A TOPological Sub-structural Molecular Design (TOPS-MODE)-QSAR approach for modeling the antiproliferative activity against murine leukemia tumor cell line (L1210)

ARTICLE *in* BIOORGANIC & MEDICINAL CHEMISTRY · JANUARY 2009

Impact Factor: 2.79 · DOI: 10.1016/j.bmc.2008.11.084 · Source: PubMed

7 AUTHORS, INCLUDING:

Reinaldo Molina-Ruiz
Central University "Marta Abreu" of Las Villas
37 PUBLICATIONS **1,124** CITATIONS
SEE PROFILE

Jose Borges
University of Porto
103 PUBLICATIONS **894** CITATIONS
SEE PROFILE

Yunierkis Perez Castillo
Universidad Técnica Particular de Loja
15 PUBLICATIONS **273** CITATIONS
SEE PROFILE

Natália D. S. Cordeiro
University of Porto
245 PUBLICATIONS **3,040** CITATIONS
SEE PROFILE

Bioorganic & Medicinal Chemistry 17 (2009) 537–547

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

A TOPological Sub-structural Molecular Design (TOPS-MODE)-QSAR approach for modeling the antiproliferative activity against murine leukemia tumor cell line (L1210)

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

[a] CIQ, Department of Chemistry, University of Porto, 4169-007 Porto, Portugal
[b] Molecular Simulation and Drug Design, Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba
[c] Departamento de Química Orgánica, Faculdade de Farmácia, Universidad de Santiago de Compostela, E-15782 Santiago de Compostela, Spain
[d] REQUIMTE, Department of Chemistry, University of Porto, 4169-007 Porto, Portugal

ARTICLE INFO

Article history:
Received 20 May 2008
Revised 25 November 2008
Accepted 29 November 2008
Available online 13 December 2008

This paper is dedicated to Maykel Pérez González *in memoriam*

Keywords:
Anticancer activity
Indolcarbazole
Rebeccamycin
Leukemia antitumor agents

ABSTRACT

Lately, Quantitative Structure–Activity Relationship (QSAR) studies have been afar used to predict anticancer activity taking into account different molecular descriptors, statistical techniques, cell lines and data set of congeneric and non-congeneric compounds. Herein we report a QSAR study based on a TOPological Sub-structural Molecular Design (TOPS-MODE) approach, aiming at predicting the anticancer leukemia activity of a diverse data set of indolocarbazoles derivatives. Finally, several aspects of the structural activity relationships are discussed in terms of the contribution of different bonds to the anticancer activity, thereby making the relationship between structure and biological activity more transparent.

1. Introduction

Cancer encompasses a group of diseases characterized by the excessive and uncontrolled growth of cells that invade and impair tissues and organs and can ultimately result in death. The search for new anticancer drugs plays a central role in the research programs of pharmaceutical companies and also of many governmental organizations.[1] Despite these efforts, the World Health Organization[2] estimates that the rate of incidence of such diseases will increase by 50% by the year 2020. For this reason, new and effective drugs are urgently needed. In recent years, 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.[4,5] The latter relates the molecular structures with their biological properties by establishing computational models able to assign activity values to new untested compounds.[6,7]

Here, the role of Quantitative Structure–Activity Relationship (QSAR) techniques should be remarked as these 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.[8–16] In particular, it should be highlighted the success of the topological substructural molecular design (TOPS-MODE) approach in producing adequate QSAR models to classify the anticancer activity of several types of organic compounds. For instance, Estrada et al.[8] developed a TOPS-MODE QSAR-based model that was able to discriminate anticancer compounds from the inactive ones in a training series. Specifically, the authors obtained an overall rate of ca. 88% and 91% correct classifications for the training set and external set compounds, respectively.

An interesting group of compounds are the indolocarbazole derivatives whose properties such as protein kinase C and topoisomerase I inhibitors have been widely studied.[17,18] Rebeccamycin (Fig. 1), a microbial metabolite isolated from cultures of *Saccharothrix aerocolonigenes* which belongs to this group, is an antitumor antibiotic that inhibits topoisomerase I by stabilizing the topoisomerase I–DNA cleavable complex.[19,20] Also, it has been shown that although topoisomerase I is a target for most rebeccamycin derivatives, the inhibition of other enzymes may also be a contributing factor to their cytotoxicity (e.g., such as protein kinase C). However, its toxicity prohibits its use in cancer chemotherapy. SAR studies

* Corresponding author. Tel.: +351 220402502; fax: +351 220402659.
E-mail address: ncordeir@fc.up.pt (M. Natália D.S. Cordeiro).

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.11.084

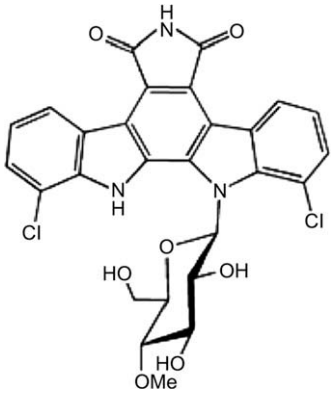


Figure 1. Rebeccamycin scaffold.

have been carried out with a view to improving the pharmacological profile of rebeccamycin,[18,21] and have led to establishing a model of a drug–topoisomerase I–DNA ternary complex. However, despite their attractiveness, no QSAR study regarding the anticancer activity of this kind of compounds has been reported so far.

For these reasons, the present work reports a QSAR modeling study for the rational selection of anticancer compounds by using the TOPS-MODE approach along with a diverse data set of indolocarbazoles derivatives.

2. Materials and methods

2.1. Data set

In the present study, we used a data set of 123 compounds whose anticancer activity against murine leukemia tumor cell line (L1210) has been reported previously. Eligible compounds were collected from several sources in the literature.[21–34]

Among this data are included 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 encompasses 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, isogranulatimide and bis-imide granulatimide analogues modified on the indole moiety and on the imide heterocycles are included as well. Cytotoxicity was measured by the microculture tetrazolium assay as described by Leonce et al.[35] Endpoint activities are expressed as IC_{50}, that is, the concentration at which the optical density of treated cells with respect to untreated controls is reduced by 50%. According to the need of more potent and less toxic new anticancer drugs, we took into consideration a threshold value of activity IC_{50} equal to 10 μM, thereby only the compounds with an activity value lower than that were considered as active.

In order to obtain validated QSAR models the dataset should be divided into training and test sets. Ideally, this division must be performed in such a way 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. In this work, we have applied the *k*-Means Cluster Analysis (*k*-MCA) technique to split the set of compounds and achieve the desired distribution.

2.2. *k*-Means cluster analysis

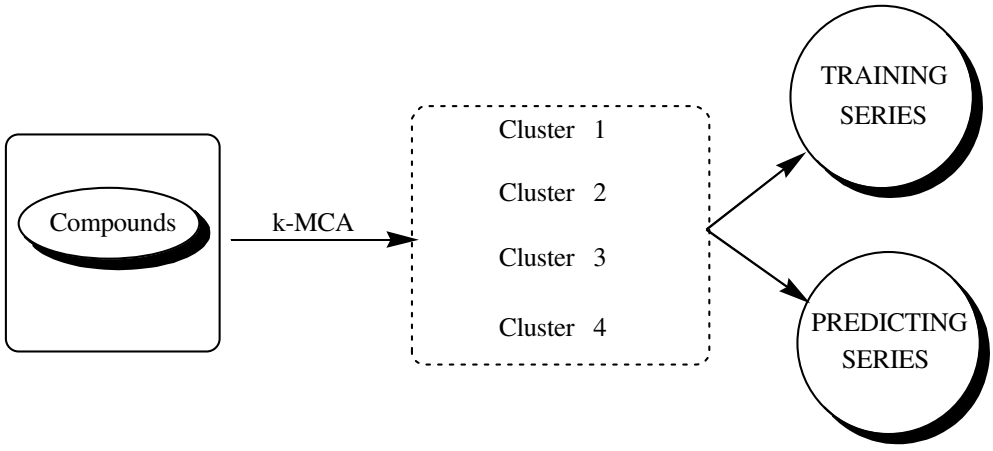
The *k*-MCA can be used in training and test set design.[36] The idea is to carry out a partition of the set of compound under study in several statistically representative classes of chemicals. The training and test sets can then be selected from the members of all these classes. This procedure ensures that any chemical class (as determined by the clusters derived from *k*-MCA) will be represented in both series' of compounds (training and test). This allows the design of both training and test sets that are representative of the entire 'experimental universe' (see Scheme 1).

The *k*-MCA was carried out for active and inactive compounds in two separated analyses. The first analysis one involved 70 active compounds, which were split into five clusters with 23, 1, 18, 11 and 17 members, whereas the second analysis yield four clusters containing 10, 18, 24 and 1 members from a total of 53 inactives compounds.

Selection of the test set was carried out by taking the compounds with the lowest 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 (as low as possible) to ensure a statistically acceptable data partition into several clusters. We also examined the following factors both between and within clusters: the variance, the Fisher ratio and the *p*-level of significance, which was considered acceptable when it was below 0.05. The variables that were finally used in the analysis showed *p*-levels <0.05 on the Fisher test. The results are shown in Tables 1 and 2.

2.3. QSAR modeling

A QSAR modeling approach seeks to uncover correlation of biological activity with structural features of the compounds (*descriptors*). Nowadays there is a vast amount of available molecular



Scheme 1. Training and predicting series design throughout *k*-MCA.

Table 1
Analysis of variance between and within clusters

Set	Analysis	Variables		
		μ_1^{Mol}	$\mu_1\mu_{15}^{Hyd}$	$\mu_1\mu_{15}^{Hyd}$
Active compounds	Between clusters	59.85	61.06	57.53
	Within clusters	9.15	7.94	11.47
	F	106.31	124.92	81.80
	p-Level	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$
Non-active compounds	Between clusters	44.42	45.25	45.36
	Within clusters	7.57	6.75	6.64
	F	95.79	109.52	111.62
	p-Level	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$

descriptors with which one can model the activity of interest. Here we have resorted to the graphed-based molecular descriptors—spectral moments, designed accordingly to the TOPS-MODE approach, whose theoretical basis has been widely described in previous reports.[37–39] To mirror fundamental physicochemical properties that might relate to the present biological endpoint (anticancer activity), the atomic properties bond hydrophobicity (*Hyd*), standard bond dipole moment (*Dip*) and molar refractivity (*MR*) were used as bond weights.

These graph-based descriptors were computed with the MODESLAB 1.5 software,[40] from the familiar SMILES inputting of chemical structures. Specifically, we have calculated the first 15 spectral moments (μ_1–μ_{15}) for each bond weight and the number of bonds in the molecules (μ_0), excluding the hydrogen atoms. Owing to the non-linearity of the biological process under study, the interactions between μ_0 or μ_1 with all variables were also evaluated.

As to the modeling technique, we opted for a discriminant-based approach; in this case, the coefficients and statistical parameters were obtained by the linear discriminant analysis (LDA) implemented in STATISTICA (version 6.0).[41]

In summary, we adopted the following 3-step procedure for establishing our QSAR model.

Step 1: Model set-up. This proceeds as follows.

(i) Select an appropriate training set of chemicals by *k*-MCA.
(ii) Input the molecular structures, through the SMILES codes.
(iii) Return the spectral-moments descriptors using an appropriate set of bond weights.
(iv) Find an adequate QSAR model from the training set by a discriminant-based approach.

Step 2: Model evaluation. Assess the performance of the QSAR model, particularly regarding its applicability and predictive power.

Step 3: Fragments activity identification. Ascertain the activity of the different fragments based on the bond contributions computed with the QSAR model.

2.4. Model set-up

The task is to obtain a mathematical function Eq. 1 that best describes the anticancer activity *P*, as a linear combination of the predictor *X*-variables (the spectral moments μ_k) weighted with the coefficients a_k. Such coefficients are to be optimized by means of LDA using the training set compounds.

$$P = a_1X_1 + a_2X_2 + \cdots + a_kX_k + a_0 \quad (1)$$

In developing the models, *P* values of +1 and −1 were assigned to active and inactive compounds[42], respectively, but a posteriori probabilities are used instead to assert the models' classification of compounds. In particular, when the probability of being active did not differ more than 5% from that of being inactive, the case was considered as unclassified by the model.

In most cases, LDA is suitably applied in QSAR studies as long as the problem of selection of variables is faced and solved. In so doing, particular care should be taken to avoid *overfitted* models, that is, models with too many included predictor variables and poor external predictivity. Also, one should be aware of possible collinearity among the variables. Moreover, one should easily detect the presence of influential outliers and then modify the data set accordingly.

In this work, the forward stepwise (FS) technique was applied to select the molecular descriptors (*X*-variables) with the highest influence on the anticancer activity. In addition, to tackle the multicollinearity problem and avoid overfitting, we have orthogonalized our descriptor variables according to the Randić procedure.[43] Such procedure was introduced by Randić, several years ago, as a way of improving the interpretation of models built from interrelated indices. Furthermore, 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. A distance based method, specifically the leverage approach[44], was employed in our QSAR modeling. So, we computed the leverage values for every compound and these were plotted against the standard residuals. From this plot (the so-called William's plot), the applicability domain

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

Set	Statistics	Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Active compounds	Mean	μ_1^{Mol}	0.411	5.269	0.212	0.250	−1.253
		$\mu_1\mu_{15}^{Hyd}$	0.133	5.347	−0.748	1.259	−0.518
		$\mu_1\mu_1^{Hyd}$	−0.070	4.404	−0.868	1.477	−0.211
	Standard deviation	μ_1^{Mol}	0.456	0	0.285	0.481	0.235
		$\mu_1\mu_{15}^{Hyd}$	0.306	0	0.320	0.571	0.234
		$\mu_1\mu_1^{Hyd}$	0.350	0	0.132	0.703	0.470
	Variance	μ_1^{Mol}	0.208	0	0.081	0.231	0.055
		$\mu_1\mu_{15}^{Hyd}$	0.094	0	0.103	0.326	0.055
		$\mu_1\mu_1^{Hyd}$	0.122	0	0.018	0.494	0.221
Inactive compounds	Mean	μ_1^{Mol}	−0.239	0.610	−0.582	5.388	
		$\mu_1\mu_{15}^{Hyd}$	0.807	−0.575	−0.142	5.684	
		$\mu_1\mu_1^{Hyd}$	1.377	−0.554	−0.336	4.263	
	Standard deviation	μ_1^{Mol}	0.461	0.441	0.320	0	
		$\mu_1\mu_{15}^{Hyd}$	0.369	0.487	0.255	0	
		$\mu_1\mu_1^{Hyd}$	0.675	0.181	0.294	0	
	Variance	μ_1^{Mol}	0.213	0.194	0.103	0	
		$\mu_1\mu_{15}^{Hyd}$	0.136	0.237	0.065	0	
		$\mu_1\mu_1^{Hyd}$	0.456	0.033	0.086	0	

was established inside a squared area within ±2 standard deviations and a leverage threshold h^* ($h^* = 3p'/n$, with p' the number of model parameters and n the number of training compounds).

2.5. Model evaluation

Various 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[7] such as the Wilk's lambda (λ), the Mahalanobis distance (D^2), the Fisher's ratio (F), and the corresponding p-level (p), the percentage of good classification and the proportion between the cases and adjustable parameters ($\rho > 4$) in the equation, as well as the Matthew's coefficient (C).[45] The Wilk's λ statistic implies perfect discrimination for $\lambda = 0$ and the absence of discrimination when $\lambda = 1$. The Mahalanobis distance indicates the separation of the respective groups, showing whether the model possesses an suitable discriminatory power for differentiating between the two respective groups. Goodness-of-prediction of the discriminant models has been assessed by means of external validation. Notice that validation of the models with compounds not used in the model setup is a crucial but necessary step to ensure generalization, and also of major relevance for future QSAR studies.

2.6. Bond contributions

One of the greatest advantages of the TOPS-MODE approach, over other traditional QSAR methods, stems from its sub-structural nature. This means that one can transform the QSAR model into a bond additive scheme and thus describe the endpoint activity as a sum of bonds contributions related to different structural fragments of the molecules.[46]

Bond contributions are based on the local spectral moments, which in turn are defined as the diagonal entries of the different powers of the weighted bond matrix:

$$\mu_k^T(i) = b_{ii}(T)^k \tag{2}$$

where $\mu_k^T(i)$ is the kth local moment of the bond i and $b_{ii}(T)$ are the diagonal entries of the weighted bond matrix, and T is the type of the bond weight (*Hyd*, *Dip*, and *MR*).

For a given molecule, one can substitute the values of the local spectral moments computed by Eq. 2 into Eq. 3 below and, thus, gather the total contribution to the anticancer activity of its different bonds:

$$P = a_o + \sum_k a_k \cdot \mu_k^T \tag{3}$$

These contributions represent the additive features of the property modeled and they can be expressed as fragment contributions, the sum of the contributions of different bonds that are inside the substructure whose contribution is under examination.

3. Results and discussion

The final partition of the data derived in a structurally representative distribution of chemicals into training and predicting series. A training set of 58 active and 42 inactive compounds was obtained, whereas a test set of 12 active and 11 inactive compounds (see Tables 1 and 2 in Supplementary data). It is worth noting that there are two clusters that contain only one member each one. These two cases are the only dimmers in the whole data set, being **109** active and included in the prediction set while **110** inactive and included in the training set. At first sight, both could be considered as potential outliers; however their inclusion in this QSAR study might be important due to the structural information that they can provide related to the large difference between their activities values—compound **109** is about 73-fold more active than **110** (see Fig. 2).

The best classification model derived from the training set, by combining the LDA and FS techniques along with the TOPS-MODE representation, followed by Randić's orthogonalization, is given below together with the statistical parameters of the LDA.

$$P = 0.64^{\cdot 1}\Omega\mu_1^{MR} - 1.83^{\cdot 2}\Omega\mu_0\mu_6^{Dip} + 0.39^{\cdot 3}\Omega\mu_1\mu_{15}^{Hyd} - 0.79^{\cdot 4}\Omega\mu_1\mu_1^{Hyd} + 0.91 \tag{4}$$

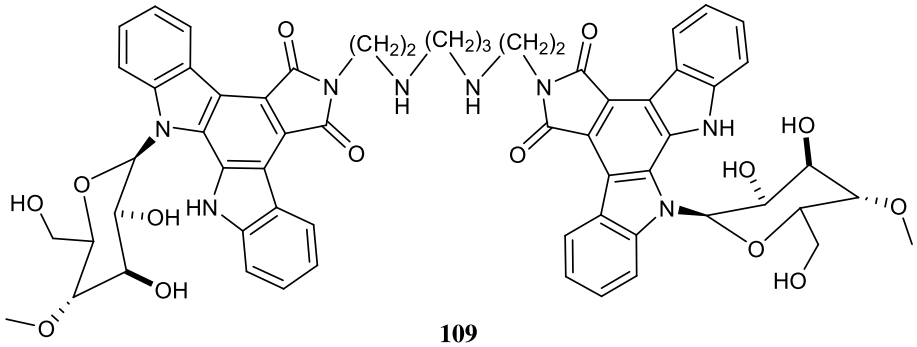
$$\boldsymbol{N} = 100 \;\; \lambda = 0.60 \;\; \boldsymbol{D^2} = 2.65 \;\; \boldsymbol{F}(4,95) = 15.625 \;\; \boldsymbol{p} < 0.0001 \;\; \boldsymbol{C} = 0.637 \;\; \rho = 8.4$$

The large $\boldsymbol{F}$ index, and small $\boldsymbol{p}$ value are indicative of the model's statistical significance. In addition, the values of the Wilks statistic (λ) and of the Mahalanobis distance ($\boldsymbol{D^2}$) show that the model displays an adequate discriminatory power for differentiating both groups. Notice also that the correlation coefficient is always between −1 and +1, and that this coefficient can often provide a much more balanced evaluation of the prediction than, for instance, the percentage.[47] Furthermore, the high value for the Matthew's coefficient indicates a strong linear relationship between the molecular descriptors and the output of the model.[48] Finally, the high value of $\boldsymbol{\rho} = 8.4$ shows that the model is not over-fitted by an excess of parameters; this parameter is expected to be >4 for linear models.[49]

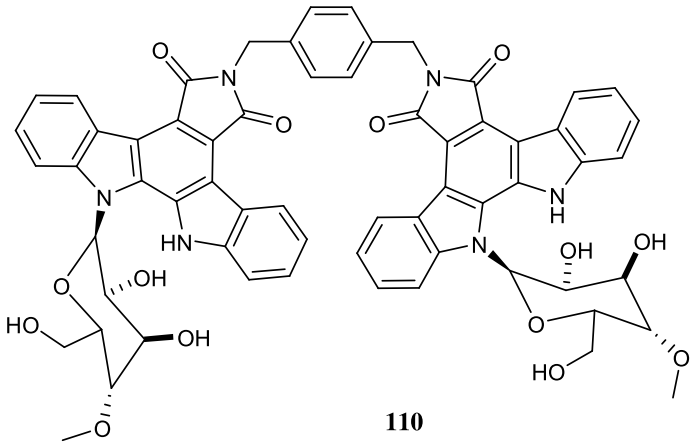
As can be seen in Table 3, this discriminant model showed excellent results in the training and external prediction series used to validate the model. A different, better threshold for the a priori classification probability can be estimated by means of the receiver operating characteristics (ROC) curve.[50] A pronounced ROC curve is depicted in Figure 3 with the area under curve markedly higher than 0.5 – the area under the curve expected for a random classifier (diagonal line).[51]

The four descriptor variables in Eq. 4 are related to the molar refractivity (μ_1^{MR}), the hydrophobicity ($\mu_1\mu_{15}^{Hyd}$ and $\mu_1\mu_1^{Hyd}$), and the bond dipole moment ($\mu_0\mu_6^{Dip}$). It is interesting to compare the role of each descriptor in modeling the compounds' activity: For instance, descriptors μ_1^{MR} and $\mu_1\mu_{15}^{Hyd}$ have a positive contribution to the anticancer activity, whereas descriptors $\mu_1\mu_1^{Hyd}$ and $\mu_0\mu_6^{Dip}$ have a negative contribution. Here, it is important to highlight the opposite contributions of $\mu_1\mu_{15}^{Hyd}$ and $\mu_1\mu_1^{Hyd}$ despite the fact that they are weighted by the same property. The main difference between these variables is the order of the spectral moments involved in the interactions and their significance in the model equation, being the former factor more significant than the latter. This finding might be related to the ability of $\mu_1\mu_{15}^{Hyd}$ to differentiate between similar chemical structures in the data more effectively than $\mu_1\mu_1^{Hyd}$, a difference that is due to the separation between the orders in the interactions.

Finally, an analysis of the applicability domain of the model was carried out, being the results shown in Figure 4. As can be seen, three compounds in the training set are out of the domain of applicability of the model as are another three compounds from the test set (that is, compounds **86**, **90**, **95**, **96**, **109** and **110**; see Figs. 1 and 5), and this is a result of their leverage values. Nevertheless, none of these compounds were considered as outliers because the values of their standardized residuals are not greater than two standard deviation units. It is important to highlight that the compounds out of the domain from the training set bear structural resemblance to those belonging to the prediction set and also have similar leverage values.



109



110

Figure 2. Structure of compounds **109** (test set) and **110** (training set).

Table 3
Results of the classification of compounds in the training and prediction series

	%	Active	Inactive	Unclassified
Training group (82.11% total)				
Active	82.46	47	10	1
Inactive	81.58	7	31	4
Prediction group (81.82% total)				
Active	83.33	10	2	0
Inactive	80.00	2	8	1

A deeper analysis showed that compounds **110** and **109** have the highest leverage values (0.816 and 0.440, respectively), which is not an unexpected result since they formed two single clusters (as explained before), and are the only two dimmers in the whole data set. However, it is noteworthy that these compounds are not to be considered as outliers in contrast to what the cluster analysis suggested.

On the other hand, compounds **90** and **86** have leverage values ($h_{90} = 0.154$ and $h_{86} = 0.158$) very close to the warning leverage ($h^* = 0.15$). The substructures of compound **90** are included in numerous molecules in the training set (see Table 1 in Supplementary data, cases **56** to **92**); therefore, the lowest leverage values can be expected for these compounds in comparison with the others, which are out of the applicability domain of the model. Conversely, compound **86** is the only one in the data set that has a glucopyranosyl tetra-*O*-acetyl substituent in its structure, and that influences its molecular descriptor values, especially μ_1^{MR}, the value of which increases, thus illustrating the bulkiness, steric and hydrophobic effects of this substituent on the activity, and leads to a slightly higher leverage value. Finally, compounds **95** and **96** ($h_{95} = 0.170$ and $h_{96} = 0.238$) are very similar structurally to the rest of the isogranulatimide analogues in the data set (see Table 1 of Supplementary data cases, cases **93** to **103**) apart from the OBn substituent. Nonetheless, this is a substructure that is not unique in the data set; there are other molecules—such as **64** and **71**—that bear this substituent even at the same position (e.g., case **71**).

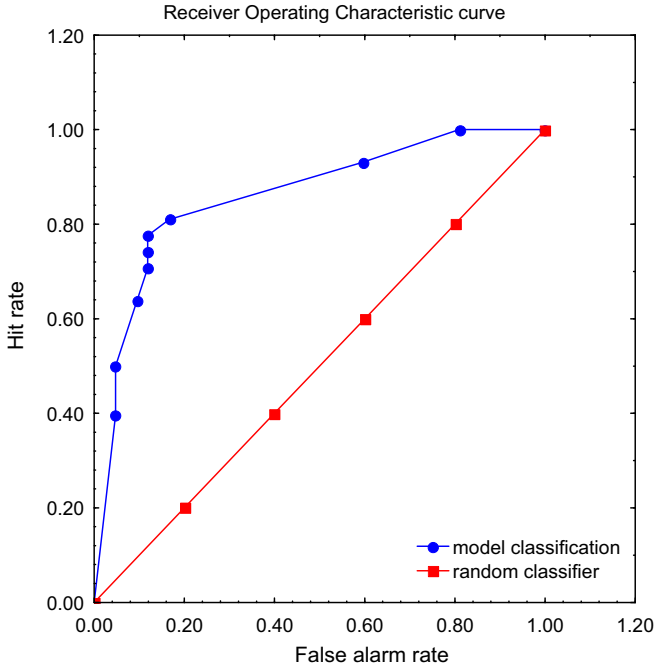


Figure 3. Receiver operating charecteristic (ROC) curve for the classification model (Eq. 4).

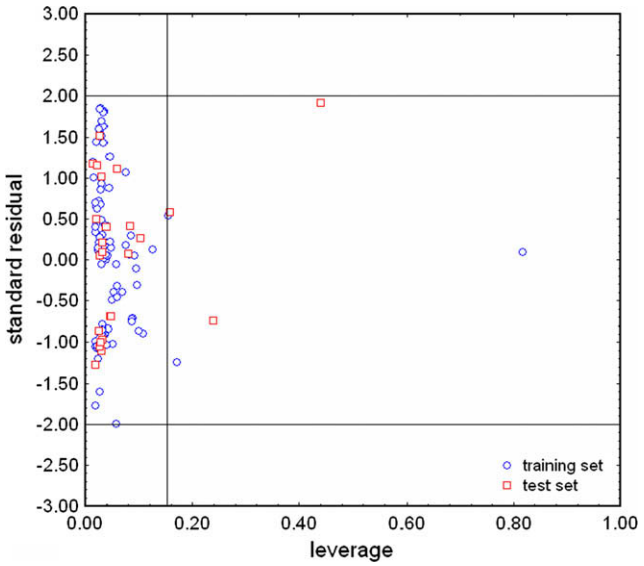


Figure 4. Williams plot based on Eq. 4, that is, a plot of the standardized residuals versus leverage values, with a warning leverage of 0.15.

Consequently, a new model was developed by removing the three compounds in the training set that were found to be out of the applicability domain. This step was carried out to check the effect of such a change on the statistical parameters of the model. As seen in Table 4, there are not significant variations in the model parameters. It follows that the influence of these compounds is not critical for the model and their exclusion is not well supported, just in agreement with the tentative conclusion mentioned above concerning the possible importance of the structural information provided by molecules **109** and **110** for the successful development of the model.

In order to gain an insight into the structure–anticancer activity relationship and to obtain a better interpretation of the model, an analysis of the contribution of the fragments to the classification was carried out. Firstly, the work was developed by taking the diversity of the data as a starting point and three groups were well established according to the core structure of the compounds and the substituents. The first group included rebeccamycin analogues, the second group bis imide granulatimide analogues and the third one isogranulatimide analogues (see Fig. 6). In some compounds in the three groups an indole unit is replaced by a 7-azaindole moiety.

Most of the rebeccamycin analogues are active and, for this reason, we will focus mainly on seeking information about the contributions of different substituents in an attempt to establish which contribute to a larger extent in making a particular compound active. It has been reported that the antitumor activity of rebeccamycin is linked to its capacity to inhibit topoisomerase I by forming a ternary DNA topoisomerase I–rebeccamycin complex that prevents the relegation of the cleaved DNA strand.[52] Additionally, it has be reported that the sugar residue attached to the indolocarbazole chromophore is critical for the drug's ability to interfere with topoisomerase I as well as for the formation of intercalation complexes, with the indolocarbazole chromophore oriented parallel to the

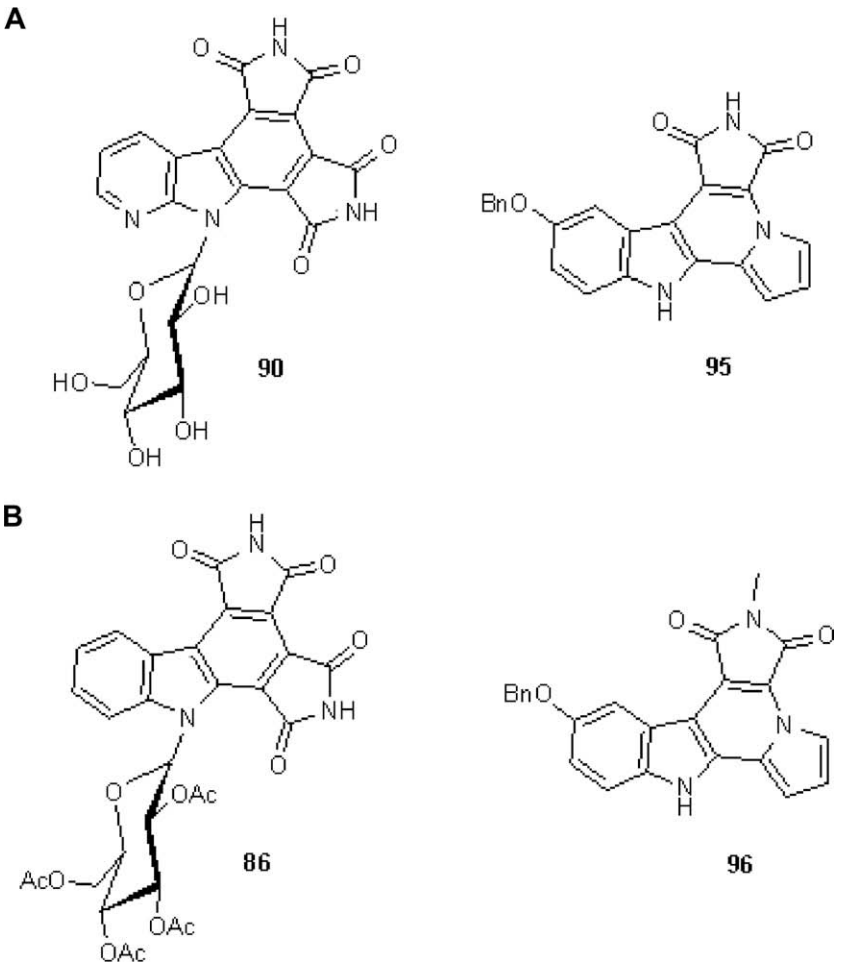


Figure 5. Other compounds out of the applicability domain due to a leverage value higher than the warning one from the training (A) and test (B) sets.

Table 4
Parameters of the non-orthogonal model and the variations after removing the potential influential compounds (cases **90**, **95** and **110**)

Model	a_0	μ_1^{MR}	$\mu_0\mu_6^{Dip}$	$\mu_1\mu_{15}^{Hyd}$	$\mu_1\mu_1^{Hyd}$	% Actives	% Inactives	Total %
Former model	0.39	4.92	−5.39	2.35	−1.79	81.03	83.33	82.00
New model	0.40	5.25	−5.72	2.03	−1.51	81.03	82.05	81.44

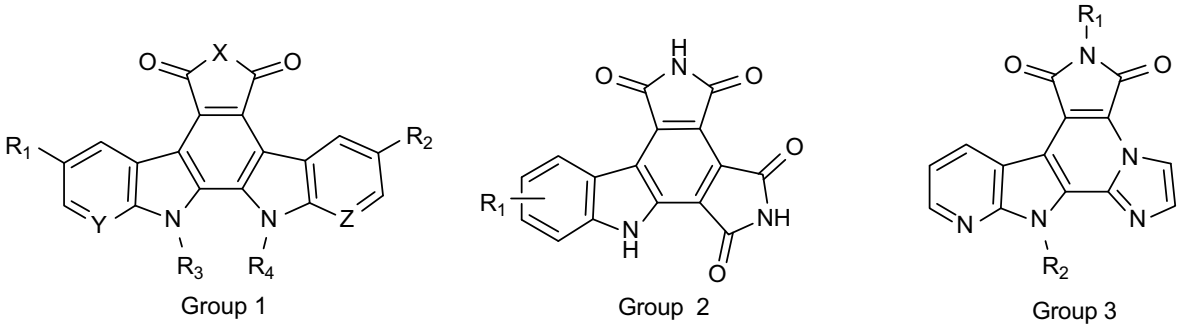


Figure 6. Core of the main structures in the data set.

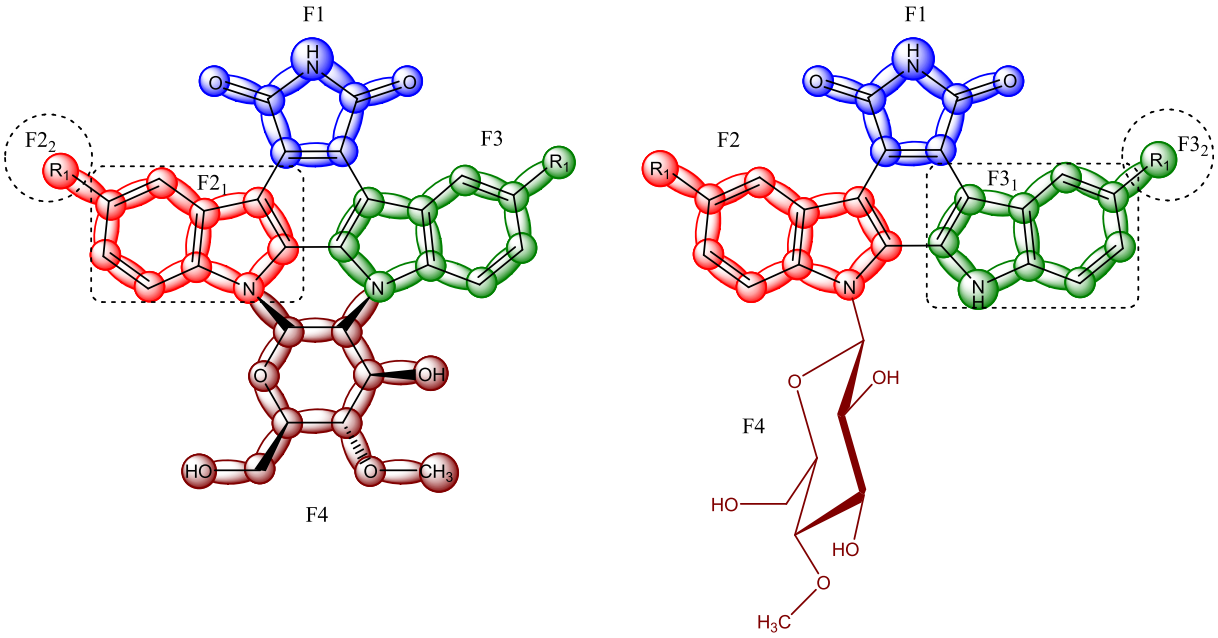


Figure 7. Main fragments studied for those rebeccamycin analogues bearing a carbohydrate unit attached to either only one indole nitrogen or both.

Table 5
Contributions of the most important fragments[a] in some rebeccamycin analogues

Case	R_1	F1	$F2_1$	$F2_2$	F2	$F3_1$	$F3_2$	F3	F4	P^b
Compounds with the sugar unit linked to both indole frameworks										
18	Br	**−0.081**	**1.424**	**0.992**	**2.416**	**1.430**	**0.992**	**2.422**	**−0.294**	**5.116**
14	CHO	−0.664	2.059	0.346	2.405	2.056	0.346	2.402	−1.373	3.316
15	OH	−0.598	2.027	0.231	2.258	1.841	0.231	2.072	−1.332	3.142
16	NO_2	**−0.621**	**1.903**	**−0.007**	**1.896**	**1.900**	**−0.007**	**1.893**	**−1.203**	**2.534**
Compounds with the sugar unit linked to one indole framework										
25	CHO	**−0.795**	**1.915**	**0.333**	**2.248**	**1.751**	**0.332**	**2.084**	**−1.710**	**2.228**
1	H	−0.504	1.827	—	1.827	1.681	—	1.681	−1.227	2.206
28	OH	−0.729	1.877	0.230	2.107	1.722	0.230	1.952	−1.686	2.065
29	COOMe	**−0.946**	**1.601**	**−0.267**	**1.334**	**1.420**	**−0.267**	**1.153**	**−1.773**	**0.118**

[a] Fragments F1, F2, F3 and F4 are those shown in Figure 6. Fragments 2 and 3 were split into two substructures, representing subscripts 1 and 2 the indole and substituent moiety in the two fragments, respectively.
[b] Anticancer activity.

base pair plane of DNA and the carbohydrate interacting with the grooves.[31] According to these studies, and taking into account the fact that most of the rebeccamycin analogues in the data set have a sugar residue attached to the indole unit, we thus focused on two kinds of analogues—those whose carbohydrate was linked to (1) only one indole nitrogen or to (2) both nitrogens.

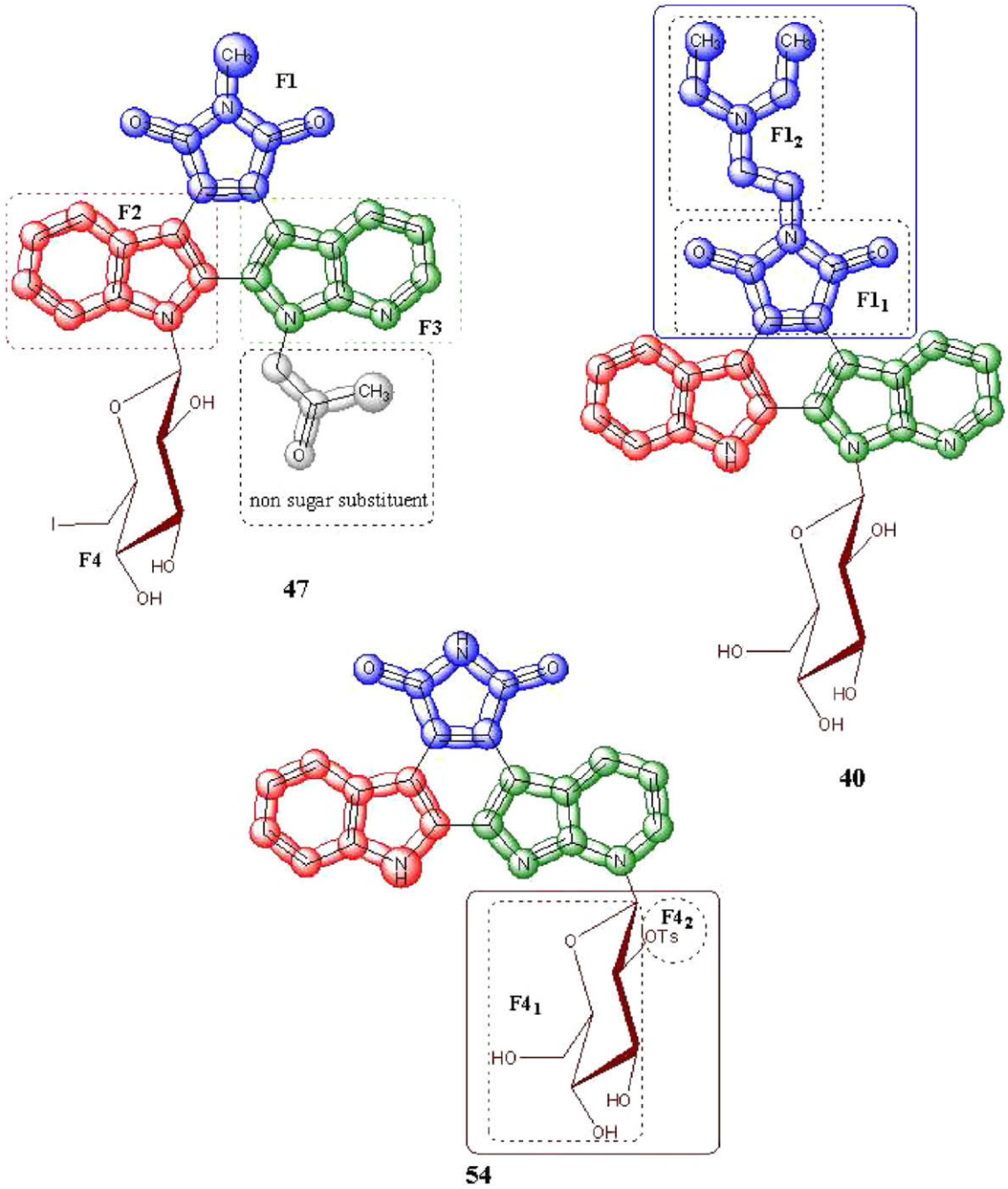


Figure 8. Study of the modifications on indole, carbohydrate and imide fragments.

Table 6
Contributions of the most important fragments[a] in rebeccamycin analogues **47**, **40** and **54**

Case	F1$_1$	F1$_2$	F1	F2	F3	Non-sugar substituent	F4$_1$	F4$_2$	F4	P^b
47	−0.540	0.210	**−0.330**	1.794	1.693	−0.269	−1.193	**2.309**	**1.115**	4.577
40	−0.463	0.084	**−0.379**	1.443	1.481	–	−0.519	–	**−0.519**	2.786
54	−0.945	–	**−0.945**	0.986	0.718	–	−1.504	−2.072	**−1.856**	**−3.460**

[a] Fragments F1, F2, F3 and F4 are those shown in Figure 7. Fragments F1 and F4 were split into two substructures, representing subscript 1 the imide substructure and sugar, while subscript 2 the substituent moiety in both fragments.
[b] Anticancer activity.

On the other hand, the functional modifications on the aromatic moieties have a significant influence on the antiproliferative activity of the compounds. We therefore studied the contribution of some substituents at the 3– and 9–positions for the two kinds of rebeccamycin analogues described above (see Fig. 7). The values of the fragments contribution were calculated as described in Section 2 and the threshold total value to classify the compounds was −0.544, in other words, if the model yielded a value higher than −0.544 the compound was considered active and otherwise it was classified as inactive.

As can be seen from the results in Table 5, there is a general decrease in the contribution of the fragments of the compounds with the carbohydrate part linked to one indole in comparison to the other analogues, an effect that is particularly marked by the influence of F3 lacking a sugar residue and F4. In addition, there is a pattern in the influence of the substituents on the aromatic rings in both kinds of rebeccamycin analogues. The higher the polarity the lower the classification value of the molecule. In this sense the positive effect of the Br substituent ($F2_2 = F3_2 = 0.992$) on the contribution values of F1 and F4 is remarkable, as exemplified by case **18**, the molecule with the highest probability to be classified as active (=0.996, see Supplementary data) and the model equation value (=5.116). Other interesting changes in the activity can be observed in the compounds with the sugar unit linked to one indole framework, where a comparison of cases **25**, **1** and **28** with case **29** is made. Note the low contribution value of the COOMe group ($F2_2 = F3_2 = -0.267$) and how this has a negative influence on the remaining fragment values.

However, there are other structural modifications that worth further study and these include the replacement of an indole unit by a 7-azaindole moiety, possible substituents on the N-imide and the position of the carbohydrate framework as well as the nature of substituents on it. To this end, analysis of these were carried out for compounds **40**, **47** and **54**.

The fragments under investigation were essentially the same as before, though with other splitting (see Fig. 8 for details), and its contribution values are included in Table 6.

The results in Table 6 show the favorable effect of a substituent in the N-imide (compare the F1 contribution values of compounds **47** and **40** with that of **54**) with a lower dipole moment. However, note that the length of the chain is very important, as is the planarity of the core of the molecule and the substituent as a whole. Indeed, this latter factor seems to be critical for the formation of intercalation DNA topoisomerase I–rebeccamycin analogue complexes, with the indolocarbazole chromophore oriented parallel to the base pair plane of DNA.[31]

Analysis of the carbohydrate unit showed the importance of the position of the rebeccamycin analogue to which it is linked and the nature of substituents on this unit. Firstly, the contribution values of the substituents on the sugar part have the following order: I (2.309) > OH (considered as zero as it is included in $F4_1$) ≫ OTs [*O*-Tosyl] (−1.856) according to the previous analyses. Moreover, the different positions of these substituents on the sugar unit show how their proximity to the rebeccamycin main core is not favorable for being classified as active.

For the sugar link position we focused on the comparison between the contributions of F2 and F3. Compound **47**, which has the sugar moiety linked to N-indole, shows the highest F2 and F3 values. In addition, the presence of a non-sugar substituent on fragment 3 of this compound decreases the activity to −0.269. It follows that this is the most favorable position to be linked by the carbohydrate. A second study compared compounds **40** and **54**, which have the sugar on the 7-azaindole. This showed that the most negative effects on the activity arise through the attachment of the sugar to the N-pyridine (see the significant difference between F3 for these compounds: =0.763). In fact, the differences in the contribution values of the frameworks between these compounds provide useful information for a rational design of new compounds as compound **47** is the second compound classified as active, **40** has a moderate value and **54** is one of the few inactive analogues.

As mentioned previously, the analysis of a second group (see Fig. 6) that includes bis imide granulatimide analogues was carried out. The fragments under investigation are shown in Figure 9, and its contribution values are included in Table 7.

An initial analysis showed that the replacement of an indole or 7-azaindole unit by another imide framework is not favorable for activity. In fact, it can be seen from Tables 5–7 how the contributions of the two first fragments mentioned are higher than those calculated for imides.

On the other hand, the influence of the substituent on the aromatic ring is very important, as it is in the other analogues analyzed before. In this case, it appears that bulky functional groups might increase the probability of a given compound being considered as active. The results in Table 7 indicate how substituents such as CH_2=CH–Ph and OBn give rise to the highest contribution values and also highlights the significant changes in activity for smaller groups. Indeed, for the methyl group the fragment contributions decrease to a large extent and can even make the molecule inactive. The slight increase in the activity contribution value of compound **63** provided by the methyl effect, in comparison to compounds **72** and **69** (with hydroxyl), suggests that hydrophobicity is a key parameter to take into account along with bulkiness.

Finally, an analysis of the third group was carried out and this concerns isogranulatimide analogues. Most of these compounds were classified inactive, with **118** being the only active example. However, the output value of the model for this molecule was not so high ($P = -0.324$) and is very close to the threshold value

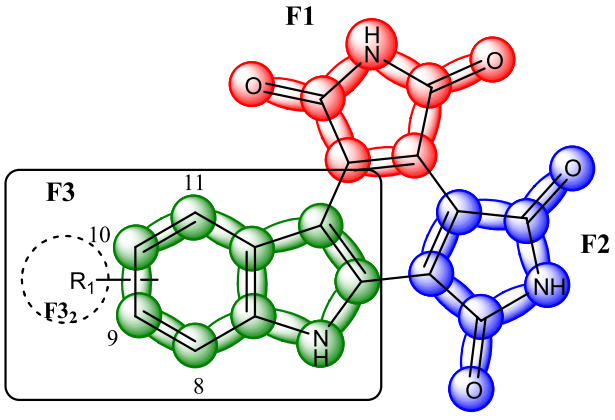


Figure 9. Main fragments studied for the bis imide granulatimide analogues.

Table 7
Contributions of the most important fragments[a] in some bis imide granulatimide analogues

Case	R_1	F1	F2	$F3_1$	$F3_2$	F3	P^b
56	10-CH_2=CH–Ph	**−0.100**	**−0.125**	**0.437**	**0.646**	**1.082**	**1.067**
64	8-OBn	**−0.242**	**−0.262**	**0.629**	**0.367**	**0.996**	**0.716**
63	8-Me	−0.707	−0.716	0.493	0.008	0.501	−0.797
72	10-OH	−0.836	−0.840	0.535	0.115	0.650	−0.891
69	9-OH	−0.836	−0.840	0.533	0.115	0.648	−0.892

[a] Fragments F1, F2 and F3 are those shown in Figure 8. Fragment F3 was split into $F3_1$, equivalent to the indole framework, and $F3_2$, in which R1 is attached to the aromatic ring in any position from 8 to 11.
[b] Anticancer activity.

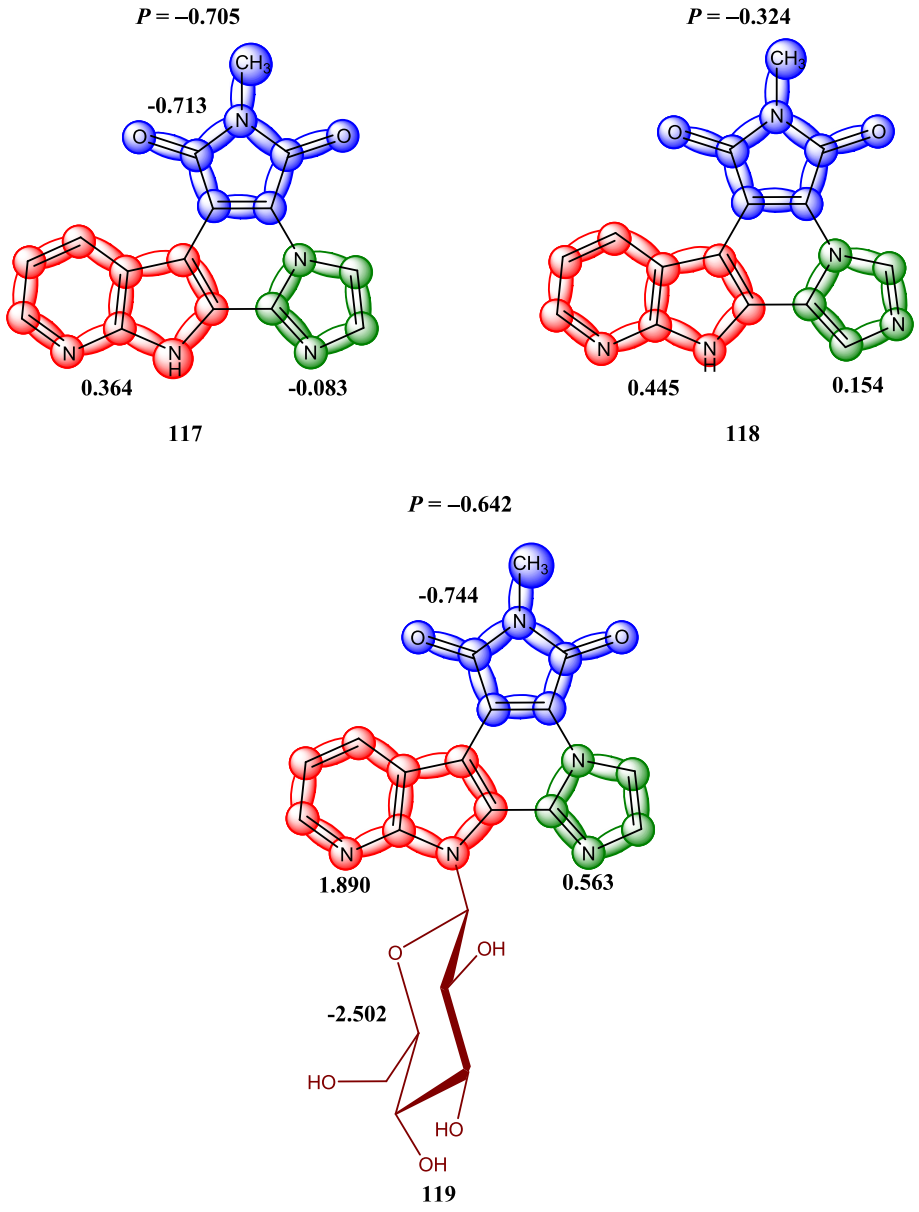


Figure 10. Contributions of the most important fragments in some isogranulatimide analogues.

(−0.544). This finding is very important as it could be considered as an unclassified compound and its prediction could be doubtful. As a consequence, this group was unattractive in terms of developing a structure-activity relationship for optimization. The results are shown in Figure 10, in which the fragments under study are highlighted in various colours and their contribution value is also shown. In addition, the label for the molecule is given (e.g., **117**) along with the global activity value (e.g., ***P* = −0.705**). Briefly, the figure illustrates that the position of the nitrogens in the imidazole ring with respect to rest of the molecule is the main factor that affects the activity, with close proximity between the N-indole and N-imidazole unfavorable for activity. On the other hand, a sugar unit linked to the 7-azaindole fragment only slightly increases the activity, and the molecule remains inactive.

4. Conclusions

A QSAR model was developed using the TOPS MODE approach towards the rational selection of anticancer compounds, from a diverse data set of indolocarbazole derivatives, for antiproliferative activity against murine leukemia tumor cell line (L1210). The model proved to have suitable statistical quality, as demonstrated by the fitting and validation parameters.

The TOPS-MODE approach afforded useful information about the contribution of different fragments in the molecules, as well as enabling a better understanding and interpretation of the developed model. Therefore, this is a particularly interesting tool that can be recommended for future studies in this field.

As for the analysis of the bond contributions, we concluded that the rebeccamycin analogues have the highest probability to be classified as active, particularly when the most significant fragments are attached to the core, that is, (1) a carbohydrate unit with one substituent at the 6′-position, (2) substituents at the 3- and 9-positions on the aromatic ring and (3) substituents on the N-imide fragment whose steric and hydrophobic properties are balanced.

Acknowledgments

This work was carried out with the support of research grants from the Higher Educational Ministry of Cuba and financial support

from Fundação para a Ciência e Tecnologia (FCT, Lisboa; **SFRH/BDP/24512/2005**). We also thank the owners of the MODESLAB for free donation of this software to our investigation group.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2008.11.084.

References and notes

1. *Anticancer drug development guide: preclinical screening, clinical trials and approval*; Teicher, B. A., Andrews, P. A., Eds., second ed.; Humana Press: Totowa, New Jersey, 2004.
2. Theisen, C. *J. Natl. Cancer Inst.* **2003**, *95*, 937.
3. Kaplow, R. *Nurs. Clin. North. Am.* **2005**, *40*, 77.
4. Cabrera, M. A.; González, I.; Fernández, C.; Navarro, C.; Bermejo, M. *J. Pharm. Sci.* **2006**, *95*, 589.
5. González, M. P.; Terán, C.; Teijeira, M.; Morales, A. H. *Curr. Med. Chem.* **2006**, *13*, 2253.
6. Van Waterbeemd, H. In *Chemometric Methods in Molecular Design*; Van Waterbeemd, H., Ed.; New York: Wiley-VCH, **1995**; Vol. 2, p 265.
7. Willett, P. *Perspect. Drug Discov. Des.* **1997**, *7–8*, 1.
8. Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. *J. Med. Chem.* **2000**, *43*, 1975.
9. Garg, R.; Denny, W. A.; Hansch, C. *Bioorg. Med. Chem.* **2000**, *8*, 1835.
10. Assefa, H.; Kamath, S.; Buolamwini, J. K. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 475.
11. González-Díaz, H.; Gia, O.; Uriarte, E.; Hernández, 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.* **2003**, *9*, 395.
12. Bhongade, B. A.; Gadad, A. K. *Bioorg. Med. Chem.* **2004**, *12*, 2797.
13. Morales, A. H.; Cabrera Peréz, M. A.; González, M. P.; Ruiz, R. M.; González-Díaz, H. *Bioorg. Med. Chem.* **2005**, *13*, 2477.
14. Amić, D.; Davidović-Amić, D.; Beslo, D.; Rastija, V.; Lučić, B.; Trinajstić, N. *Curr. Med. Chem.* **2007**, *14*, 827.
15. Saíz-Urra, L.; González, M. P.; Teijeira, M. *Bioorg. Med. Chem.* **2007**, *15*, 3565.
16. Helguera, A. M.; Rodriguez-Borges, J. E.; Garcia-Mera, X.; Fernández, F.; Cordeiro, M. N. D. S. *J. Med. Chem.* **2007**, *50*, 1537.
17. Pindur, U.; Kim, Y. S.; Mehrabani, F. *Curr. Med. Chem.* **1999**, *6*, 29.
18. Prudhomme, M. *Curr. Med. Chem.* **2000**, *7*, 1189.
19. Nettleton, D. E.; Doyle, T. W.; Krishnan, B.; Matsumoto, G. K.; Clardy, J. *Tetrahedron Lett.* **1985**, *26*, 4011.
20. Bush, J. A.; Long, B. H.; Catino, J. J.; Bradner, W. T.; Tomita, K. *J. Antibiot. (Tokyo)* **1987**, *40*, 668.
21. Voldoire, A.; Sancelme, M.; Prudhomme, M.; Colson, P.; Houssier, C.; Bailly, C.; Leonce, S.; Lambel, S. *Bioorg. Med. Chem.* **2001**, *9*, 357.
22. Moreau, P.; Sancelme, M.; Bailly, C.; Leonce, S.; Pierre, A.; Hickman, J.; Pfeiffer, B.; Prudhomme, M. *Eur. J. Med. Chem.* **2001**, *36*, 887.
23. Marminon, C.; Facompre, M.; Bailly, C.; Hickman, J.; Pierre, A.; Pfeiffer, B.; Renard, P.; Prudhomme, M. *Eur. J. Med. Chem.* **2002**, *37*, 435.
24. Marminon, C.; Anizon, F.; Moreau, P.; Leonce, S.; Pierre, A.; Pfeiffer, B.; Renard, P.; Prudhomme, M. *J. Med. Chem.* **2002**, *45*, 1330.
25. Marminon, C.; Pierre, A.; Pfeiffer, B.; Perez, V.; Leonce, S.; Renard, P.; Prudhomme, M. *Bioorg. Med. Chem.* **2003**, *11*, 679.
26. Moreau, P.; Gaillard, N.; Marminon, C.; Anizon, F.; Dias, N.; Baldeyrou, B.; Bailly, C.; Pierre, A.; Hickman, J.; Pfeiffer, B.; Renard, P.; Prudhomme, M. *Bioorg. Med. Chem.* **2003**, *11*, 4871.
27. Marminon, C.; Pierre, A.; Pfeiffer, B.; Perez, V.; Leonce, S.; Joubert, A.; Bailly, C.; Renard, P.; Hickman, J.; Prudhomme, M. *J. Med. Chem.* **2003**, *46*, 609.
28. Messaoudi, S.; Anizon, F.; Leonce, S.; Pierre, A.; Pfeiffer, B.; Prudhomme, M. *Eur. J. Med. Chem.* **2005**, *40*, 961.
29. Henon, H.; Anizon, F.; Golsteyn, R. M.; Leonce, S.; Hofmann, R.; Pfeiffer, B.; Prudhomme, M. *Bioorg. Med. Chem.* **2006**, *14*, 3825.
30. Messaoudi, S.; Anizon, F.; Peixoto, P.; David-Cordonnier, M. H.; Golsteyn, R. M.; Leonce, S.; Pfeiffer, B.; Prudhomme, M. *Bioorg. Med. Chem.* **2006**, *14*, 7551.
31. Conchon, E.; Aboab, B.; Golsteyn, R. M.; Cruzalegui, F.; Edmonds, T.; Leonce, S.; Pfeiffer, B.; Prudhomme, M. *Eur. J. Med. Chem.* **2006**, *41*, 1470.
32. Hugon, B.; Anizon, F.; Bailly, C.; Golsteyn, R. M.; Pierre, A.; Leonce, S.; Hickman, J.; Pfeiffer, B.; Prudhomme, M. *Bioorg. Med. Chem.* **2007**, *15*, 5965.
33. Conchon, E.; Anizon, F.; Aboab, B.; Golsteyn, R. M.; Leonce, S.; Pfeiffer, B.; Prudhomme, M. *Eur. J. Med. Chem.* **2007**.
34. Henon, H.; Messaoudi, S.; Anizon, F.; Aboab, B.; Kucharczyk, N.; Leonce, S.; Golsteyn, R. M.; Pfeiffer, B.; Prudhomme, M. *Eur. J. Pharmacol.* **2007**, *554*, 106.
35. Leonce, S.; Perez, V.; Casabianca-Pignede, M. R.; Anstett, M.; Bisagni, E.; Pierre, A.; Atassi, G. *Invest. New Drugs* **1996**, *14*, 169.
36. Dillon, W. R.; Goldstein, M. *Multivariate Analysis: Methods and Applications*; Wiley: New York, 1984.
37. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 844.
38. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 320.
39. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 23.
40. Gutierrez, Y.; Estrada, E. MODESLAB 1.0 (Molecular DEScriptors LABoratory) for Windows, Universidad de Santiago de Compostela, Spain, 2002.
41. Statsoft, I., STATISTICA (data analysis software system), version 6.0; 2002.
42. Kowalski, R. B.; Wold, S. *Handbook of Statistics*; North Holland Publishing: Amsterdam, 1982.
43. Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
44. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361.
45. Matthews, B. W. *Biochim. Biophys. Acta* **1975**, *405*, 442.
46. Estrada, E.; Patlewicz, G.; Gutierrez, Y. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 688.
47. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. *Bioinform. Rev.* **2000**, *16*, 412.
48. Yuan, Z. *FEBS Lett.* **1999**, *451*, 23.
49. García-Domenech, R.; de Julián-Ortiz, J. V. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445.
50. Provost, F.; Fawcett, T. In Analysis and Visualization of Classifier Performance Comparison Under Class and Cost Distributions, Third International Conference on Knowledge Discovery and Data Mining (KDD); American Association for Artificial Intelligence Press: 1997.
51. Toivonen, H.; Srinivasan, A.; King, R. D.; Kramer, S.; Helma, C. *Bioinformatics* **2003**, *19*, 1183.
52. Bailly, C.; Colson, P.; Houssier, C.; Rodrigues-Pereira, E.; Prudhomme, M.; Waring, M. *J. Mol. Pharmacol.* **1998**, *53*, 77.