12th International Electronic Conference on Synthetic Organic Chemistry (ECSOC-12) 1-30 November 2008
http://www.usc.es/congresos/ecsoc/12/ECSOC12.htm & http://www.mdpi.org/ecsoc-12

[G0006]

Theoretical Prediction of Antiproliferative Activity against Murine Leukemia Tumor Cell Line (L1210). 3D-Morse Descriptors and its Application in Computational Chemistry.

Liane Saíz-Urra [a,c] ,Yunierkis Pérez-Castillo [a,c], Maykel Pérez González[c], Reinaldo Molina Ruiz [a,b], M. Natália D. S. Cordeiro [a] , J. Enrique Rodríguez-Borges [b,*] , Xerardo García-Mera[d]

[a]*REQUIMTE,* [b]*CIQ,* Chemistry Department, Faculty of Sciences, University of Porto, 4169-007 Porto Portugal. [c]Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, Villa Clara, C.P. 54830, Cuba. [d]Departamento de Química Orgánica, Facultade de Farmacia, Universidade de Santiago de Compostela, E-15706 Santiago de Compostela, Spain.

ABSTRACT: Cancer is among the top ten causes of death in the world but in spite of the efforts of the pharmaceutical companies and many governmental organizations, new and more effective drugs are urgently needed. Computer assisted studies have been widely used to predict anticancer activity taking into account different molecular descriptors, statistical techniques, cell lines and data sets of congeneric and non-congeneric compounds. This paper describes a QSAR study and the successful application of 3D-MoRSE descriptors for developing Linear Discriminant Analysis (LDA) to predict the anticancer potential of a diverse set of indolocarbazoles derivatives. Despite the structural complexity of this sort of compounds the used descriptors are able to identify the most remarkable features like the incidence of polarizability of the substituents and the interatomic distance in the 7-azaindole moiety in the antiproliferative activity. A comparison with other approaches such as the Getaway, Randić molecular profile, Geometrical, RDF descriptors, was carried out showing the model with 3D-MoRSE descriptors resulted in the best accuracy and predictive capability. An LDA based desirability analysis was conducted to select the levels of the predictor variables which should generate more desirable drugs, i.e. with higher posterior probability to be classified cytotoxic.

Keywords: QSAR; Anticancer activity; Indolocarbazoles derivatives; 3D-MoRSE.

*Corresponding author. Fax: +351 226082959. e-mail: jrborges@fc.up.pt

Introduction

One of the most important issues in Medicinal Chemistry is cancer, which encompasses a group of diseases characterized by the excessive and uncontrolled growth of cells invading and impairing tissues and organs and can, eventually, result in the death. In 2005, 7.6 million of the 58 million deaths registered in the world were caused by cancer. Over 70% of these deaths were in countries with low or average incomes where the resources for the diagnosis and the treatment of the disease are limited or even nonexistent [1].

The search for new anticancer drugs plays a central role in the research programs of pharmaceutical companies but also those of many governmental organizations [2]. However, it is estimated that the rate of incidence of cancer far from decreasing will rise to about 9 million in 2015 and 11.4 million in 2030. Hence, new and effective drugs are increasingly and urgently needed. A large number of anticancer agents have been discovered that act at different levels [3] and have higher efficacy and lower toxicity than existing treatments. These databases can be exploited with the help of automated and multivariate data analysis methods [4-6]. The latter relates the molecular structures with their biological properties by establishing computational models able to assign activity values to new untested compounds [7, 8].

QSAR techniques in anticancer activity studies have previously reported the use of different molecular descriptors, statistical techniques, cell lines and data sets of congeneric and non-congeneric compounds as well as the respective toxicological assays of these compounds [4, 9-17].

An interesting group of compounds is indolocarbazole derivatives whose properties as protein kinase C and topoisomerase I inhibitors have been widely studied [18, 19]. Rebeccamycin, a microbial metabolite isolated from cultures of Saccharothrix aerocolonigenes, which belong to this group, is an antitumor antibiotic that inhibits topoisomerase I by stabilizing the topoisomerase I-DNA cleavable complex [20, 21]. Also, it has been shown that although topoisomerase I is a target for most of the rebeccamycin derivatives, the inhibition of other enzymes may also be a contributing factor to their cytotoxicity. However, its toxicity prohibits its use in cancer chemotherapy. Structure–activity relationship studies have been carried out with the purpose of improving the pharmacological profile of rebeccamycin,[19, 22] and have led to the development of a schematic representation of a drug-topoisomerase I-DNA ternary complex.

In spite of its promise and the diversity and quantity of derivatives developed, no anticancer QSAR studies taking into account these types of compounds have been reported. In this paper we report a QSAR model for the rational selection of anticancer compounds which involves a diverse data set of indolocarbazoles derivatives. The use of the 3D descriptors is reported as well, specifically the 3D-MoRSE, owing to the flexibility of these descriptors. 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 to compare the datasets comprising of molecules of different sizes, and number of atoms [23, 24].

Materials and methods

Data sets:

In the present study we used a data set of 125 compounds whose anticancer activity against murine leukemia tumor cell line (L1210) has been previously reported. Eligible compounds were determined by reviewing the literature [22, 25-37].

The data encompasses rebeccamycin analogues from indolo[2,3-*c*]carbazole, indolocarbazoles bearing amino acid residues, sugar units linked to both indole nitrogens, 7-azaindole moieties or different substituents on the indolocarbazole framework. Another group consists of dipyrrolo[3,4-a:3,4-c]carbazole-1,3,4,6-tetraones, substituted with various saturated and unsaturated side chains, indolylpyrazolones and indolylpyridazinedione. Finally, we studied isogranulatimide and bis-imide granulatimide analogues modified on the indole moiety and on the imide heterocycles. Cytotoxicity was measured by the microculture tetrazolium assay as described by Leonce, S. *et. al.* [38]. Results are expressed as IC_{50}, the concentration at which the optical density of treated cells with respect to untreated controls is reduced by 50%.

Resulting from the need for more potent and less toxic new anticancer drugs, we established the threshold value of activity IC_{50} equal to 10 µM, thereby only the compounds with an activity value lower than the aforementioned were considered as active.

In order to obtain validated QSAR models the dataset was divided into training and test sets. Ideally, this division should be performed such that points representing both training and test sets are distributed within the whole descriptor space occupied by the entire dataset, and each point of the test set is close to at least one point of the training

set. This partitioning ensures that a similar principle can be employed for the activity prediction of the test set. For this reason, k-Means Cluster Analysis (k-MCA) was employed to split the set of compounds and achieve the desired distribution.

k-Means cluster analysis

The k-MCA may be used in training and test sets design [39]. The idea is to partition the set of compounds under study into several statistically representative classes of chemicals. Then the training and test sets can be selected from the members of these classes. This procedure ensures that any chemical class (as determined by the clusters derived from k-MCA) will be represented in both compound series (training and test). It allows one to design both, training and test sets, which are representative of the entire "experimental universe". Figure 1 illustrates graphically the above-described procedure.

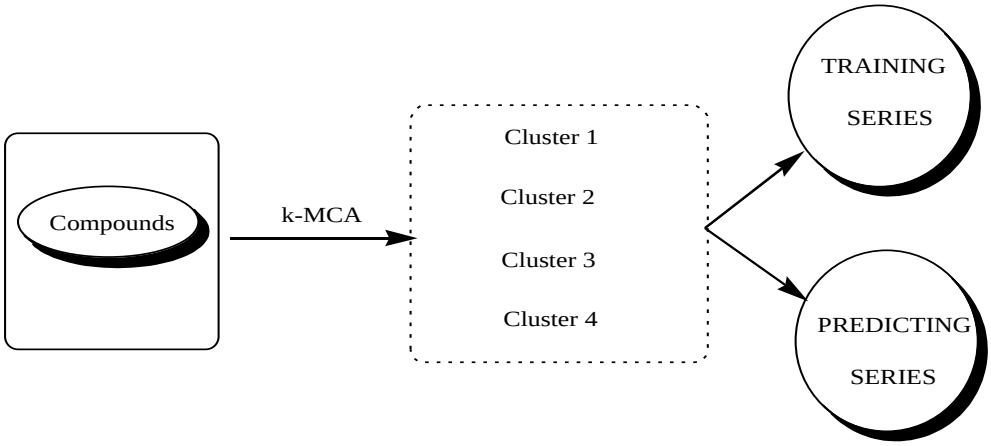

Figure 1. Training and Predicting series design throughout k-MCA.

The k-MCA was carried out for active and non-active compounds by two separate analyses. The first involved 68 active compounds, which were split into five clusters with 1, 17, 18, 16 and 16 members respectively, whereas the second analysis yielded four clusters containing 4, 16, 10 and 27 members for a total of 57 non-active compounds.

Selection of the test set was carried out by taking the compounds with the minor Euclidean distance in each cluster. We took into account the number of members in each cluster and the standard deviation of the variables in the cluster (with the goal of making it as low as possible) to ensure a statistically acceptable partitioning of the data into several clusters. We also examined among and within clusters for variance the Fisher ratio and their p-level of significance which were considered to be lower than

0.05. The variables which were finally used in the analysis showed p-levels < 0.05 for Fisher test. The results are depicted in Table 1 and Table 2.

Table 1. Analysis of variance between and within clusters.

Active compounds set

	Between clusters	**Within clusters**	**F**	**Significance**
*Mor*11*v*	52.96	14.04	59.41	$<10^{-6}$
*Mor*25*v*	57.12	9.88	91.01	$<10^{-6}$
*Mor*25*p*	58.03	8.97	101.86	$<10^{-6}$

Non-active compounds set

	Between clusters	**Within clusters**	**F**	**Significance**
*Mor*11*v*	36.53	19.47	33.15	$<10^{-6}$
*Mor*25*v*	46.91	9.09	91.22	$<10^{-6}$
*Mor*25*p*	46.65	9.35	88.14	$<10^{-6}$

Table 2. Analysis of the descriptive statistics of the variables in each cluster.

Active compounds set

Descriptive Statistics	**Variables**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Mean	*Mor*11*v*	3.447	0.952	-0.480	0.401	-1.088
	*Mor*25*v*	2.578	-0.065	-0.268	-1.130	1.339
	*Mor*25*p*	2.806	0.010	-0.270	-1.175	1.293
Standard Deviation	*Mor*11*v*	0	0.497	0.282	0.463	0.607
	*Mor*25*v*	0	0.335	0.358	0.370	0.508
	*Mor*25*p*	0	0.332	0.345	0.340	0.479
Variance	*Mor*11*v*	0	0.247	0.080	0.214	0.368
	*Mor*25*v*	0	0.112	0.128	0.137	0.258
	*Mor*25*p*	0	0.110	0.119	0.116	0.230

Non-active compounds set

Descriptive Statistics	Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Mean	*Mor*11*v*	-0.811	-0.186	1.703	-0.400
	*Mor*25*v*	2.454	0.085	0.886	-0.742
	*Mor*25*p*	2.397	0.062	0.940	-0.740
Standard Deviation	*Mor*11*v*	1.164	0.664	0.601	0.461
	*Mor*25*v*	0.976	0.281	0.680	0.183
	*Mor*25*p*	1.005	0.299	0.663	0.199
Variance	*Mor*11*v*	1.355	0.441	0.361	0.213
	*Mor*25*v*	0.953	0.079	0.463	0.034
	*Mor*25*p*	1.010	0.089	0.439	0.040

Computational strategies:

The DRAGON [40] computer software, version 5.4, was employed to calculate all the molecular descriptors included in this study. We carried out geometry optimization calculations for each compound using the quantum chemical semi-empirical method AM1 [41] included in MOPAC 6.0 [42] before calculating the DRAGON descriptors.

Mathematical models were obtained by means of Linear Discriminant Analysis (LDA) as implemented in STATISTICA software version 6.0 [43]. Forward stepwise was employed as the variable selection strategy. The quality of the model was determined by examining: the Wilk's lambda (λ), the Mahalanobis distance (D^2), the Fisher's ratio (F), and the corresponding p-level (p(F)). The percentage of good classifications and the proportion between the cases and adjustable parameters ($\rho>4$) in the equation were examined as well. The Mahalanobis distance indicates the separation of the respective groups, showing whether the model possesses an appropriate discriminatory power for differentiating between the two groups. It establishes a perfect discrimination for $\lambda = 0$ and not discrimination when $\lambda = 1$

The values 1 and -1 were used to classify compounds as active and inactive respectively. Finally, the posteriori probabilities were used to classify the compounds as anticancer or not against murine leukemia tumor cell line (L1210).

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 which had been built by using interrelated indices [44-48]. The main tenet of this approach is to avoid the exclusion of descriptors on the basis of their collinearity with other variables previously included in the model. In our view, the collinearity of the variables should be as low as possible because interrelatedness among the different descriptors can result in a highly unstable regression coefficient. Making it impossible to know the relative importance of an index and underestimating the utility of the regression model coefficients. The Randić method of orthogonalization has been described in detail in several publications [44-48].

Identifying outliers:

An analysis of the applicability domain of the model was carried out to explore the presence of potential outliers and compounds that influence model parameters resulting in an unstable model. The test set was included to check how adequate the model was for the external prediction.

The leverage values were calculated for every compound and plotted vs. standard residuals (Y-axis). Then, the domain of applicability of the model was defined as a squared area within the ±2 band for residuals and a leverage threshold [49, 50]

The leverage (h) of a compound in the original variable space which measures its influence on the model is defined as:

$$h_i = x_i^T (X^T X)^{-1} x \qquad (i = 1, \ldots, n) \qquad \text{Eq. 1}$$

where x_i is the descriptor vector of the considered compound and $\boldsymbol{X}$ is the model matrix derived from the training set descriptor values. The warning leverage h^* is defined as follows:

$$h^* = 3 \times p'/n \qquad \text{Eq. 2}$$

where n is the number of training compounds and p' is the number of model adjustable parameters.

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 Getaway [51],

Randić molecular profile [52, 53], Geometrical [51], RDF [54] and WHIM [55-61] descriptors were calculated.

In order to make the comparison on the same basis, all models were developed using the same data set and included four variables from the 3D descriptors of the DRAGON software. Also, we focussed on the quality of the statistical parameters of the discriminant established above and the predictive capability of the models generated.

Desirability analysis:

A common problem in drug design is to choose a set of conditions or levels for the independent variables, in our case being molecular descriptors, which generates the most desirable product in terms of output values of the predictor variables. The procedure of optimization involves two main steps: (1) predicting responses on the property under study (in this case, the anticancer potency), by fitting the observed responses using an equation based on the levels of the predictive variables (in this case, the 3D-MoRSE descriptors), and (2) finding the levels of the X-variables that simultaneously produce the most desirable predicted responses on the studied property [62]. In the present study, the desirability analysis was carried out with STATISTICA 6.0 [43], by setting the current level of each predictor variable to the respective mean. Curvature s and t parameters were fixed at 1.00 considering the linear form of the function used to perform the discriminant analysis. Spline method was selected for fitting the desirability function and surface/contours maps.

Results and discussion:

The final partition of the data resulted in a structurally representative distribution of chemicals into training and predicting series. A training set of 55 active and 45 non-active compounds was created and a test set of 13 active and 12 non-active compounds (see Tables 1 and 2 in Supplementary Material). It is worth noting that cluster 1 from the active compounds involves only one member; case **110,** which has been included in the training set.

At first sight, it could be considered a potential outlier; however its inclusion in this QSAR study might be important due to the structural information that it can provide (see Figure 2).

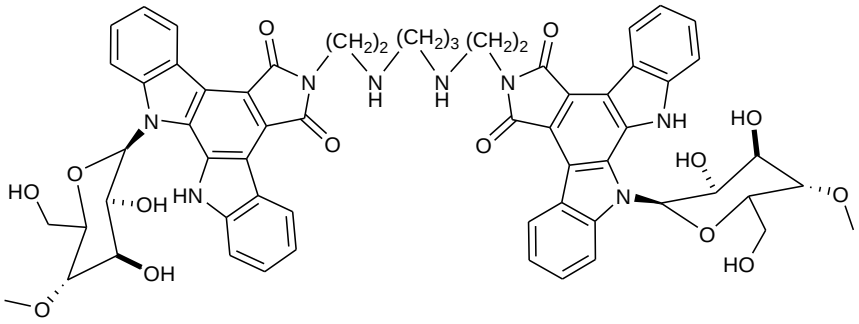

Figure 2. Compound number **110**.

After the application of the LDA, the best model obtained was:

$$A_{act} = -1.43 \cdot Mor31p - 1.41 \cdot Mor11v + 15.16 \cdot Mor25p - 14.60 \cdot Mor25v + 0.60 \qquad \text{Eq. 3}$$

λ=0.610 **F**(4.95) = 15. 187 **p** < 0.0001 **C** = 0.698 $\boldsymbol{\rho}$ = 9.0

Where A_{act} is a dummy variable [63] with A_{act} =1 for active compounds and A_{act} = -1 for non-active ones. The Wilk's statistic (λ=0.610) Fischer Ratio (**F**(4.95) = 15.187), and significance level (**p** < 0.0001) of the parameters were assessed [7]. In addition, we controlled the Matthew's coefficient (**C** = 0.698) [64] and the cases/adjustable parameters ratio ($\boldsymbol{\rho}$ = 9.0) taking into account the smallest group into the classification [65]. It can be seen that the model is statistically significant (**p** < 0.05) and the correlation coefficient is always between -1 and +1. A value of -1 indicates total disagreement, +1 total agreement and 0 for completely random predictions. The correlation coefficient may often provide a much more balanced evaluation of the prediction than, for instance, the percentage [66]. In this case the high value for the Matthew's indicates a strong linear relationship between the molecular descriptors and the output of the model [67].

Finally, the high value of $\boldsymbol{\rho}$ = 9.0 shows that the model is not over-fitted by an excess of parameters; this parameter is expected to be >4 for linear models [65]. This discriminant model showed excellent results in the training and external prediction series used to validate the model, as can be seen in Table 3.

Table 3. Training and Predictability analysis results

Including non classified compounds

	Training (85.00% total)			Prediction (84.00% total)		
	%	Actives	Non-actives	%	Actives	Non-actives
Actives	85.45	47	8	84.62	11	2
Non-actives	84.44	7	38	83.33	2	10

Without non classified compounds

	Training (87.37% total)			
	%	Actives	N. actives	N. classified
Actives	88.68	47	6	2
N. actives	83.72	6	36	2
	Prediction (84.00% total)			
	%	Actives	N. actives	N. classified
Actives	84.62	11	1	1
N. actives	83.33	2	9	1

In spite of achieving adequate statistical results from the 3D-MoRSE descriptors family, we thought that it was not enough to say that our model design was appropriate. Therefore, we carried out a comparison of different methodologies to validate our model. The results obtained from this comparison are given in Table 4. The variables derived using DRAGON are given in Table 3 of Supplementary Material including the 3D-MoRSE descriptors employed. All models were developed by using the same training and test sets.

Table 4. The statistical parameters of the linear discriminant models obtained for all methodologies included in the comparison.

	METHODOLOGIES					
PARAMETERS	**Geometrical**	**Getaway**	**Randić molecular profile**	**RDF**	**WHIM**	**3D-MoRSE**
Variables	J3D, QZZv, QYYe, DISPp	H5v, HATS8e, HATS8p, RCON	DP02, DP15, DP16, SP08	RDF020e, RDF060p, RDF080p, RDF090p	G1u, G2u, E1u, L2m	Mor31p, Mor11v, Mor25p, Mor25v

λ	0.694	0.816	0.851	0.777	0.765	0.610
F(4.95)	10.45	5.35	4.15	6.81	7.29	15.19
p	<0.0000	<0.0006	<0.0038	<0.0001	<0.0000	<0.0001
C	0.477	0.421	0.352	0.389	0.556	0.698
% Total	74.00	71.00	68.00	70.00	78.00	85.00
% Actives	74.55	89.09	85.45	78.18	80.00	85.45
% Inactives	73.33	48.89	46.67	60.00	75.56	84.44

From an inspection of Table 4 it is quite clear that the best results are provided by the 3D-MoRSE descriptors since they have the highest accuracy (**% Total** = 85.00) and discriminatory power (**λ** = 0.610) as well as the best correlation between the molecular descriptors and the output of the model (**C** = 0.698).

After demonstrating the superiority of the 3D-MoRSE descriptors using other methodologies, we used Randić orthogonalization to avoid collinearity between the variables. The QSAR model obtained (Eq. 4) after this procedure is given below, together with the statistical parameters.

$$A_{act} = -1.08 \cdot^{1} \Omega Mor31p - 1.18 \cdot^{2} \Omega Mor11v + 0.73 \cdot^{3} \Omega Mor25p - 14.60 \cdot^{4} \Omega Mor25v + 0.60 \quad \text{Eq. 4}$$

λ=0.610 **F**(4.95) = 15. 187 **p** < 0.0001 **C** = 0.698 ***ρ*** = 9.0

We also performed ROC curve analysis to examine our classifier against a random one. A pronounced ROC curve is depicted in Figure 3 with an area under curve markedly higher than 0.5- which is the threshold value expected for a random classifier (diagonal line)[68].

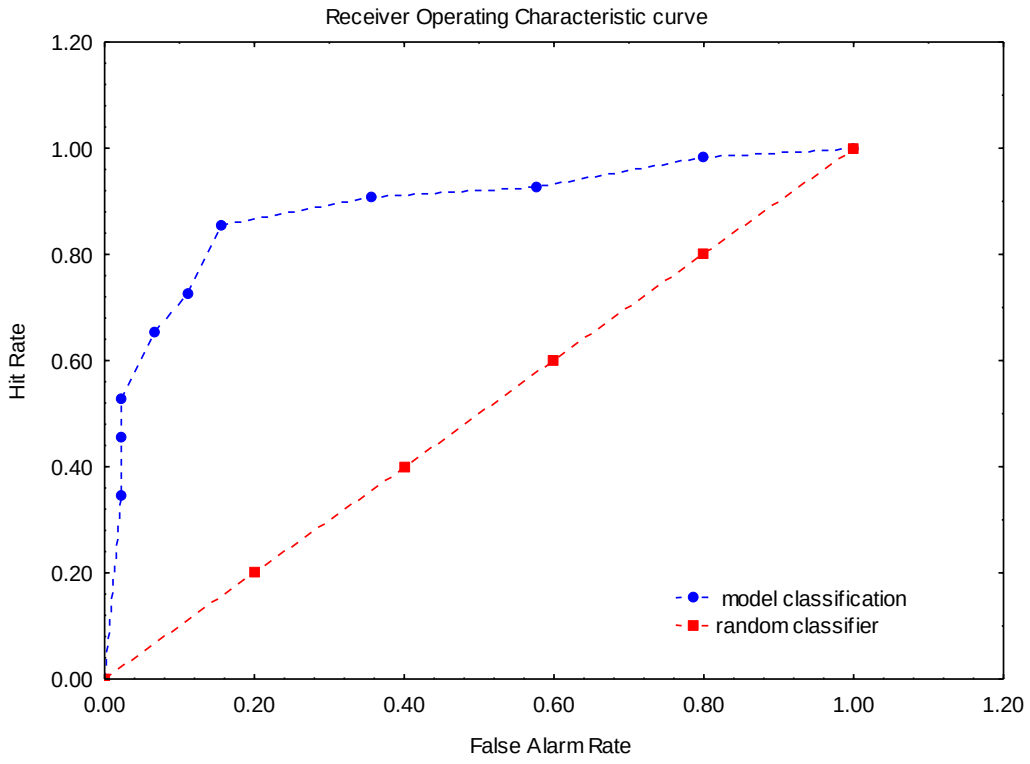

Figure 3. ROC curve of the training set. Comparison with the random classifier.

Finally, an analysis of the applicability domain of the model was carried out and the results can be seen in the next figure.

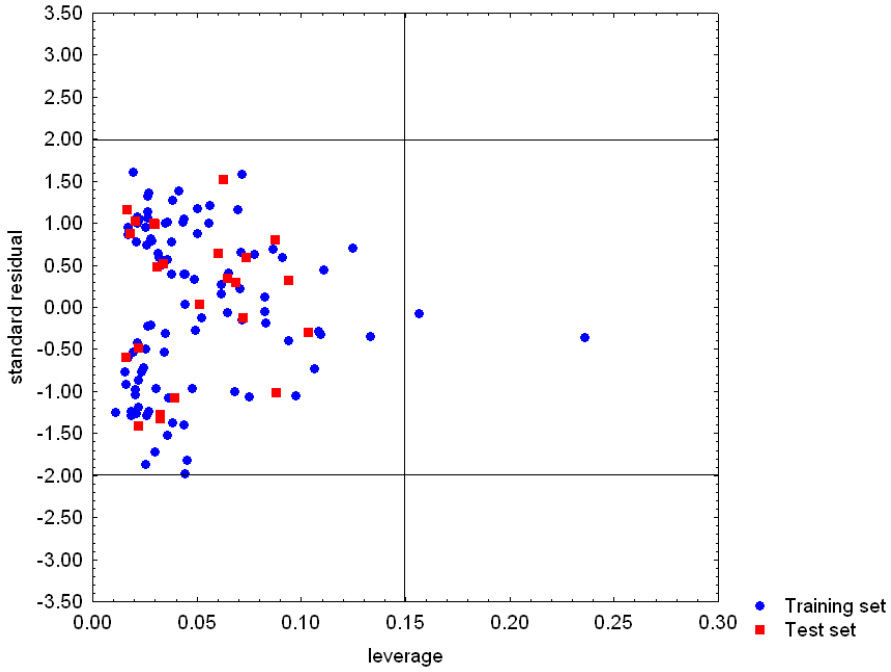

Figure 4. Applicability domain of the model.

As it can be seen from the chart, two compounds in the training set are outside the domain of applicability of the model due to their leverage values (see figure 5 for structural details). Nevertheless, none of these compounds were considered as outliers

because their values of standardized residuals are not greater than two standard deviation units.

A deeper analysis showed that compound **14** has a very similar leverage value to the threshold value established previously $h^* = 0.15$ ($h_{14} = 0.156$) while compound **110** shows the highest value ($h_{110} = 0.236$). Both compounds have similar chains attached to the N-imide since they encompass an aliphatic chain with amine groups which are also the longest ones in the data but in the specific case of compound **110** it has been explained before how it is included in a single cluster due to its chemical structure. However, it is worthy noting that it is not considered an outlier as the results of the cluster analysis had suggested.

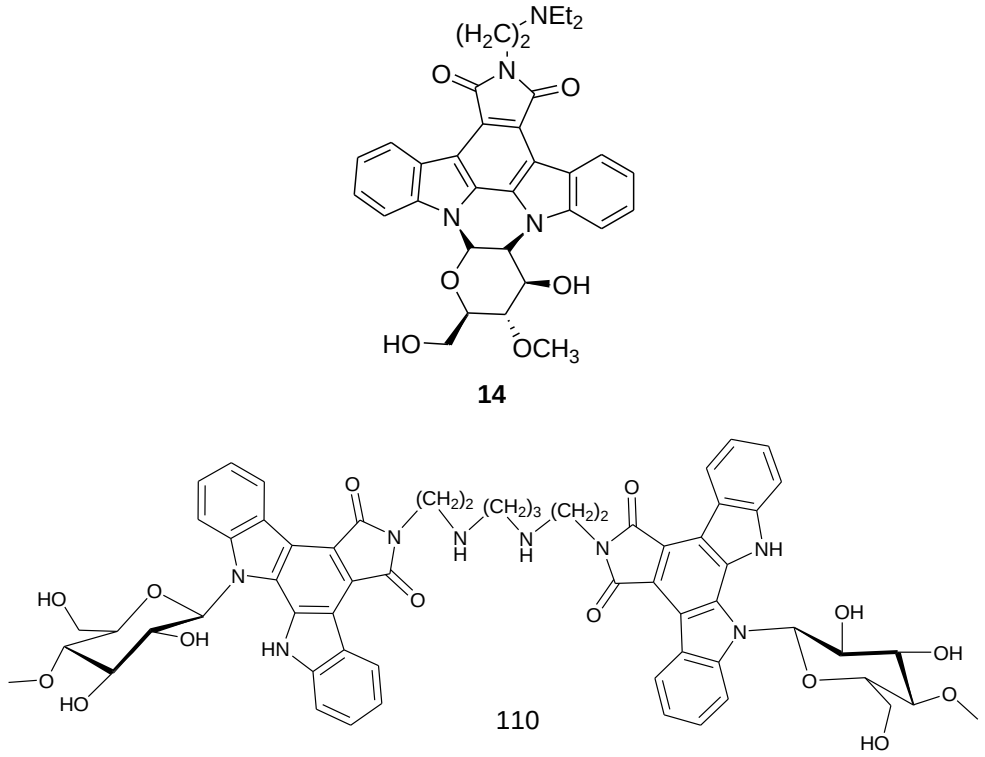

Figure 5. Compounds from the training set that fall out of the domain of applicability of the model.

Consequently, a new model was developed by removing the two compounds of the training set that were determined to be out of the domain to examine its effect on the statistical parameters of the model (see Table 5 for comparison). As a result of the analysis, no significant variations resulted in the model parameters. It follows that, the influence of these compounds is not critical for the model and they were not excluded because of the aforementioned conclusion about the possible importance of the structural information provided by the molecule **110** for the successful development of the model.

Table 5. Parameters of the former model and the variations after removing the potential influential compounds (**14** and **110**)

Model	*b*	*Mor*31*p*	*Mor*11*v*	*Mor*25*p*	*Mor*25*v*	% Total	% Actives	% Inactives
Eq. 4	0.60	-1.08	-1.18	0.73	-14.60	85.00	85.45	84.44
Under study	0.62	-1.06	-1.19	0.68	-14.12	84.69	86.79	82.22

Once equation 4 was determined to be the best model, a desirability analysis was performed based on the levels of the predictor variables used in the model. The optimal values for obtaining the highest probability to be classified as a potential anticancer compound should be about -0.354, -0.277, 1.071 and -0.108 for Mor31p, Mor11v, Mor25p and Mor25v, respectively, fixing the other three variables at their present mean values (see Table 6 and Figure 6). However, if the current values for the four variables (-0.128, 0.235, 0.485 and 0.413 respectively) are used, it is possible to obtain a desirability value for the anticancer potency of 0.630.

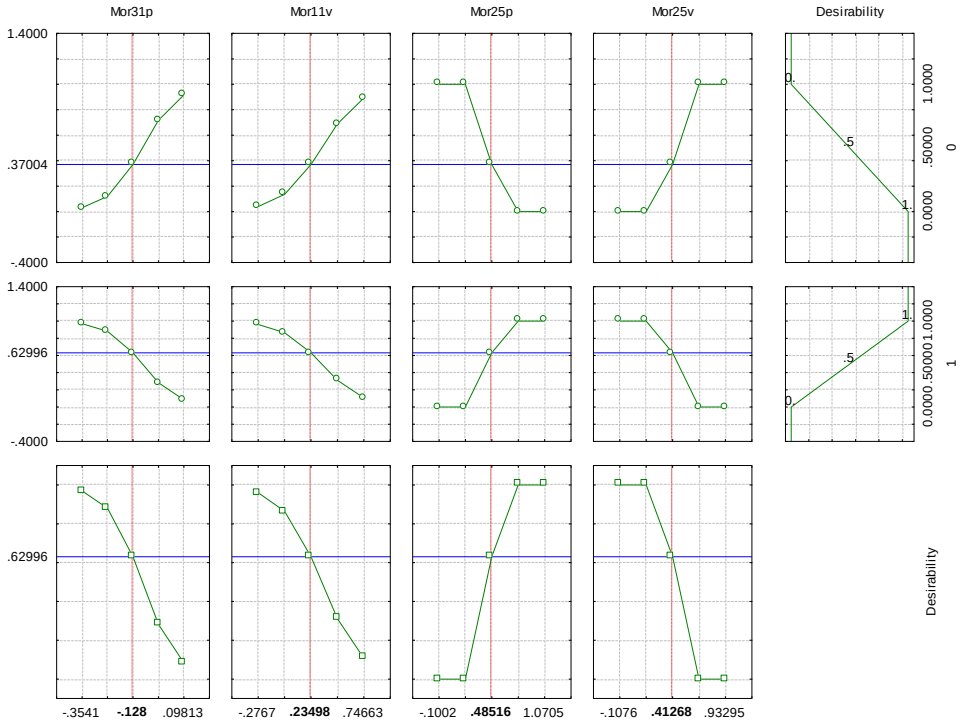

Figure 6. Profiles for posterior probabilities and desirability.

Table 6. Optimal values for every variable in the model to obtain the highest desirability values remaining the mean values of the other three variables

Variables	Factor levels	P. P. (inactive)	P. P. (active)	Desirability values
Mor31p	-0.354	0.030	0.970	0.970
Mor11v	-0.277	0.040	0.960	0.960
Mor25p	1.071	0.000	1.000	1.000
Mor25v	-0.108	0.000	1.000	1.000

On the other hand, the contour plots in Figure 7 show the overall response desirability produced by different level combinations of the two independent variables (fixing the value of the remaining two variables at their mean values). The plots were created by transforming the previous scores of each of the four variables into desirability scores (they could range from 0.0 - undesirable (in green) to 1.0 - very desirable (in red). The red zone in the contour plots represents the higher probability of obtaining a drug with the best anticancer profile.

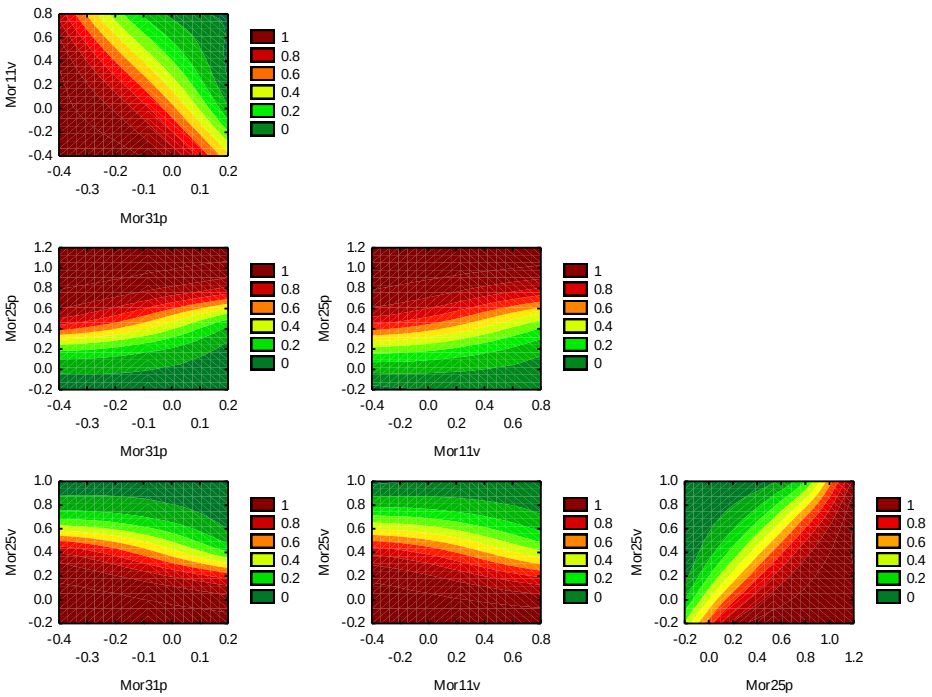

Figure 7. Desirability Surface/Contours of the posterior probabilities and desirability resulting from the classification model (Eq. X); Method: Spline Fit.

To obtain an insight into the activity-structure relationship we interpreted the descriptors in the model and their influence in the anticancer profile of some molecules to find possible patterns.

The 3D-MoRSE descriptors are based on the following equation:

$$I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i \cdot w_j \cdot \frac{\sin(s \cdot r_{ij})}{s \cdot r_{ij}} \qquad \text{Eq. 5}$$

where $I(s)$ is the scattered electron intensity, w is an atomic property, r_{ij} are the interatomic distances between the ith and jth atoms respectively, s measures the scattering angle and A is the number of atoms.

In this study, the variables in the model are related to the atomic polarizabilities and atomic van der Waals volumes but at different scattering angle values. It is important to note the sinusoidal nature of the relation between $I(s)$ and s showing how s is the determining parameter of the 3D-MoRSE descriptor's sign and its contribution to the classification when the atomic properties are the same. It can be observed from the mean values of the variables described in the desirability analysis explained above, how the Mor31p tends to have negative values while the rest of them show positive values. Consequently, both variables weighted by polarizability Mor31p and Mor25p; contribute positively to the classification of the antiproliferative activity. From the analysis of figure 8 it can be seen how the higher the polarizability of the substituent, the higher the posterior probability to be considered active is for the rebeccamycin analogues in the data.

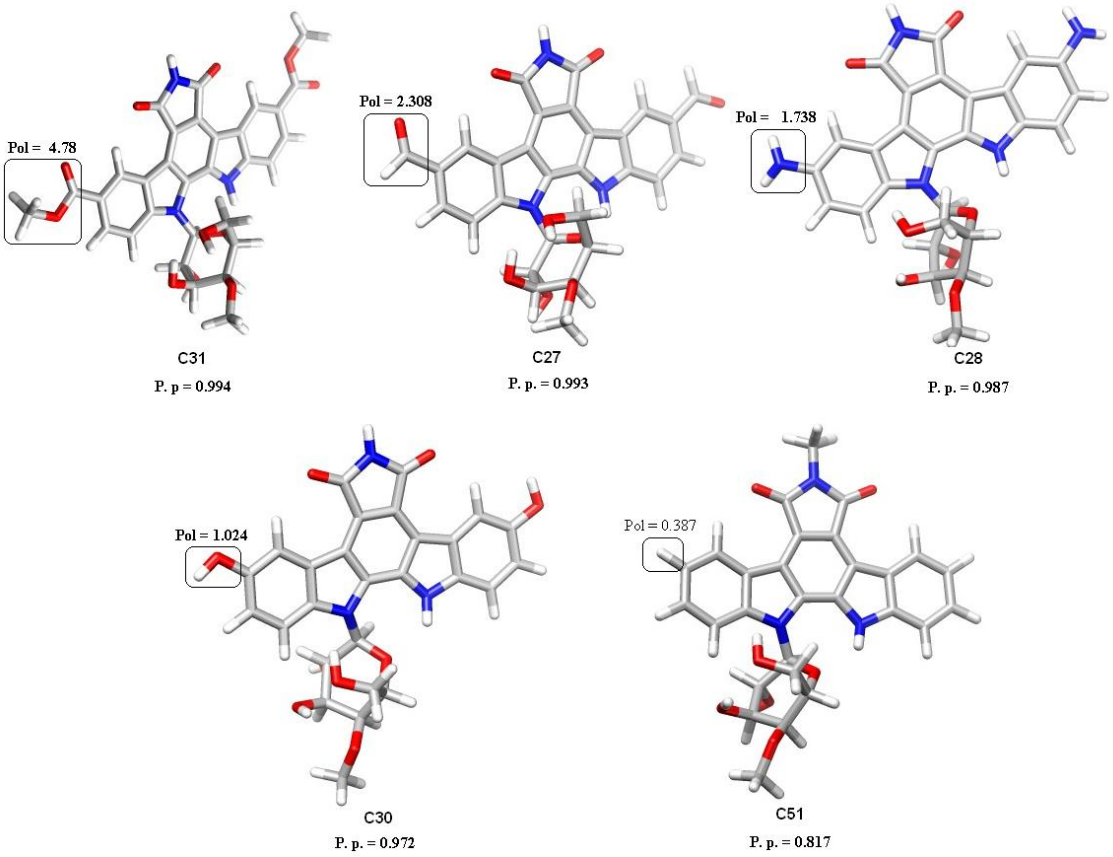

Figure 8. Influence of the polarizability in the classification

However, when the nature of the substituent is the same, their position is critical for the activity. This behavior is observed in a comparison between rebeccamycin analogues bearing one 7-azaindole moiety, which are position isomers such as compounds **53** and **57,** resulting in a lost of activity when the sugar framework is linked to the N-pyridine.

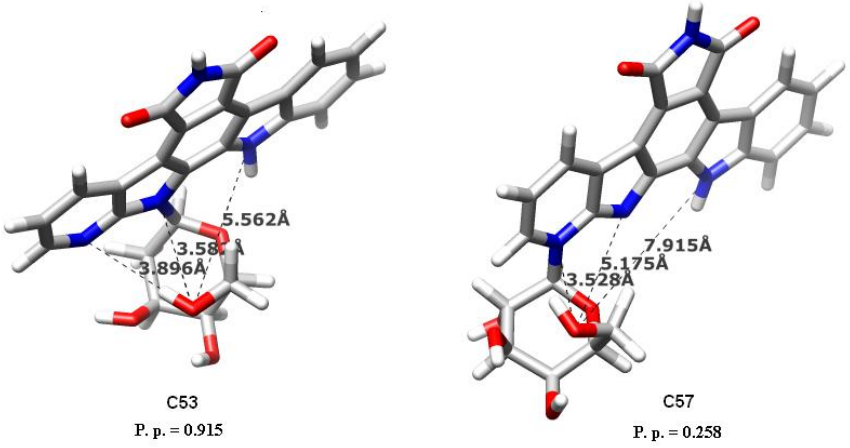

Figure 9. Influence of the interatomic distance in the classification.

Conclusions

In summary, a QSAR model was developed by using 3D-MoRSE descriptors towards the rational selection of anticancer compounds considering the antiproliferative activity against murine leukemia tumor cell line (L1210) of a structurally and pharmacologically diverse data set of indolocarbazoles derivatives.

A comparison with the Getaway, Randić molecular profile, Geometrical, RDF and WHIM descriptors was carried out and the model with 3D-MoRSE descriptors had the best accuracy and predictive capability.

Additionally, desirability analysis based on the LDA model yielded the optimal descriptor values that a drug candidate should have for guaranteeing antiproliferative activity. The weights of the variables that were found to be most significant in describing the model were atomic polarizabilities and atomic van der Waals volumes.

However, despite the ability of the alternative QSAR model proposed here to predict accurately the anticancer potential of drugs, further study is recommended. New QSAR models employing Multiple Regression Linear technique should be developed in order to establish quantitative relations more specific to the antiproliferative activity since it is not possible to discern the difference in activity between compounds but only the probability to be classified as active.

Acknowledgments

The authors acknowledge the Portuguese *Fundação para a Ciência e a Tecnologia* (FCT) (**SFRH/BDP/24512/2005**) and Cuban Higher Education Ministry (R&D project number 6.181-2006) for financial support.

References

[1] Acción mundial contra el cáncer - Versión rev., 2005.

[2] Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials and Approval. Teicher BA, Andrews PA, editors. Second Edition. Totowa, New Jersey: Humana Press; 2004. Biomedecine & Pharmacotherapy, vol. 59, 2005. pp. 137.

[3] R. Kaplow, Nurs. Clin. North. Am. 40 (2005) 77-94.

[4] L. Saiz-Urra, M. P. GonzalezM. Teijeira, Bioorg Med Chem. 14 (2006) 7347-7358.

[5] A. H. Morales, M. A. Perez, R. D. CombesM. P. González, Toxicology. 220 (2006) 51-62.

[6] A. H. Morales, P. R. Duchowicz, M. A. Cabrera, E. A. Castro, N. CordeiroM. P. González, Chemom Intell Lab Syst. 81 (2006) 180 - 187.

[7] H. Van Waterbeemd. Discriminant Analysis for Activity Prediction. In: H. Van Waterbeemd, editor. Chemometric methods in molecular design, vol. 2. New York: Wiley-VCH, 1995. pp. 265-282.

[8] P. Willett, Perspect. Drug Discov. Des. 7-8 (1997) 1-11.

[9] A. M. Helguera, J. E. Rodriguez-Borges, X. Garcia-Mera, F. FernandezM. N. Cordeiro, J Med Chem. 50 (2007) 1537-1545.

[10] H. Gonzales-Diaz, O. Gia, E. Uriarte, I. Hernadez, R. Ramos, M. Chaviano, S. Seijo, J. A. Castillo, L. Morales, L. Santana, D. Akpaloo, E. Molina, M. Cruz, L. A. TorresM. A. Cabrera, J Mol Model. 9 (2003) 395-407.

[11] A. H. Morales, M. A. Cabrera Perez, M. P. González, R. M. RuizH. Gonzalez-Diaz, Bioorg Med Chem. 13 (2005) 2477-2488.

[12] A. H. Morales, M. A. Cabrera, R. D. Combes P. González M, Current Computer-Aided Drug Design. 1 (2005) 237-255.

[13] D. Amic, D. Davidovic-Amic, D. Beslo, V. Rastija, B. LucicN. Trinajstic, Curr Med Chem. 14 (2007) 827-845.

[14] H. Assefa, S. KamathJ. K. Buolamwini, J Comput Aided Mol Des. 17 (2003) 475-493.

[15] B. A. BhongadeA. K. Gadad, Bioorg Med Chem. 12 (2004) 2797-2805.

[16] R. Garg, W. A. DennyC. Hansch, Bioorg Med Chem. 8 (2000) 1835-1839.

[17] L. Saiz-Urra, M. P. GonzalezM. Teijeira, Bioorg Med Chem. 15 (2007) 3565-3571.

[18] U. Pindur, Y. S. KimF. Mehrabani, Curr Med Chem. 6 (1999) 29-69.

[19] M. Prudhomme, Curr Med Chem. 7 (2000) 1189-1212.

[20] D. E. Nettleton, T. W. Doyle, B. Krishnan, G. K. MatsumotoJ. Clardy, Tetrahedron Letters. 26 (1985) 4011-4014.

[21] J. A. Bush, B. H. Long, J. J. Catino, W. T. BradnerK. Tomita, J Antibiot (Tokyo). 40 (1987) 668-678.

[22] A. Voldoire, M. Sancelme, M. Prudhomme, P. Colson, C. Houssier, C. Bailly, S. LeonceS. Lambel, Bioorg Med Chem. 9 (2001) 357-365.

[23] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. SteinhauerV. Steinhauer, J. Chem. Inf. Comput. Sci. 36 (1996) 1030-1037.

[24] J. H. Schuur, P. SelzerJ. Gasteiger, J. Chem. Inf. Comput. Sci. 36 (1996) 334-344.

[25] P. Moreau, M. Sancelme, C. Bailly, S. Leonce, A. Pierre, J. Hickman, B. PfeifferM. Prudhomme, Eur J Med Chem. 36 (2001) 887-897.

[26] C. Marminon, M. Facompre, C. Bailly, J. Hickman, A. Pierre, B. Pfeiffer, P. RenardM. Prudhomme, Eur J Med Chem. 37 (2002) 435-440.

[27] C. Marminon, F. Anizon, P. Moreau, S. Leonce, A. Pierre, B. Pfeiffer, P. RenardM. Prudhomme, J Med Chem. 45 (2002) 1330-1339.

[28] C. Marminon, A. Pierre, B. Pfeiffer, V. Perez, S. Leonce, P. RenardM. Prudhomme, Bioorg Med Chem. 11 (2003) 679-687.

[29] P. Moreau, N. Gaillard, C. Marminon, F. Anizon, N. Dias, B. Baldeyrou, C. Bailly, A. Pierre, J. Hickman, B. Pfeiffer, P. RenardM. Prudhomme, Bioorg Med Chem. 11 (2003) 4871-4879.

[30] C. Marminon, A. Pierre, B. Pfeiffer, V. Perez, S. Leonce, A. Joubert, C. Bailly, P. Renard, J. HickmanM. Prudhomme, J Med Chem. 46 (2003) 609-622.

[31] S. Messaoudi, F. Anizon, S. Leonce, A. Pierre, B. PfeifferM. Prudhomme, Eur J Med Chem. 40 (2005) 961-971.

[32] H. Henon, F. Anizon, R. M. Golsteyn, S. Leonce, R. Hofmann, B. PfeifferM. Prudhomme, Bioorg Med Chem. 14 (2006) 3825-3834.

[33] S. Messaoudi, F. Anizon, P. Peixoto, M. H. David-Cordonnier, R. M. Golsteyn, S. Leonce, B. PfeifferM. Prudhomme, Bioorg Med Chem. 14 (2006) 7551-7562.

[34] E. Conchon, B. Aboab, R. M. Golsteyn, F. Cruzalegui, T. Edmonds, S. Leonce, B. PfeifferM. Prudhomme, Eur J Med Chem. 41 (2006) 1470-1477.

[35] B. Hugon, F. Anizon, C. Bailly, R. M. Golsteyn, A. Pierre, S. Leonce, J. Hickman, B. PfeifferM. Prudhomme, Bioorg Med Chem. 15 (2007) 5965-5980.

[36] E. Conchon, F. Anizon, B. Aboab, R. M. Golsteyn, S. Leonce, B. PfeifferM. Prudhomme, Eur J Med Chem. (2007).

[37] H. Henon, S. Messaoudi, F. Anizon, B. Aboab, N. Kucharczyk, S. Leonce, R. M. Golsteyn, B. PfeifferM. Prudhomme, Eur J Med Chem. 554 (2007) 106-112.

[38] S. Leonce, V. Perez, M. R. Casabianca-Pignede, M. Anstett, E. Bisagni, A. PierreG. Atassi, Invest New Drugs. 14 (1996) 169-180.

[39] W. R. DillonM. Goldstein. Multivariate analysis: Methods and applications. N. Y.: Wiley, 1984.

[40] R. Todeschini, V. ConsonniM. Pavan. Dragon Software 2002.

[41] M. J. S. Dewar, E. G. Zoebisch, E. F. HealyJ. J. P. Stewart, J Am Chem Soc. 107 (1985) 3902-3909.

[42] J. Frank. MOPAC. Seiler Research Laboratory, US Air Force Academy, Colorado Springs CO, 1993.

[43] I. Statsoft. STATISTICA (data analysis software system). 2002.

[44] D. J. Klein, M. Randić, D. Babić, B. Lučić, S. NikolićN. Trinajstić, Int J Quant Chem. 63 (1991) 215-222.

[45] M. Randić, J Mol Struct (Teochem). 233 (1991) 45-59.

[46] M. Randić, New J Chem. 15 (1991) 517-525.

[47] M. Randić, J Chem Inf Comput Sci. 31 (1991) 311-320.

[48] B. Lučić, S. Nikolić, N. TrinajstićD. Jurić, J Chem Inf Comput Sci. 35 (1995) 532-538.

[49] H. Gonzalez-Díaz, S. Vilar, L. SantanaE. Uriarte, Bioorganic & Medicinal Chemistry. 15 2544-2550.

[50] L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowellP. Gramatica, Environ Health Perspect. 111 (2003) 1361-1375.

[51] R. TodeschiniV. Consonni. Handbook of Molecular Descriptors, 1. Edition ed.: Wiley-VCH, Mannheim 2000.

[52] M. Randić, New J. Chem. 19 (1995) 781-791.

[53] M. Randić, J Chem Inf Comp Sci. 35 (1995) 373-382.

[54] J. Gasteiger, J. Schuur, P. Selzer, L. SteinhauerV. Steinhauer, Fresen J Anal Chem. 359 (1997) 50-55.

[55] P. Gramatica, V. ConsonniR. Todeschini, Chemosphere. 38 (1999) 1371 - 1378.

[56] P. Gramatica, M. CorradiV. Consonni, Chemosphere. 41 (2000) 763 - 777.

[57] P. Gramatica, N. NavasR. Todeschini, Chemom Intell Lab Syst. 40 (1998) 53 - 63.

[58] R. TodeschiniP. Gramatica, Quant Struct - Act Relat. 16 (1997) 113 -119.

[59] R. TodeschiniP. Gramatica, Quant Struct - Act Relat. 16 (1997) 120 - 125.

[60] R. Todeschini, P. Gramatica, R. ProvenzaniE. Marengo, Chemom Intell Lab Syst. 27 (1995) 221 - 229.

[61] R. Todeschini, M. LasagniE. Marengo, J Chemom. 8 (1994) 263 - 273.

[62] G. DerringerR. Suich, J. Quality Technol. 12 (1980) 214–219.

[63] R. B. KowalskiS. Wold. In Handbook of Statistics. Amsterdam: North Holland Publishing, 1982.

[64] B. W. Matthews, Biochim Biophys Acta. 405 (1975) 442-451.

[65] R. Garcia-DomenechJ. V. de Julian-Ortiz, J Chem Inf Comput Sci. 38 (1998) 445-449.

[66] P. Baldi, S. Brunak, Y. chauvin, C. A. F. AndersenH. Nielsen, Bioinformatics Review. 16 (2000) 412-424.

[67] Z. Yuan, FEBS Lett 451 (1999) 23-26.

[68] P. Póvoa, L. Coelho, E. Almeida, A. Fernandes, R. Mealha, P. MoreiraH. Sabino, Clin Microbiol Infect 11 (2005) 101-108.