the increase of R^2 value is only 0.02, whereas the other increases are, at least, higher than 0.059.

These two variables were eliminated from the previous model, which resulted in the following equation:

$$\begin{aligned}-\log(\mathrm{IC}_{50}) = {} & -0.21(\pm 0.04)\cdot \Omega^1\mathrm{Mor07m}\\ & -0.22(\pm 0.04)\cdot \Omega^2\mathrm{Mor11m}\\ & +0.16(\pm 0.04)\cdot \Omega^3\mathrm{Mor16m}\\ & -0.14(\pm)\cdot \Omega^4\mathrm{Mor26v}\\ & -0.17(\pm 0.04)\cdot \Omega^5\mathrm{Mor03e}\\ & -0.20(\pm 0.04)\cdot \Omega^6\mathrm{Mor12p}\\ & +0.14(\pm 0.04)\cdot \Omega^7\mathrm{Mor18p}\\ & -2.24(\pm 0.04) \qquad (5)\end{aligned}$$

$N = 64$, $R^2 = 0.701$, $S = 0.329$, $F = 18.8$, $p < 10^{-5}$, $\rho = 8.0$, AIC = 0.114, FIT = 1.162, $q^2_{\text{CV-LOO}} = 0.620$, $S_{\text{CV-LOO}} = 0.372$, $q^2_{\text{CV-LGO}} = 0.593$, $S_{\text{CV-LGO}} = 0.391$.

To further test the QSAR model, it was important to examine the data outliers. The level of outliers can become a serious problem because the model is unable to predict 'real' biological activity. In this context, we looked for the presence of outliers in Eq. 5. The number of outliers extracted ranged from 0 to 6. The extraction of 10% of the general data is classically accepted in the literature as the threshold for this procedure. The two tests that we used to detect the presence of outliers were a standard residual higher than 2 × δ, where δ is equivalent to the standard deviation, and deleted residual, The structure of these outliers is shown below and the new

Table 2. The statistical parameters of the linear regression models obtained for the nine kinds of descriptors involved in the comparison

Kind of descriptor	Variables	R^2	S	F	$q^2_{\text{CV-LOO}}$	$S_{\text{CV-LOO}}$	AIC	FIT
Topological	9	0.629	0.374	10.163	0.488	0.439	0.192	0.631
BCUT	9	0.539	0.417	7.006	0.390	0.480	0.238	0.435
Galvez topological charge indexes	9	0.342	0.498	3.118	0.049	0.599	0.340	0.194
2D autocorrelations	9	0.601	0.388	9.024	0.445	0.457	0.206	0.561
Randic molecular profile	9	0.418	0.468	4.308	0.163	0.562	0.300	0.267
Geometrical	9	0.505	0.432	6.121	0.309	0.51	0.256	0.380
RDF	9	0.624	0.376	9.945	0.526	0.423	0.194	0.618
3D-MoRSE	**9**	**0.726**	**0.321**	**15.946**	**0.619**	**0.379**	**0.141**	**0.986**
WHIM	9	0.475	0.445	5.424	0.324	0.504	0.271	0.337

All models contained nine variables.

Table 3. Forward stepwise step-by-step analysis

Mor11m	Mor07m	Mor12p	Mor03e	Mor16m	Mor18p	Mor26v	Mor24e	Mor12v	b_o	R^2	S	p-level
−0.218									−2.24	0.148	0.529	2×10^{-6}
−0.218	−0.210								−2.24	0.284	0.489	3×10^{-6}
−0.218	−0.210	−0.201							−2.24	0.408	0.448	7×10^{-6}
−0.218	−0.210	−0.201	−0.167						−2.24	0.495	0.417	1.2×10^{-4}
−0.218	−0.210	−0.201	−0.167	0.164					−2.24	0.578	0.385	1.67×10^{-4}
−0.218	−0.210	−0.201	−0.167	0.164	0.144				−2.24	0.642	0.357	7.83×10^{-4}
−0.218	−0.210	−0.201	−0.167	0.164	0.144	−0.138			−2.24	0.701	0.329	1.213×10^{-3}
−0.218	−0.210	−0.201	−0.167	0.164	0.144	−0.138	−0.080		−2.24	0.721	0.321	**0.053**
−0.218	−0.210	−0.201	−0.167	0.164	0.144	−0.138	−0.080	0.041	−2.24	0.727	0.321	**0.310**

statistical parameters for each successive extraction are given in Table 4.

Taking into account the two tests employed, the **44** and **47** compounds showed the highest potentialities to be considered as outliers. Compound **44** is structurally similar to compound **43** with an important difference, the position of the OMe substituent in relation to the side chain of the same ring as can be seen in Figure 1.

This fact caused us to explore if the ortho effect might be present in the compound **44**. This effect could be one of the causes of the change in its conformation, causing the substituent be out of plane. On the other hand, the presence of an intra-molecular hydrogen bond, between the amide carbonyl and the protonated acridine nitrogen, and in the case of molecules more acidic, between the amide N–H and the phenazine nitrogen has been reported previously for active conformation of phenazines.[9]

In compound **44**, the ortho effect could impede the formation of a possible intra-molecular hydrogen bond owing to the change in its conformation. It is possible that this compound interacts in another unknown way with the receptor.

In order to understand what properties of compound **47** caused it to be an outlier we compared compound **47** with compounds **46** and **48** in order to understand why it is an outlier (see Fig. 2).

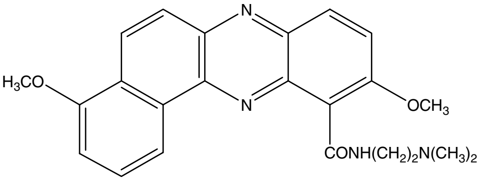
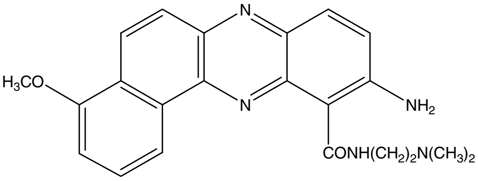
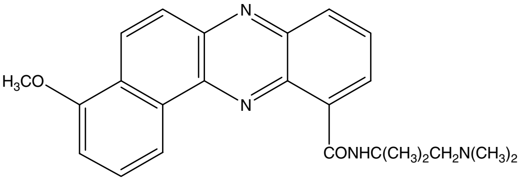
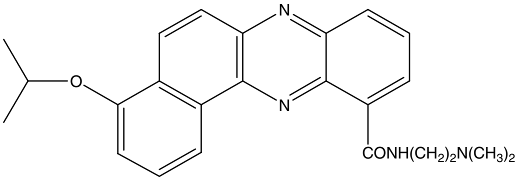
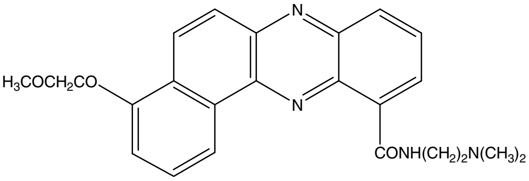
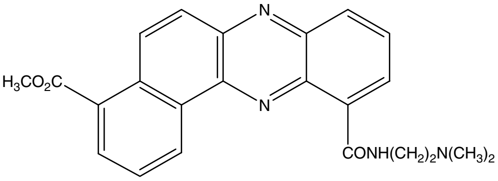
As we can see, in this set of compounds the major difference is the substituent in ortho position with respect to the side chain. All these structures are able to form an intra-molecular hydrogen bond between the substituent in ortho and the side chain (to the oxygen atom from C═O), and also between the acridine nitrogen and the side chain (to the amide N–H). Therefore, these molecules might have the necessary conformation to consider them activate in front of the receptor, in spite of the ortho effect present in these molecules.

Compound **47** presented one interesting property which differentiates it with respect to the other ones. This is the dual possibility of the NH_2 group to form hydrogen bonds due to it having two hydrogen atoms. Moreover, it has been reported before for acridine derivative, the interaction with the ADN by a specific hydrogen bond between the protonated *N*, *N*-dimethylamino group of the side chain ($\text{–CONH(CH}_2)_2\text{N(CH}_3)_2$) and the N7 from the guanine in the major groove.[23] Then, it could be possible an interaction between the second hydrogen atom from the amine group in the compound **47** and the receptor in a different mode.

The following equation was obtained without the outliers:

$$\begin{aligned} -\log(\text{IC}_{50}) = {} & -0.21(\pm 0.03) \cdot \Omega^1\text{Mor07m} \\ & - 0.26(\pm 0.03) \cdot \Omega^2\text{Mor11m} \\ & + 0.14(\pm 0.03) \cdot \Omega^3\text{Mor16m} \\ & - 0.11(\pm 0.03) \cdot \Omega^4\text{Mor26v} \\ & - 0.16(\pm 0.03) \cdot \Omega^5\text{Mor03e} \\ & - 0.18(\pm 0.03) \cdot \Omega^6\text{Mor12p} \\ & + 0.17(\pm 0.03) \cdot \Omega^7\text{Mor18p} \\ & - 2.21(\pm 0.03) \qquad (6) \end{aligned}$$

Table 4. Structures and statistics parameters of the outliers

Compound	Structure	R^2	S	F	q^2_{LOO}	S_{LOO}	AIC	FIT
44		0.721	0.307	20.28	0.637	0.35	0.102	1.279
47		0.748	0.290	22.91	0.666	0.333	0.100	1.444
59		0.770	0.277	25.372	0.697	0.318	0.098	1.613
25		0.787	0.265	27.481	0.716	0.306	0.091	1.763
35		0.806	0.255	30.202	0.737	0.297	0.086	1.962
18		0.822	0.246	33.001	0.761	0.286	0.080	2.158

$N = 58$, $R^2 = 0.822$, $S = 0.246$, $F = 33.001$, $p < 10^{-5}$, $\rho = 8.0$, AIC = 0.080, FIT = 2.158, $q^2_{\text{CV-LOO}} = 0.761$, $S_{\text{CV-LOO}} = 0.286$, $q^2_{\text{CV-LGO}} = 0.731$, $S_{\text{CV-LGO}} = 0.304$.

This final model yielded the best statistical results with just seven variables; hence its interpretation becomes simpler still.

In order to make the analysis easier, we developed Table 5 where the individual contribution of the variables to the percentage of the explained variance and their definition are shown. It highlights the contribution of the variables weighted by atomic mass, accounting for the 48.5% of the experimental variance, and being this one, higher than the experimental variance explained by the rest of descriptors with other weights like atomic polarizabilities (21.0%), atomic Sanderson electronegativities (8.3%) and atomic van der Waals volumes (4.4%).

Nevertheless, a deeper analysis is necessary to interpret the application of these kinds of descriptors. In this context, we can say that 3D-MoRSE descriptors could return a negative value because within the original equation is the following function:

$$\frac{\sin(s \cdot r_{ij})}{s \cdot r_{ij}} \tag{7}$$

where s measures the scattering angle and r_{ij} represents the interatomic distances between atoms i and j. As we can see in the chart (Fig. 3), the descriptor values as well as the sign depend, to a large extent, on the values of s and r_{ij}. For these reasons, we cannot say that certain variable has a particular influence on the biological activity, either negative or positive, only taking into account the coefficient sign at the final regression equation reported above, as the case when the indepen-

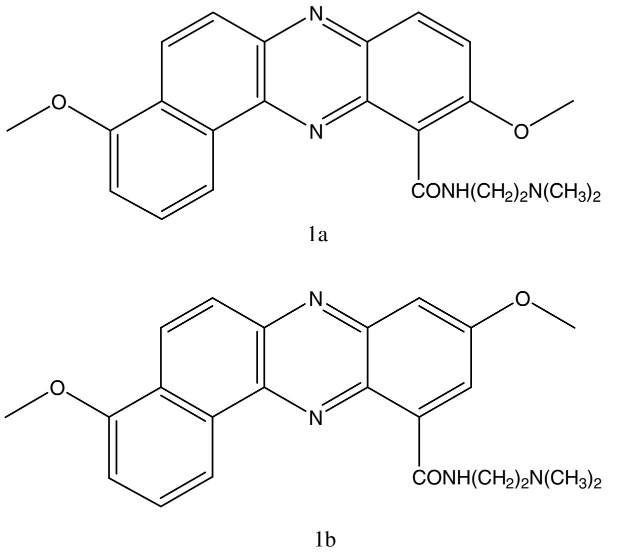


Figure 1. (a) Compound **44** and (b) compound **43**.

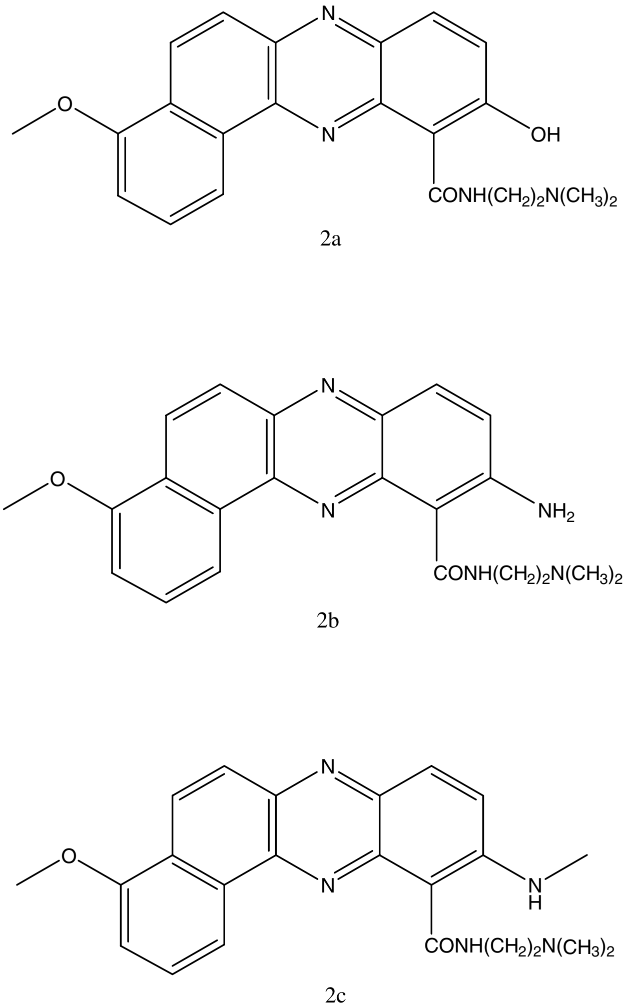


Figure 2. (a) Compound **46**; (b) compound **47**; and (c) compound **48**.

dent variables used for the biological activity modeling are only positive. In this study, when the coefficient and the independent variable (descriptor) have the same sign, the contribution of this last one is positive, else, negative.

All these things considered, we might focus the analysis on the variable group weighted by the atomic mass, since two of these, Mor11m and Mor07m, are the most meaningful. With the aim of making this analysis in the most useful way, we developed the following table.

Note in Table 1 of the supplementary material, the particular structural features of these compounds and how useful they will be for a possible comparison among them, so as to interpret the descriptors that are present in the final regression equation. In other words, in Table 6, we only showed the backbone of the data and also nine compounds more. These nine compounds were chosen because they form three groups with comparable structural features. For example, in every group the compounds are isomers, taking into account the position of the substituent in the benzene ring fused to the phenazine. When we compare across groups, the difference is the nature of the substituent, while the positions in the benzene ring is the same. The substituents used were CH_3–O, OH, and NO_2 as can be seen in Table 1 of the supplementary material.

It can be seen in Table 6 that the values for the descriptors Mor11m and Mor07m were the highest for the compounds **8**, **9**, and **10**, when apparently the structural features were not so different. Besides, it can be seen how the above-mentioned descriptors' values are even higher relative to others, when weighted by other atomic properties, for these specific compounds set with the NO_2 group. This fact allows us to hypothesize about the importance of their contribution when the difference in regard to the atomic mass, among the different substituents, is persistent, no matter what kind of contribution is exerted, either negative or positive.

We made a comparison among the compounds contained in Table 6, splitting them into groups in regard to the position of the substituents. We can make the generalization that the compounds with NO_2 are the most active for all positions, except, for position number four, in spite of presence of the highest value for the Mor07m descriptor. This fact could be supported by the contribution of other weightings, such as atomic polarizability, that increase the importance of its contribution to the biological activity that it exerts. The values reported for these kinds of descriptors for compound **10** (Mor12p = 1.81 and Mor18p = −0.14) are worse than for compounds **4** (Mor12p = 0.48 and Mor18p = 1.09) and **7** (Mor12p = −0.16 and Mor18p = 0.32). It is important to note the negative sign in the coefficient of the Mor12p descriptor, and the positive one in Mor18p and to remember that they were also considered the second most meaningful group according to its contribution to the explained variance, with 21% as we reported above.

Nonetheless, not only is weighting important in the analyses of the results obtained with the 3D-MoRSE descriptors, but also others which were previously men-

Table 5. Contribution from the variables to the pattern

Variables	Definition	R^2 global (step by step)	R^2 for every variable in the model
Mor11m	3D-MoRSE—signal 11/weighted by atomic masses	0.244	0.244
Mor07m	3D-MoRSE—signal 07/weighted by atomic masses	0.416	0.172
Mor12p	3D-MoRSE—signal 12/weighted by atomic polarizabilities	0.527	0.111
Mor18p	3D-MoRSE—signal 18/weighted by atomic polarizabilities	0.626	0.099
Mor03e	3D-MoRSE—signal 03/weighted by atomic Sanderson electronegativities	0.709	0.083
Mor16m	3D-MoRSE—signal 16/weighted by atomic masses	0.778	0.069
Mor26v	3D-MoRSE—signal 26/weighted by atomic van der Waals volumes	0.822	0.044

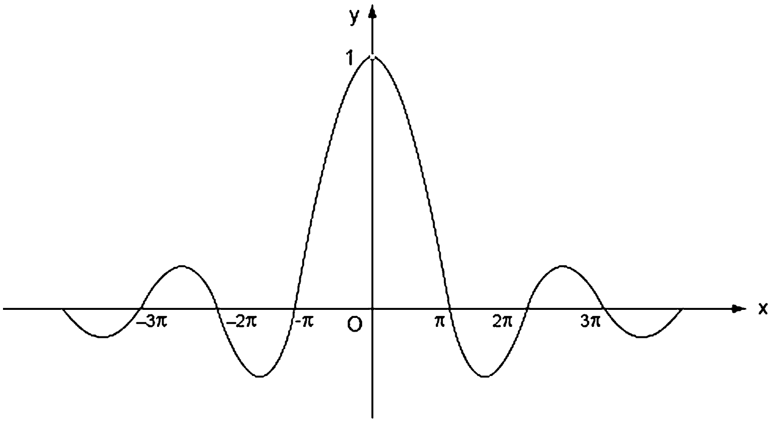


Figure 3. Chart relative to the function $y = \sin(x)/x$, where $x = s \cdot r_{ij}$.

Table 6. Descriptor values with the observed and predicted values for a sample of compounds from the data set

Compound	Mor11m	Mor07m	Mor12p	Mor18p	Obsd values	Pred values	Residual
1	0.11	−0.39	−0.47	−0.17	6.89	6.94	−0.05
2	0.75	0.26	−0.67	0.12	6.95	6.65	0.30
3	0.61	0.63	−0.12	−0.73	6.73	6.36	0.37
4	−0.02	0.27	0.48	1.09	7.31	7.01	0.30
5	0.10	0.04	−0.14	0.06	6.55	6.56	−0.01
6	0.35	0.49	−0.03	0.93	6.71	6.80	−0.08
7	−0.51	−0.22	−0.16	0.32	7.32	7.08	0.24
8	2.04	−3.69	−1.17	1.07	6.99	7.28	−0.28
9	2.20	−2.83	0.58	0.73	7.16	7.00	0.16
10	2.23	−2.59	1.81	−0.14	6.92	6.76	0.16

tioned such as the parameter s and the interatomic distances r_{ij}.

For a compound, if the weighting and the structure are kept constant, there are thirty-two possible values of $I(s)$, (3D-MoRSE descriptor) and will quantify the intensity of the scattered radiation considering these 32 different values of s for this molecule. In other words, these 32 descriptors will describe the diffraction pattern for the molecule that has been predetermined or imposed with the aim to establish a common parameter useful in several different studies. In this way, we would see the signs more marked or typical for the compound under certain conditions. However, when a model is developed, the most significant thing is that a group with certain signs correlates with a biological activity and then, their values take on a greater importance according to the specific molecule described.

The influence of the interatomic distance is related to a large extent to the structure of the different compounds and plays a significant role in the 3D coding of the molecules. For instance, when a comparison is made of the results obtained for compounds **8**, **9**, and **10**, which only differ in the substituent position, we can see that for this subset of compounds, the Mor07m descriptor contributes to a greater extent, with the reported activity being −3.69, −2.83, and −2.59, respectively.

As we can see in Figure 4, the distances among the different atoms will be the principal means to separate the

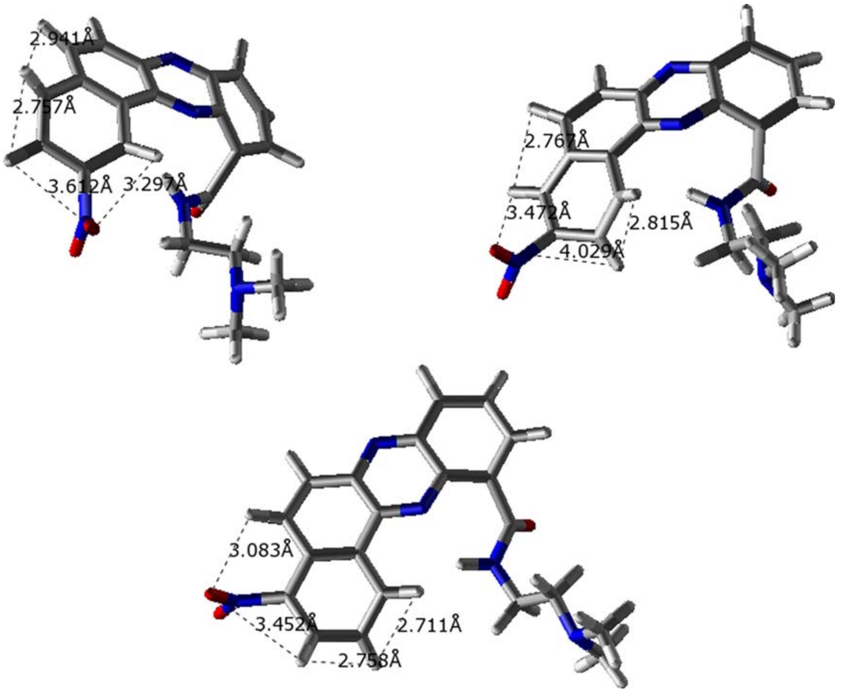


Figure 4. Different interatomic distances that show their influence in the descriptor values at the compounds **8**, **9**, and **10**.

molecules according to their structural features when the other parameters remain constant. In short, the essence of the difference among the molecules resides in this parameter that includes the 3D part of the descriptor.

If we observe once again Table 6, it can be seen that for the same sign *s* and weighting, there are great differences among the values of the descriptors; therefore, the r_{ij} is the parameter in which we should focus our efforts to better interpret these descriptors.

3. Conclusion

In this study we modeled the anticancer activity, specifically the inhibition of topoisomerases I and II, expressed by the cytotoxicity of sixty-four compounds (IC_{50}) against drug-resistant human small cell lung carcinoma line cell H69/LX4 which overexpresses P-glycoprotein (P-gp). For this purpose, we employed the 3D-MoRSE descriptors.

The results produced by the methodology we proposed were superior to the other descriptors such as; Topological, BCUT, Galvez topological charge indexes, 2D autocorrelations, Randić molecular profile, Geometrical, RDF, and WHIM; taking into account the statistical parameters of the model and the cross-validation results.

The variables that were found to be most significant in describing the model were atomic masses, atomic polarizabilities, atomic Sanderson electronegativities, and atomic van der Waals volumes.

The analysis of the outliers present also suggested to us that a certain conformation is necessary, as plane as possible in the aromatic rings regarding the substituents, for the interaction between the compounds and the receptor.

4. Materials and methods

4.1. Data set

Our data set consisted of 64 benzophenazines with dual topoisomerase I and topoisomerase II inhibitors as anticancer agents, with the necessary concentration to reduce the cell number to 50% (IC_{50}) reported.[9]

We used only 64 compounds of a total set of 72, because the other 8 benzophenazines were reported with an inexact IC_{50} concentration measurement higher than 5000 nM and could not be used in the multiple regression analysis. The cytotoxicity was measured using the drug-resistant human small cell lung carcinoma line cell H69/LX4 which over expresses P-glycoprotein (P-gp). After addition of the cytotoxics, the cells were incubated for 5–6 days before adding Alamar blue (H69/LX4) to measure cell proliferation.[9]

4.2. Computational strategies

The DRAGON[24] computer software, version 2.1, was employed to calculate all the molecular descriptors included in this work. We carried out geometry optimization calculations for each compound using the quantum chemical semi-empirical method AM1[25] included in MOPAC 6.0[26] before calculating the DRAGON descriptors.

The mathematical models were obtained by means of the multiple regression analysis (MRA) as implemented in the STATISTICA software version 6.0.[27] The genetic algorithm (GA) was used as the variable selection strategy, in order to include in the equation the most significant parameters from the data set.

The GA is a class of methods based on biological evolution rules.[28–30] The first step is to create a population of linear regression models. These regression models mate with each other, mutate, cross-over, reproduce, and then evolve through successive generations toward an optimal solution. The GA simulation conditions were 10,000 generations, the number of crossovers = 5000, smoothness factor = 1, mutation probability for adding new term = 50%, and 300 model populations.

Analysis of residuals and deleted residuals from the regression equations was used to identify outliers. The statistical significance of the models was determined by examining the squared regression coefficient (R^2), the standard deviation (S), the Fisher ratio (F), and the ratio between cases and adjustable parameters (ρ). This last statistical parameter was a very important criterion to determine when to stop the introduction of variables in the development of the model. The formula of ρ is given below.

$$\rho = \frac{N}{X} \tag{8}$$

where N is the number of compounds included in the model and X is the number of adjustable parameters in the equation of the model. A necessary condition for the development of a linear model is $\rho > 4$ and needs to be kept in mind when introducing variables.[31]

4.3. Orthogonalization of descriptors

The orthogonalization process of molecular descriptors was introduced by Randić several years ago as a way of improving the statistical interpretation of the model built by using interrelated indices.[32–36] The main philosophy of this approach is to avoid the exclusion of descriptors on the basis of their collinearity with other variables previously included in the model. The acceptable level of collinearity to avoid is a more subjective issue. In our view, the collinearity of the variables should be as low as possible because the interrelatedness among the different descriptors can result in a highly unstable regression coefficient, which makes it impossible to know the relative importance of an index and underestimates the utility of the regression coefficient model.

The Randić method of orthogonalization has been described in detail in several publications,[32–36] thus, we will only give a general overview here. The first step in orthogonalizing the molecular descriptors in a model is to select the appropriate order of orthogonalization, which, in this case, is the order in which the variables were selected in the genetic algorithm search procedure of the linear regression analysis. The first variable Mor07m is taken as the first orthogonal descriptor Ω^1Mor07m, and the second one is orthogonalized with respect to it by taking the residual of its correlation with Ω^1Mor07m. The process is repeated until all the variables are completely orthogonalized, and the orthogonal variables are then used to obtain the new model.

4.4. Validation of the models

The models obtained were validated calculating the cross-validated squared regression coefficient (q^2) values. The q^2 values are calculated from 'leave-one-out' (LOO) test and from 'leave-group-out' (LGO) test, also known as cross-validation.

For LOO a data point is removed from the set, and the regression recalculated; the predicted value for that point is then compared to its actual value. This is repeated until each datum has been omitted once; the sum of squares of these deletion residuals can then be used to calculate q^2, an equivalent statistic to R^2. In the LGO method, 25% of the data was eliminating every time in one hundred different forms. In this way, we guaranteed not to use the same group of compounds for every validation. The q^2 values can be considered a measure of the predictive power of a regression equation: whereas R^2 can always be increased artificially by adding more parameters (descriptors), q^2 decreases if a model is overparameterized[20] and is therefore a more meaningful summary statistic for QSAR models.

4.5. Comparison with other approaches

The use of 3D-MoRSE descriptors for the prediction of anticancer activity, explained in the previous section, was compared with other methodologies. The Topological,[37] BCUT,[38–40] Galvez topological charge indexes,[41–44] 2D autocorrelations,[45–47] Randić molecular profile,[48,49] Geometrical,[37] RDF,[50] and WHIM[51–57] descriptors were calculated.

Eight models using these methodologies were developed with the same data set as the QSAR models. The comparison was made of the quality of the statistical parameters of the regression and the predictive capability of the models generated.

These yielded additional criteria to compare the quality of different models. One of these criteria was formulated by Akaike in 1973.[58,59] Akaike's information criterion (AIC) takes into account the statistical goodness of fit and the number of parameters that have to be estimated to achieve that degree of fit. This criterion is calculated using the following equation:

$$\mathrm{AIC} = \mathrm{RSS} \cdot \frac{(n + p\prime)}{(n - p\prime)^2} \tag{9}$$

where as RSS is the sum of the squared differences between the observed (y) and estimated response (y'), n is the number of compounds in the training set, and p' is the number of adjustable parameters in the model.

When comparing models, the model that produces the minimum AIC value should be considered potentially the most useful.

Another criterion that we used to compare the quality of the developed models was the Kubinyi function (FIT) (Eq. 10). This function, closely related to the F value, was created and proved to be a useful parameter assessing the quality of the models.[60,61]

$$\mathrm{FIT} = \frac{R^2 \cdot (n-k-1)}{(n+k^2)\cdot(1-R^2)} \quad (10)$$

where n is the number of compounds in the training set, k is the number of variables in the equation that describe the model, and R^2 is the squared correlation coefficient.

The main disadvantage of the F value is its sensitivity to changes in k, if k is small, and its lower sensitivity if k is large. The FIT criterion has a low sensitivity toward changes in k values, as long as they are small numbers, and a substantially increasing sensitivity for large k values.[60,61] Finally, the best model will present the highest value of this function.

Acknowledgments

We acknowledge the Universidad de Vigo and Programa de cooperación inter-institucional VLIR-UCLV for financial support. Marta Teijeira thanks the Xunta de Galicia for the Parga Pondal contract. Last but not least, we acknowledge to the Alan Saiz for your useful comments and suggestions.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2006.05.081.

References and notes

1. Corbett, A. H.; Osheroff, N. *Chem. Res. Toxicol.* **1993**, *6*, 585–597.
2. Dancey, J.; Eisenhauer, E. A. *Br. J. Cancer* **1996**, *74*, 327–338.
3. Seeber, S.; Osieka, R.; Schmidt, C. G.; Achterrath, W.; Crooke, S. T. *Cancer Res.* **1982**, *42*, 4719–4725.
4. Hendricks, C. B.; Rowinsky, E. K.; Grochow, L. B.; Donehower, R. C.; Kaufmann, S. H. *Cancer Res.* **1992**, *52*, 2268–2278.
5. Riou, J. F.; Fosse, P.; Nguyen, C. H.; Larsen, A. K.; Bissery, M. C.; Grondard, L.; Saucier, J. M.; Bisagni, E.; Lavelle, F. *Cancer Res.* **1993**, *53*, 5987–5993.
6. Finlay, G. J.; Riou, J. F.; Baguley, B. C. *Eur. J. Cancer* **1996**, *A*, 708–714.
7. Utsugi, T.; Aoyagi, K.; Asao, T.; Okazaki, S.; Aoyagi, Y.; Sano, M.; Wierzba, K.; Yamada, Y. *Jpn. J. Cancer Res.* **1997**, *88*, 992–1002.
8. Wang, S.; Miller, W.; Milton, J.; Vicker, N.; Stewart, A.; Charlton, P.; Mistry, P.; Hardick, D.; Denny, W. A. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 415–418.
9. Vicker, N.; Burgess, L.; Chuckowree, I. S.; Dodd, R.; Folkes, A. J.; Hardick, D. J.; Hancox, T. C.; Miller, W.; Milton, J.; Sohal, S.; Wang, S.; Wren, S. P.; Charlton, P. A.; Dangerfield, W.; Liddle, C.; Mistry, P.; Stewart, A. J.; Denny, W. A. *J. Med. Chem.* **2002**, *45*, 721–739.
10. González, M. P.; Dias, L. C.; Helguera, A. M.; Rodriguez, Y. M.; de Oliveira, L. G.; Gomez, L. T.; Diaz, H. G. *Bioorg. Med. Chem.* **2004**, *12*, 4467–4475.
11. González, M. P.; Gonzalez Diaz, H.; Molina Ruiz, R.; Cabrera, M. A.; Ramos de Armas, R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1192–1199.
12. Gonzalez-Diaz, H.; Bastida, I.; Castañedo, N.; Nasco, O.; Olazabal, E.; Morales, A.; Serrano, H. S.; de Armas, R. R. *Bull. Math. Biol.* **2004**, *66*, 1285–1311.
13. Gonzalez-Diaz, H.; Gia, O.; Uriarte, E.; Hernadez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Model. (Online)* **2003**, *9*, 395–407.
14. González, M. P.; Caballero, J.; Tundidor-Camba, A.; Helguera, A. M.; Fernandez, M. *Bioorg. Med. Chem.* **2006**, *14*, 200–213.
15. Garg, R.; Denny, W. A.; Hansch, C. *Bioorg. Med. Chem.* **2000**, *8*, 1835–1839.
16. Mekapati, S. B.; Denny, W. A.; Kurupm, A.; Hansch, C. *Bioorg. Med. Chem.* **2001**, *9*, 2757–2762.
17. Verma, R. P. *Bioorg. Med. Chem.* **2005**, *13*, 1059–1067.
18. Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030–1037.
19. Schuur, J. H.; Selzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334–344.
20. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.
21. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
22. Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238–1244.
23. Todd, A. K.; Adams, A.; Thorpe, J. H.; Denny, W. A.; Wakelin, L. P.; Cardin, C. J. *J. Med. Chem.* **1999**, *42*, 536–540.
24. Todeschini, R.; Consonni, V.; Pavan, M. Dragon Software version 2.1, 2002.
25. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
26. Frank, J. Seiler Research Laboratory, US Air Force Academy, Colorado Springs CO, 1993.
27. Statsoft, I. STATISTICA (data analysis software system)version 6.0, 2002.
28. Vedani, A.; Dobler, M. *Prog. Drug Res.* **2000**, *55*, 105–135.
29. Tropsha, A.; Zheng, W. *Curr. Pharm. Des.* **2001**, *7*, 599–612.
30. Hasegawa, K.; Funatsu, K. *SAR QSAR Environ. Res.* **2000**, *11*, 189–209.
31. Garcia-Domenech, R.; de Julian-Ortiz, J. V. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445–449.
32. Klein, D. J.; Randić, M.; Babić, D.; Lučić, B.; Nikolić, S.; Trinajstić, N. *Int. J. Quant. Chem.* **1991**, *63*, 215–222.
33. Randić, M. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45–59.
34. Randić, M. *New J. Chem.* **1991**, *15*, 517–525.
35. Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–320.
36. Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, D. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532–538.
37. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Mannheim, 2000.
38. Burden, F. R. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.
39. Burden, F. R. *Quant. Struct.-Act. Relat.* **1997**, *16*, 309–314.

40. Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28–35.
41. Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520–525.
42. Galvez, J.; Garcia-Domenech, R.; de Gregorio Alapont, C.; de Julian-Ortiz, J. V.; Popa, L. *J. Mol. Graph.* **1996**, *14*, 272–276.
43. Galvez, J.; Garcia-Domenech, R.; de Julian-Ortiz, J. V.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 272–284.
44. Rios-Santamarina, I.; Garcia-Domenech, R.; Galvez, J.; Cortijo, J.; Santamaria, P.; Morcillo, E. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 477–482.
45. Moran, P. A. P. *Biometrika* **1950**, *37*, 17–23.
46. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359–360.
47. Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 757–764.
48. Randic, M. *New J. Chem.* **1995**, *19*, 781–791.
49. Randic, M. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 373–382.
50. Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. *Fresenius J. Anal. Chem.* **1997**, *359*, 50–55.
51. Gramatica, P.; Consonni, V.; Todeschini, R. *Chemosphere* **1999**, *38*, 1371–1378.
52. Gramatica, P.; Corradi, M.; Consonni, V. *Chemosphere* **2000**, *41*, 763–777.
53. Gramatica, P.; Navas, N.; Todeschini, R. *Chemom. Intell. Lab. Syst.* **1998**, *40*, 53–63.
54. Todeschini, R.; Gramatica, P. *Quant. Struct.-Act. Relat.* **1997**, *16*, 113–119.
55. Todeschini, R.; Gramatica, P. *Quant. Struct.-Act. Relat.* **1997**, *16*, 120–125.
56. Todeschini, R.; Gramatica, P.; Provenzani, R.; Marengo, E. *Chemom. Intell. Lab. Syst.* **1995**, *27*, 221–229.
57. Todeschini, R.; Lasagni, M.; Marengo, E. *J. Chemom.* **1994**, *8*, 263–273.
58. Akaike, H. In *Second International Symposium on Information Theory*; B. N. Petrov; F. Csaki, Eds.; Akademiai Kiado: Budapest, 1973, pp 267–281.
59. Akaike, H. *IEEE Trans. Autom. Control* **1974**, *AC-19*, 713–716.
60. Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285–294.
61. Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393–401.