Probing the Anticancer Activity of Nucleoside Analogues: A QSAR Model Approach Using an Internally Consistent Training Set

Aliuska Morales Helguera,[†,§,||] J. E. Rodríguez-Borges,[‡] Xerardo García-Mera,[⊥,*] Franco Fernández,[⊥] and
M. Natália D. S. Cordeiro[†,*]

REQUIMTE and CIQ, Department of Chemistry, University of Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal, CBQ and Department of Chemistry, Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba, and Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15706 Santiago de Compostela, Spain

Received December 19, 2006

The cancer research community has begun to address the *in silico* modeling approaches, such as quantitative structure−activity relationships (QSAR), as an important alternative tool for screening potential anticancer drugs. With the compilation of a large dataset of nucleosides synthesized in our laboratories, or elsewhere, and tested in a single cytotoxic assay under the same experimental conditions, we recognized a unique opportunity to attempt to build predictive QSAR models. Here, we report a systematic evaluation of classification models to probe anticancer activity, based on linear discriminant analysis along with 2D-molecular descriptors. This strategy afforded a final QSAR model with very good overall accuracy and predictability on external data. Finally, we search for similarities between the natural nucleosides, present in RNA/DNA, and the active nucleosides well-predicted by the model. The structural information then gathered and the QSAR model *per se* shall aid in the future design of novel potent anticancer nucleosides.

Introduction

The nucleoside analogues (NAs) were among the first chemotherapeutic agents to be introduced for the medical treatment of cancer.[1−4] These chemicals include several analogues of physiological pyrimidine or purine nucleosides and nucleobases, which are thought to interact with various intracellular targets involved in the metabolism of physiological nucleosides and DNA synthesis. Their cytotoxic activity against a panel of tumors *in vitro* has been extensively studied in the past, in particular against murine leukemia L1210/0 cells.[5,6] For instance, it has been shown that NAs such as the 5-substituted-2′-deoxyuridines suppress tumor cell proliferation by inhibition of thymidylate synthase,[6] an enzyme essential in the synthesis of DNA. The clinical use of NAs, however, is limited by important side-effects and primary or acquired drug resistance,[4] thus providing opportunities for the development of new, more efficient analogues of this sort.

Finding new drugs is a complex, expensive, and very time-consuming task, as there is no single systematic way to automatically discover a drug even when the disease, targets, and molecular mechanism(s) of drug activity are well understood.[3,4,7] There are literally millions of candidate molecules and experiments that cannot be carried out on every one due to prohibitive costs both in terms of time and money. Lately, the rational drug design strategies, especially the *in silico*-based approaches, have emerged as a promising alternative or complementary tool toward the effective screening of potential drugs. *In silico* approaches include for example the quantitative structure−activity relationship (QSAR) modeling techniques, which are increasingly attracting the attention of medicinal chemists as well as of the pharmaceutical industry.[8−19] QSAR modeling may be better regarded as an exercise to filter drug candidates, before they are subjected to more intensive calculations such as docking or an experimental measurement of activity (*in vitro*) and under real conditions (*in vivo*) last.

Almost all QSAR techniques rely upon the use of molecular descriptors, which aim at encoding useful pshysicochemical information to enable correlation of molecular structure with biological activity. There may be thousands of molecular descriptors nowadays with potential for being applied in drug design,[20,21] particularly in anticancer activity predictions. Among others, the topological indices (TIs) have become widely applicable due to their great success in many diverse QSAR studies including cancer related research.[8,9,11,12] TI descriptors are based on graph-theoretical concepts and derived mainly from 2D-molecular connectivity but also, to a certain extent, from atom types and electronic environment. They account for a large number of molecular properties, mostly of the steric attributes of the molecule such as molecular size, branching, and to a certain degree, shape.[20,21]

For years our research group has been engaged in the design, synthesis, and evaluation of nucleoside analogues in an effort to develop potential anticancer or antiviral agents.[22−32] We recognized a unique opportunity when we could assemble a set of over 200 NA compounds, previously synthesized in one our laboratories,[25,26,33−37] or elsewhere,[8,38−58] and measured in a single, consistent cytotoxic assay. With the desire to build a reliable predictive QSAR model from such data that could be used to probe anticancer activity, we examined the use of 2D descriptors along with linear discriminant analysis and feature selection algorithms. Our final QSAR model exhibits very good cross-validation statistics and perform well on an external validation set comprising other NA chemicals designed by us,[30−37] many of which with unknown activity (that are being reported here for first time). Finally, we analyzed the key structural features present in active NAs by means of a similarity study.

[*] To whom correspondence should be addressed. E-mail: ncordeir@fc.up.pt; fax: +351 226082959 (M.N.D.S.C.). E-mail: qoxgmera@usc.es; fax : +34 981594912 (X.G.M.).
[†] REQUIMTE, University of Porto.
[‡] CIQ, Department of Chemistry, University of Porto.
[§] CBQ, Central University of Las Villas.
[||] Department of Chemistry, Central University of Las Villas.
[⊥] University of Santiago de Compostela.

10.1021/jm061445m CCC: $37.00 © 2007 American Chemical Society
Published on Web 03/07/2007

Table 1. The Seven 2D-Molecular Descriptors Used in the Initial Classification Model (model 1; eq 1)

symbol	definition
D/Dr06	distance/detour ring index of order 6
AAC	mean information index on atomic composition
MATS8e	Moran autocorrelation − lag 8/weighted by atomic Sanderson electronegativities
X5A	average connectivity index chi-5
piPC10	molecular multiple path count of order 10
MPC10	molecular path count of order 10
piPC09	molecular multiple path count of order 9

Table 2. Intercorrelation among the Seven Descriptors Selected as Statistically Significant by LDA[a]

	AAC	D/Dr06	piPC10	X5A	MATS8e	piPC09	MPC10
AAC	1.00	−0.27	-0.37	0.18	−0.24	−0.34	-0.43
D/Dr06		1.00	**0.57**	0.09	0.26	0.65	**0.74**
piPC10			1.00	−0.22	0.21	**0.90**	**0.84**
X5A				1.00	−0.12	−0.24	−0.34
MATS8e					1.00	0.23	0.35
piPC09						1.00	**0.91**
MPC10							1.00

[a] Significant correlations are marked in bold.

Results and Discussion

Model Calibration. The best classification model derived from the training set (from now on denoted as model 1), by combining the LDA and FS techniques along with a 2D topological structure representation, is given below together with the statistical parameters of the LDA, while the selected descriptors are shown in Table 1.

$$P = 0.027\mathrm{D/Dr06} + 11.338\mathrm{AAC} + 4.104\mathrm{MATS8e} - 191.215\mathrm{X5A} + 0.011\mathrm{piPC10} - 0.030\mathrm{MPC10} - 0.013\mathrm{piPC09} - 6.480 \quad (1)$$

$$N = 241 \quad \rho = 34.40 \quad F(7{,}233) = 25.65 \quad p < 10^{-5} \quad \lambda = 0.56 \quad D^2 = 3.25$$

The ρ value, large sample size, large F index, and small p value are indicative of the model's statistical significance. In addition, the values of the Wilks λ statistic (λ can take values from zero, perfect discrimination, to one, no discrimination) and of the Mahalanobis distance (a measure of the separation between the active and inactives groups) show that the model displays an adequate discriminatory power for differentiating both groups. The latter is also confirmed by the classification results; the model correctly classified 82.4% of the 91 cytotoxic compounds and 84.7% of the 150 inactives, giving rise to an overall 83.8% effective discrimination of the 241 training set compounds.

Further analysis of this classification model should only be carried out after checking the reliability of preadopted assumptions. First, LDA establishes a linear, additive relation between the molecular descriptors and the underlying bioactivity, and, in fact, this is the simplest mathematical form that might be envisaged for the model in absence of any *a priori* information. Nevertheless, by looking at the distribution of the standardized residuals (observed minus predicted divided by the square root of the residual mean square) for all cases (Figure 1), no specific pattern is seen, thereby reinforcing the idea that the model does not exhibit a nonlinear dependence.

Another aspect deserving special attention is the degree of collinearity among the variables of the model, but that may easily be diagnosed by analyzing the cross-correlation matrix. As seen in Table 2, the pairs of descriptors (D/Dr06; piPC10), (D/Dr06; MPC10), (piPC10; piPC09), (piPC10; MPC10), and (piPC09; MPC10) are strongly correlated with each other. Rather than deleting any of such descriptors, it is of interest to examine the performance of orthogonal complements in modeling the anticancer activity.

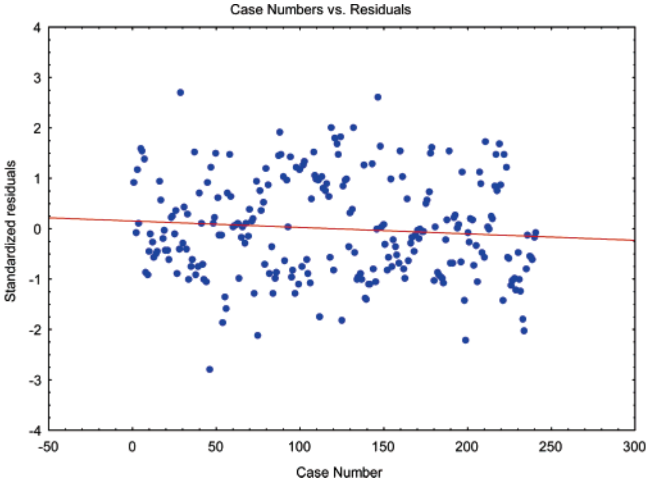


Figure 1. Distribution of the standardized residuals for all cases studied.

Following the Randić technique, we determined orthogonal complements for all variables in model 1, which in turn were further standardized, to then find the best five-variable equation (model 2):

$$P = 1.828\cdot{}^{2}\Omega\mathrm{D/Dr06} + 3.646\cdot{}^{3}\Omega\mathrm{piPC10} + 0.455\cdot{}^{5}\Omega\mathrm{MATS8e} - 3.708 \times {}^{6}\Omega\mathrm{piPC09} - 0.549 \times {}^{7}\Omega\mathrm{MPC10} + 0.161 \quad (2)$$

$$N = 241 \quad \rho = 48.2 \quad F(7{,}233) = 35.74 \quad p < 10^{-5} \quad \lambda = 0.57 \quad D^2 = 3.21$$

where the symbol ${}^{i}\Omega X$ means the orthogonal complement of variable X, the superscript referring to the variable in the equation employed to obtain the residuals.

As can be noticed, the descriptors ACC and X5A have been excluded, as they were found to be not statistically significant. That had however no effects on the overall fitness of the model as the statistics are as robust as before (see eq 1), though the classifications of the active/moderate-active slightly improved (84.62%; see eq 1 and also the Supporting Information), which therefore increased the percentage of overall discrimination (86.65%). Yet there are significant differences between both models as regards the interpretation of the results. By comparing eq 1 with eq 2, one can see that there are no changes in the sign of the coefficients save for the one associated to the constant. The relative contributions of the variables in the orthogonal-descriptor model are nevertheless significantly different to those in the nonorthogonalized model. For example, the variables MATS8e and MPC10 have similar contributions (in absolute terms) in model 2, while in model 1 the contribution of MATS8e is 9 times larger than that of MPC10. Therefore, for purposes of QSAR interpretability, we shall consider the orthogonal-descriptor model defined in eq 2.

Let us now check another important parametric assumption of LDA, i.e., multivariate normality. Traditionally, one starts by plotting the individual frequency distributions of each variable. Figure 2 shows the plots for the histograms of frequency distributions of the different variables in model 2, divided by active and nonactive groups, respectively. Attached

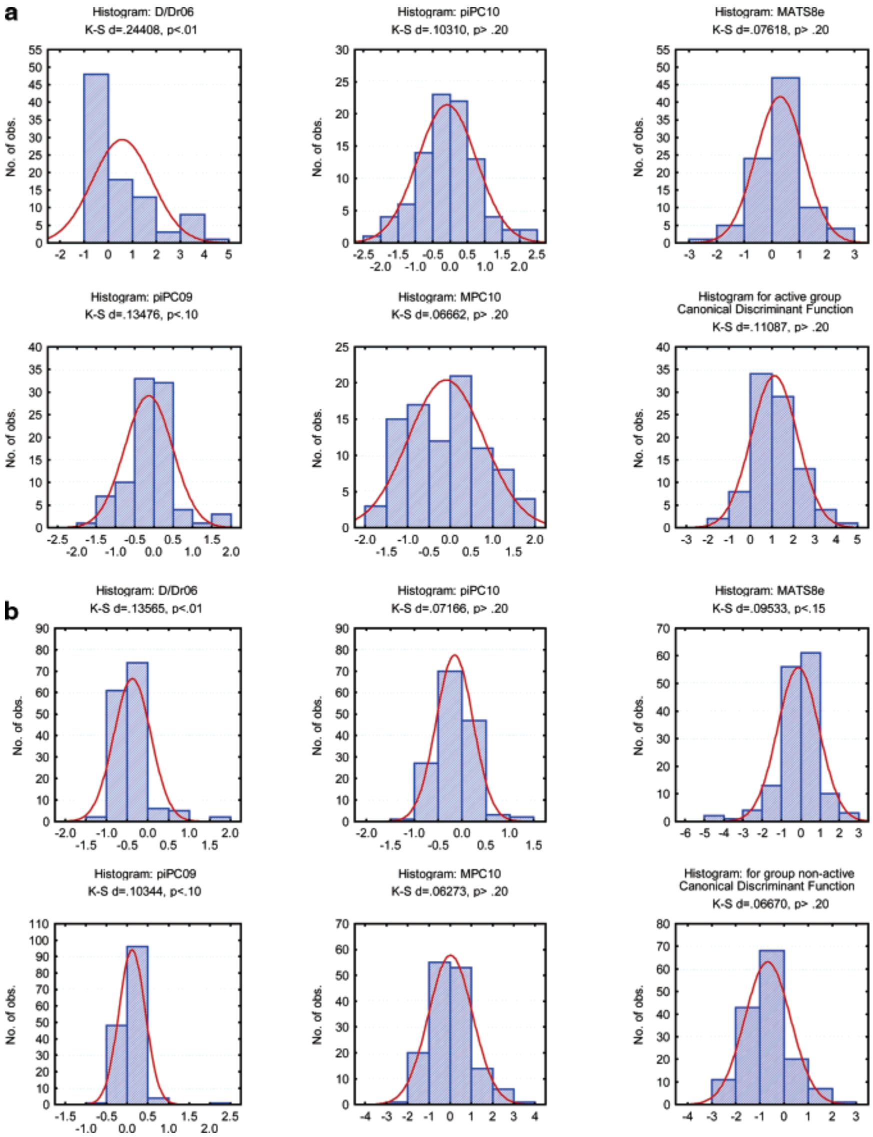


Figure 2. Histograms for the frequency distributions of all variables in model 2 (eq 2), considering active/moderate-active groups **(a)** and inactive groups **(b)**.

to each plot are also the results derived from the Kolmogorov–Smirnov statistical test (*d*). At first sight, these plots and the *d* values suggest that not all variables exhibit adequate normal distribution, say at least the D/Dr06 and piPC09 descriptors. However, lack of individual normality is not by itself enough for rejecting the hypothesis of multivariate normality. In fact, for checking multivariate normality, one should instead examine the discriminant function, which takes into account the interactions among variables. Accordingly, a visual inspection of the normality plots and frequency distributions for the discriminant function (see Figure 2), as well as the calculated *d* values for both groups (actives: $d = 0.067$; inactives: $d = 0.111$; $p > 0.200$ in both cases), lead us to accept that hypothesis.

Moving on now to the hypothesis of homocedasticity, a possible problem regarding the homogeneity of the (co)variances is suggested by the Box's M statistical test ($p < 0.01$), although this test can be overly sensitive to large data files[59] which is likely what happened here. This nevertheless increases the likelihood that a case belongs to the higher dispersion group, and, in this sense, adjusting the *a priori* probabilities can greatly improve the overall classification rate of the discriminant model.

A different, better threshold for the *a priori* classification probability can be estimated by means of the receiver operating characteristics (ROC) curve.[60] This is a useful technique not only for obtaining the best thresholds but also for organizing classifiers.[61,62] As Figure 3 shows, the optimal threshold for predicting the active chemicals with the present QSAR model is 0.52. Further, one can see that the model is not a random, but a truly statistically significant, classifier, since the area under the ROC curve is significantly higher than the area under the random classifier curve (diagonal line).

Model Interpretation and Validation. With regard to QSAR modeling, our main goal was not only to establish a robust explanatory model but also to ensure good generalization performance. This was first accomplished by means of internal cross-validation (CV) of the model. The statistics and classification results reported in Table 3 correspond to five independent leave-20%-out CV runs, each involving a different, randomly chosen partition into training and a test set.

As can be seen, the model is robust and shows little dependence on the composition of the training and test sets. It also shows good predictive power, judging from the averages

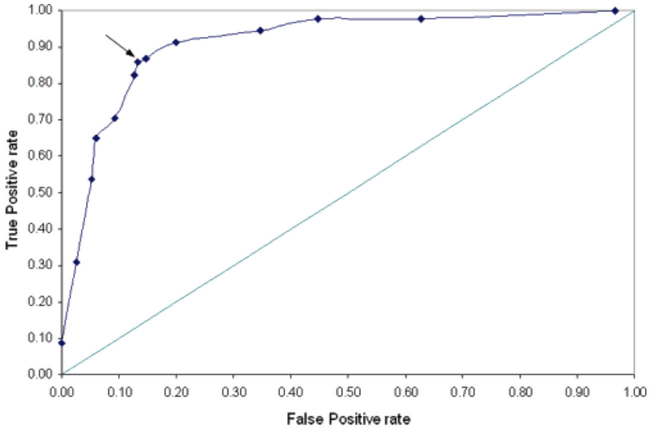


Figure 3. Receiver operating characteristic (ROC) curve for the classification model (model 1; eq 1).

Table 3. Results from the Cross-Validation Leave-group-out Procedure[a]

CV-run	λ	D^2	F	%AC$_G$ (T)[b]	%AC$_G$ (P)[b]
1	0.529	3.750	33.360	85.49	77.08
2	0.528	3.760	33.430	89.12	79.17
3	0.551	3.430	30.459	85.49	83.33
4	0.607	2.715	24.133	83.42	81.25
5	0.602	2.786	24.547	83.85	91.84
average	0.563	3.288	29.186	85.48	82.53

[a] Results obtained with model 2 (eq 2) after removing ~20% of compounds from the original data set (48 out of 241). [b] %AC$_G$ (T) and %AC$_G$ (P) are the percentage of good overall classification in the training and predicting sets, respectively.

computed for the overall classification results in the training and predictive sets generated in each CV run (85.48% and 82.53%, respectively).

For a final validation, an external test set (20 compounds) was assembled from other NAs synthesized previously in one of our laboratories. Some of these compounds have already been evaluated experimentally in terms of cytotoxic activity,[23,24] but most have not. For this whole external set, the percentage of overall discrimination is 75% (15 out of 20), and, simultaneously, the model correctly classifies 80% (4 out of 5) of actives and 73.33% (11 out of 15) of inactives (Table 4), which is quite impressive given the diversity of such set and the complexity of the biologic response being modeled. It also reveals the good predictive ability of the present model.

Finally, the model 2 interpretation was made on the basis of structural differences between noncytostatic and cytostatic carbocyclic nucleosides included in training and external test set (well-predicted by model 2), respectively, and based on the analysis of model 2. For this propose, we consider that three of the most important variables are related with molecular path counts (MPC10, piPC10, and piPC09). The path counts are molecular descriptors obtained from an H-depleted molecular graph and are vertex invariants encoding that molecular environment, defined as the number of path length m starting from the ith vertex to any other vertex in the graph. A path (or self-avoiding walk) is a walk without any repeated vertices.[20] The path length is the number of edges associated with the path, and this value is increased with the ring size, ring numbers, and the ramification number.[63]

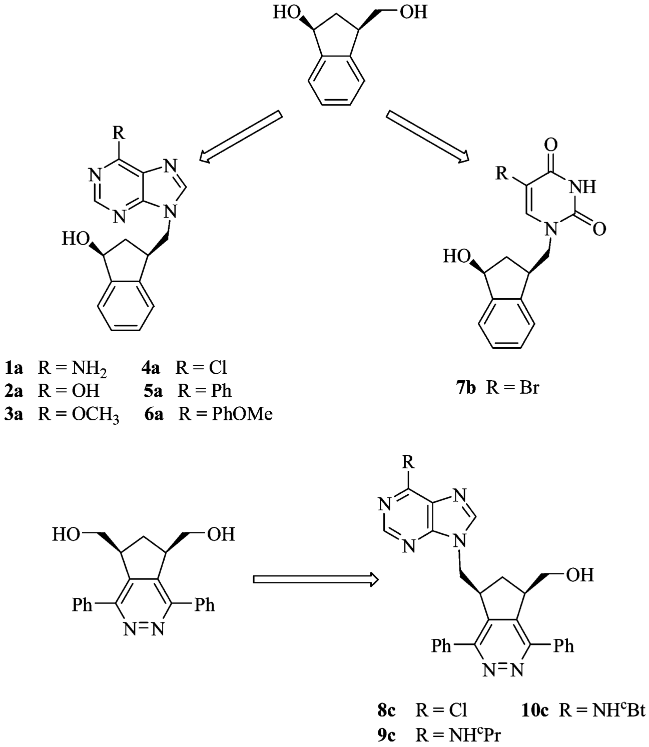
With the interpretation and application of model 2, we can design an active structure from an inactive one. For this purpose, Scheme 3 shows the structural representation of chemical 195 (inactive) and chemical 11d (active). As can be seen, both NAs have the same purine base, but their carbocycles differ. For improving biological activity, we increased the number of molecular path counts on the order of 10, through increasing of length of the $N-$glycoside bond, size of carbocyclic ring (from cyclobutyl to cyclopentyl), and its branchings. Thus we achieved a significant increase of piPC10 for chemical 11d with

Table 4. The 20 Carbanucleosides Used in the External Prediction Set along with the Observed Cytotoxicity against the Cellular Line L1210/0 and Classification (*a posteriori* Probabilities) According to Model 2 (eq 2)

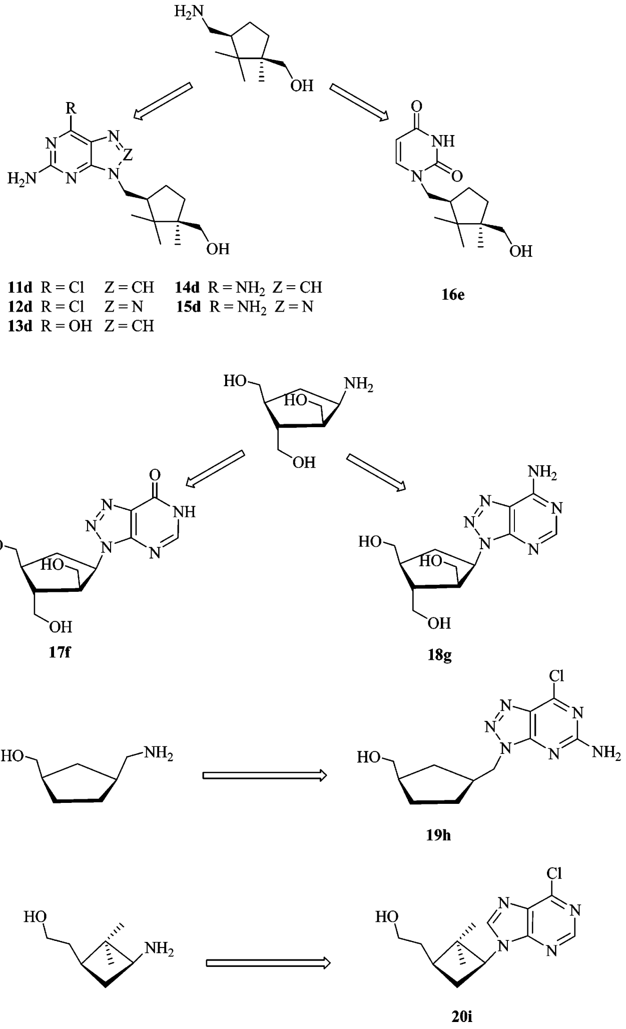
no.	R	Z	IC$_{50}$ (μM)[a]	prob (%)[b]	predicted class[c]
		Active Chemicals			
10c	NH-cyclobutyl	-	90.80	100	+1
11d	-	CH	50.03	100	+1
12d	-	N	44.64	100	+1
19h	-	-	169.66	5.29	−1
20i	-	-	105.07	97.68	+1
		Inactive Chemicals			
1a	NH$_2$	-	>200	10.89	−1
2a	OH	-	>200	15.21	−1
3a	OCH$_3$	-	>200	14.64	−1
4a	Cl	-	>200	24.76	−1
5a	Ph	-	>200	78.79	+1
6a	PhOCH$_3$	-	>200	98.95	+1
7b	-	-	>200	27.61	−1
8c	Cl	-	>200	100	+1
9c	NH-cyclopropyl	-	>200	100	+1
13d	OH	CH	>200	2.7	−1
14d	NH$_2$	CH	>200	2.43	−1
15d	NH$_2$	N	>200	3.45	−1
16e	-	-	>200	8.56	−1
17f	-	-	>200	17.34	−1
18g	-	-	>200	15.13	−1

[a] 50% inhibitory concentration or compound concentration required to reduce proliferation of tumors cells by 50%. [b] A posteriori probability of classifying a chemical as active, according to model 2. [c] Values of +1 and −1 stand for compounds with and without cytotoxic activity, respectively.

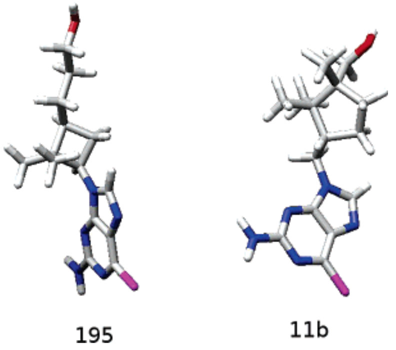
Scheme 1. General Procedures for the Preparation of the Test Set Nucleoside Analogues Belonging to Series **a, b**, and **c**



Scheme 2. General Procedures for the Preparation of the Test Set Nucleoside Analogues Belonging to Series **d, e**, **f**, **g**, **h**, and **i**

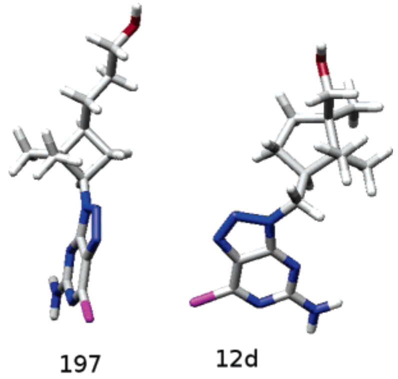


Scheme 3. Molecular Representation of Chemicals 195 (Non-cytotoxic) and 11b (Cytotoxic)

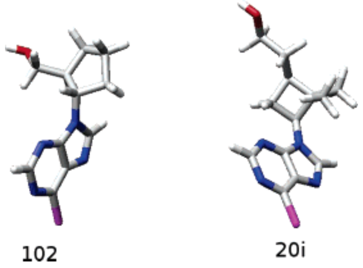


respect to chemical 195; in contrast, piPC09 does not show an experimentally significant increment, probably because these factors exert major influence over a path count of higher order. It is important to remark that piPC10 has a contrary effect over biological activity if it is compared with piPC09. Thus, we can explain the increment in cytostatic activity of chemical 11d, based on our model 2. A similar behavior could be analyzed for the chemicals 197 (no active) and 12d (active) (Scheme 4), as well as for chemicals 102 (no active) and 20i (active) (Scheme 5). In the last case, we achieve an improvement of cytostatic activity by increasing of branching attached to carbocycle ring, in spite of the fact that the ring size decreased.

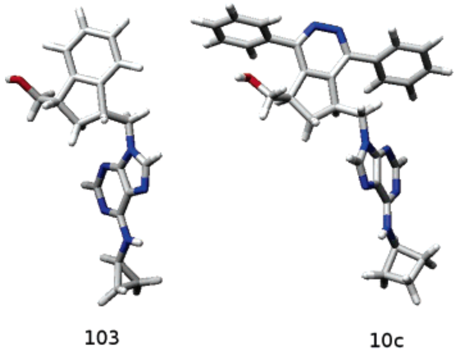
Scheme 4. Molecular Representation of Chemicals 197 (Non-cytotoxic) and 12d (Cytotoxic)



Scheme 5. Molecular Representation of Chemicals 102 (Non-cytotoxic) and 20i (Cytotoxic)



Scheme 6. Molecular Representation of Chemicals 103 (Non-cytotoxic) and 10c (Cytotoxic)



Another important variable in model 2 is D/Dr06, which has positive contribution to the anticancer property and means distance/detour ring index of order 6. D/Dr06 is based the vertex row sums of D/DD (D, distance matrix and DD, detour matrix) matrix, allowing one to build local structural invariants of cyclic systems corresponding to individual atoms or individual fragments of a molecule.[64] The ring descriptor is obtained by summing local contributions of carbon atoms making up the benzene ring. The indices derived from benzene rings are descriptors that reflect the local geometrical rather than local electronic environment of the benzene rings. However, in our present considerations no electronic factors have been involved; thus, we end with descriptors that reflect solely the geometrical features of these systems. This variable increased with number of condensed and noncondensed cycle systems.[64] On this basis, we can explain the differences between chemical 103 (inactive) and 10c (active) (Scheme 6a). Both chemicals have two condensed rings as carbocyclic, but in the case of chemical 10c

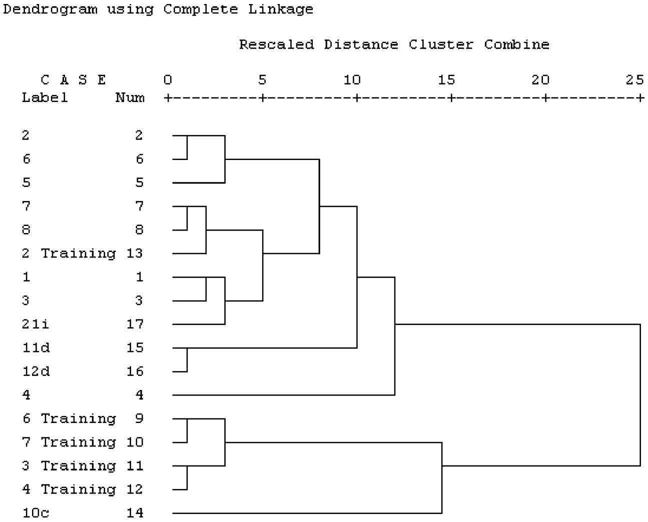


Figure 4. Dendogram and similarities between the active carbanucleosides well-classified (model 2; eq 2) and natural nucleosides after a hierarchical cluster analysis.

the addition of two benzene rings increased the D/Dr06 value and then the cytostatic activity. In addition, for the chemical 10c, the values of piPC10 and piPC09 descriptors increase in similar order; the first one increases the biological activity while the second one decreases it. Therefore, piPC10 and piPC09 do not have a global contribution to cytostatic activity in this structure.

Similarity/Dissimilarity Searching. As a final exercise, we have searched for similarities between the natural nucleosides present in ribonucleic acid (RNA) or deoxyribonucleic acid (DNA) and the active carbocyclic nucleosides well-predicted by model 2 (eq 2). The eight natural nucleosides used in this similarity/dissimilarity searching are (**1**) deoxythymidine, (**2**) deoxycytidine, (**3**) deoxyadenosine, (**4**) deoxyguanosine, (**5**) uridine, (**6**) cytidine, (**7**) adenosine, (**8**) guanosine. Figure 4 shows the dendogram resulting from the hierarchical cluster analysis, while Table 5 lists the calculated Euclidean distances.

In general, one can say that the present carbocyclic nucleosides are closely related to adenosine and guanosine (see Figure 4 and Table 5), in particular the indan derivatives (chemicals **6**, **7**, and **4**). In fact, the indan derivatives studied here are substituted at position 2 and 3 of the cyclopentane ring, and a similar topological pattern is observed in adenosine and guanosine in which these positions are substituted by hydroxyl groups.

On the other hand, chemicals **2** and **21i** from the training and validation set, respectively, display great similarity to natural nucleosides as seen in Table 5. Chemical **2**, which is a cyclohexane derivative, is more similar to adenosine and guanosine, while chemical **21i**, a cyclobutane derivative, is more closely related to deoxyadenosine.

Conclusions

Here we have examined the ability of a large, diverse, and consistently tested training set to provide predictive QSAR models for probing anticancer activity. The training set included 241 nucleoside analogues, derived from purinic and pyrimidinic bases, and was assembled from literature compounds with published cytotoxic activity against L1210/0 cancer cells, whereas another set of 20 NAs, which have been previously designed by us and synthesized, acted as an external validation set. The cytotoxic activitiy of 14 compounds from such external set is being reported here for the first time. This is of great relevance to the cancer research community, as new compounds enrich the structural diversity of related databases and could be used in forthcoming QSARs.

With regard to the QSAR modeling, the combination of LDA in conjunction with a 2D topological structure representation and feature selection algorithms was found to produce a final classification model with good accuracy, internal cross-validation statistics, and predictability on the external data. This was followed by a clustering search analysis to identify the similarities to natural nucleosides of the well-predicted active carbocyclic nucleosides (from both the training and test sets), and structural interpretation of the results took into account the mechanism of action, substitution of DNA bases, responsible for cytotoxic activity. The information provided by this analysis showed us that the present carbocyclic nucleosides are, in general, closely related to adenosine and guanosine, say, in particular, the indan derivatives. But similarity is more remarkably strong for chemical **2**, a cyclohexane derivative, while in the case of chemical **21i**, a cyclobutane, is strongly related to deoxythymidine and deoxyadenosine. Overall, that structural information and the QSAR model *per se* shall aid in future design of novel potent anticancer drugs.

Experimental Section

Data Set. All the compounds used here are primarily nucleoside analogues, derived from purinic and pyrimidinic bases, and were experimentally assayed for their inhibitory effects (IC_{50}) in the proliferation of L1210/0 cancer cells. These experiments have been conducted at the Rega Institute for Medical Research of the Katholieke Universiteit Leuven in Leuven, Belgium, following the same *in vitro* assay protocol.[53] One can rely on the quality of such biological data, which has been measured by a single protocol, at the same laboratory, by even the very same staff.

Table 5. Euclidean Distances between the Natural Nucleosides and the Active Carbocyclic Nucleosides, Derived with the Final 2D Orthogonal Descriptors[a]

	bases[b]							
	DNA				RNA			
chemical	1	2	3	4	5	6	7	8
6_training	3.363	2.980	3.268	3.919	2.992	2.946	**2.765**	**2.731**
7_training	3.526	3.235	3.403	3.995	3.218	3.200	**2.955**	**2.911**
3_training	3.976	**3.277**	3.939	4.657	3.438	**3.254**	3.334	3.322
4_training	4.001	3.427	3.930	4.596	3.537	3.401	**3.379**	**3.355**
2_training	1.060	1.221	1.209	2.207	0.750	1.165	**0.474**	**0.566**
10c	7.173	7.073	6.925	7.175	7.041	7.039	**6.686**	**6.608**
11d	2.453	2.600	2.514	3.080	2.387	2.574	**2.294**	**2.303**
12d	2.392	2.718	2.419	2.914	2.444	2.689	**2.293**	**2.291**
21i	**0.675**	2.205	**0.584**	1.211	1.551	2.159	1.090	1.050

[a] The five topological indices included in the orthogonal-descriptor model (eq 2). The less Euclidean distance values between the natural nucleosides and the active carbocyclic nucleosides are marked in bold. [b] 1: deoxythymidine, 2: deoxycytidine, 3; deoxyadenosine, 4: deoxyguanosine, 5: uridine, 6: cytidine, 7: adenosine, 8: guanosine.

The compounds were first clustered into two groups according to their IC_{50} values. The first group, actives, includes all chemicals with $IC_{50} < 200\ \mu M$, while the second one, inactives, includes those with $IC_{50} \geq 200\ \mu M$. This classification criterion was adopted not only because over that concentration chemicals can be too toxic and therefore lack biological value but also to get a reasonable ratio of active/inactive chemicals in the dataset. We have also discarded all chemicals with disconnected structures like salts and polymers. In addition, there are other compounds pairs, geometrical isomers, which could not be distinguished by the present 2D descriptors but had nevertheless similar IC_{50} values; in such cases, one of the isomers was discarded. Finally, we managed to assemble a large, balanced dataset of 261 chemicals comprising 96 actives and 165 inactives.

A necessary but delicate task in any QSAR modeling is predictive validation, i.e., to assess model adequacy for new compounds. The most reliable way to predictively validate a model is by external validation,[65−67] which consists of making predictions for an independent set of data not used in the model setup. Here we select a small subset (20 compounds) of the chemicals from the entire dataset to act as an external test set. These compounds have been synthesized in one of our laboratories following usual procedures (see description below),[30−37] and six of them have already been assayed experimentally in L1210 cells and their activity reported[23,24] while the others were evaluated here for the first time. The remaining chemicals (241 compounds) form the training set. A complete list of the training set compounds along with the reported experimental cytotoxicity (IC_{50} values expressed in μM) is given as Supporting Information.

Molecular Descriptors. Our study is based on the different sets of 2D-descriptors available in the DRAGON software (version 2.1),[68] which in turn have a long history in structure−activity and structure−property correlation. They include, for instance, pure topological descriptors, walk and path counts, connectivity indices, information indices, or 2D-autocorrelations. Taking into account the compounds' structural diversity, an initial subset of 462 descriptors was computed for each molecule from the SMILES (simplified molecular input line entry specification) inputting of chemical structures. By disregarding descriptors with constant or near constant values inside each class, a final subset of 259 descriptors was then used for building the QSAR models.

Modeling Technique. Linear discriminant analysis (LDA), specifically the LDA technique implemented in the STATISTICA software (version 6.0),[69] will be used here to find classification models (eq 3), which best describe the cytostatic activity P, as a linear combination of the predictor X-variables (2D descriptors) weighted by the a_n coefficients:

$$P = a_1X_1 + a_2X_2 + ...a_nX_n + a_0 \qquad (3)$$

In developing the models, P values of +1 and −1 were assigned to active and inactive compounds, 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 not classified (NC) by the model.

Feature Selection. The forward stepwise (FS) technique was applied to select the molecular descriptors (X-variables) with the highest influence on the anticancer activity.[70] This technique begins by including the variable which yields the best linear fit in terms of explaining the response. The next variable is included as that variable which most significantly improves the existing model. Once this new model is determined, the variables included are tested to see if the model can be improved by dropping them from the model. If the model can be improved, the variable is removed and the stepwise procedure is repeated until no further variables are either included or removed.

In any multiple linear-based QSAR it is desirable that the variables included in the model are not interrelated to each other. Highly correlated variables clearly contain redundant information that might be more effectively encoded by a single variable. Further, and most importantly from the point of view of a QSAR model, correlated independent variables lead to multicollinearity, which can cause problems in interpreting the individual estimated coefficients.[71−74] One very useful and informative approach of avoiding multicollinearity is the orthogonal descriptors technique suggested by Randić some year's ago.[71−73] In the Randić's approach, after choosing a starting descriptor, subsequent descriptors are added only as their orthogonal complements to the descriptors already present. This approach has the advantages that the equation coefficients are stable (i.e., they do not change as new descriptors are added), and the new information supplied by each additional descriptor is clearly distinguished in the final equation statistics. Here, to tackle the multicollinearity problem, we have applied the Randi'c's approach and orthogonalized the variables following the order selected by the FS scheme. The resulting orthogonal-descriptor model was standardized afterward.

Model Evaluation. Several diagnostic statistical tools were used for evaluating our model equations, in terms of the criteria goodness-of-fit and goodness-of-prediction. Measures of goodness-of-fit have been estimated by standard statistics such as the Wilks' lambda (λ), the Mahalanobis's distance (D^2), the Fisher ratio (F), and the corresponding p-level (p) as well as the percentage of good classifications and the ratio between cases and adjustable parameters (ρ). We have also checked the validity of the preadopted assumptions, parametric (normality, homocedasticity, and noncollinearity) and linearity of the model, which is another important aspect in the application of multiple linear statistical-based approaches[75] such as the LDA technique. Goodness-of-prediction of the final model has been assessed by means of internal cross-validation (CV), specifically by the leave-group-out (LGO) technique.[76] Basically, CV-LGO consists of forming several subsets from the entire dataset, each missing a small group of k cases ($k = 48$, in this case). These k cases are used to validate a new model that is trained with the corresponding subset. Quality (goodness-of-fit) of the new models gives then a measure of the predictive ability of the full model. As mentioned, we also evaluated the predictability of our final discriminant model by using an external set of compounds not used in the model setup. Validation of the final model with compounds, which are not part of the training set, is a crucial but necessary step to ensure generalization, and also of great relevance to future QSAR studies.

Similarity/Dissimilarity Analysis. The task here is to cluster the active NAs well-predicted by our final classification model (either from the training set or from the external set), accordingly to their similarity to natural nucleosides. To this end, we applied a hierarchical agglomerative cluster analysis based on the Euclidean distance, using all variables included in the model.

Synthesis. The test set compounds have been obtained from suitable precursors, such as a diol or an amino alcohol, accordingly to the classic procedures of nucleosic preparation.[1] The chemicals in series **a**, **b,** and **c** were obtained by a coupling reaction like as Mitsunobu reaction or nucleophlic substitution of respective diol, suitably functionalized with puric or pyrimidinic bases (Scheme 1). The remaining chemicals, series **d, e**, **f**, **g**, **h**, and **i**, were obtained from an amino alcohol by transformation of the amine group into the corresponding puric or pyrimidinic base (Scheme 2).

Cytostatic Assay. All assays were carried out in flat-bottomed 96-well microtiter plates. To each well were added 5×10^4 murine leukemia cells (L1210/0) and a given amount of the test compound. The cells were allowed to proliferate for 48 h at 37 °C a humidified, CO_2-controlled atmosphere. The growth of the cells was linear during this 48 h incubation period. At the end of the incubation period, the cells were counted in a coulter counter (Coulter Electronics Ltd, Harpenden Herts, England), and the number of dead cells was evaluated by staining with trypan blue. The IC_{50} (50% inhibitory concentration) was defined as the compound concentration that inhibit cell proliferation by 50%, as compared to untreated control.[53]

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

Supporting Information Available: A complete list of compounds used in training set with their inhibitory activity against the cellular line L1210/0, and their classification (*a posteriori* probabilities) according to the final model (eq 1). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Simons, C. *Carbocyclic Nucleosides*; Gordon & Breach Science Publisher: Reading, UK, 2001; Vol. 3, pp 137−153.

(2) Jordheim, L.; Galmarini, C. M.; Dumontet, C. Nucleoside analogues and nucleobases in cancer treatment. *Lancet Oncol.* **2002**, *3*, 415−424.

(3) Jordheim, L.; Galmarini, C. M.; Dumontet, C. Drug resistance to cytotoxic nucleoside analogues. *Curr. Drug Targets* **2003**, *4* (6), 443−460.

(4) Jordheim, L.; Galmarini, C. M.; Dumontet, C. Recent developments to improve the efficacy of cytotoxic nucleoside analogues. *Anti-Cancer Drug Discovery* **2006**, *1*, 163−170.

(5) Balzarini, J.; De Clercq, E.; Torrence, P. F.; Mertes, M. P.; Park, J. S.; Schmidt, C. L.; Shugar, D.; Barr, P. J.; Jones, A. S.; Verhelst, G.; Walker, R. T. Role of thymidine kinase in the inhibitory activity of 5-substituted-2′-deoxyuridines on the growth of human and murine tumor cell lines. *Biochem. Pharmacol.* **1982**, *31* (6), 1089−95.

(6) Balzarini, J.; de Clercq, E.; Ayusawa, D.; Seno, T. Thymidylate synthetase-positive and -negative murine mammary FM3A carcinoma cells as a useful system for detecting thymidylate synthetase inhibitors. *FEBS Lett* **1984**, *173* (1), 227−32.

(7) Dixit, K. S.; Mitra, S. N. Bioinformatics in drug discovery. *Curr. Res. Inf. Pharm. Sci.* **2002**, *3*, 2−6.

(8) Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. A novel approach for the virtual screening and rational design of anticancer compounds. *J. Med. Chem.* **2000**, *43* (10), 1975−85.

(9) Gálvez, J.; García-Domenech, R.; Gómez-Lechón, M. J.; Castell, J. V. Use of molecular topology in the selection of new cytostatic drugs. *J. Mol. Struct. (THEOCHEM)* **2000**, *504*, (1−3.), 241−248.

(10) Estrada, E.; Uriarte, E. Recent advances on the role of topological indices in drug discovery research. *Curr. Med. Chem.* **2001**, *8*, 1573−1588.

(11) Xiao, Z.; Xiao, Y.-D.; Feng, J.; Golbraikh, A.; Tropsha, A.; Lee, K.-H. Antitumor agents. 213. Modeling of epipodophyllotoxin derivatives using variable selection k nearest neighbor QSAR method. *J. Med. Chem.* **2002**, *45*, 2294−2309.

(12) Gonzalez-Diaz, 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. Markovian chemicals “in silico” design (MARCH-INSIDE), a promising approach for computer-aided molecular design I: discovery of anticancer compounds. *J. Mol. Model. (Online)* **2003**, *9* (6), 395−407.

(13) Ren, S. S.; Lien, E. J. Anticancer agents: tumor cell growth inhibitory activity and binary QSAR analysis. *Curr. Pharm. Des.* **2004**, *10*, 1399−1415.

(14) Saczewski, F.; Reszka, R.; Gdaniec, M.; Grünert, R.; Bednarski, P. J. Synthesis, X-ray crystal structures, stabilities, and in vitro cytotoxic activities of new heteroarylacrylonitriles. *J. Med. Chem.* **2004**, *47*, 3438−3449.

(15) Kozikowski, A. P.; Roth, B.; Tropsha, A. Why academic drug discovery makes sense. *Science* **2006**, *313*, 1235−1236.

(16) Huang, R.; Wallqvist, A.; Covell, D. G. Assessment of in vitro and in vivo activities in the National Cancer Institute’s anticancer screen with respect to chemical structure, target specificity, and mechanism of action. *J. Med. Chem.* **2006**, *23* (49), 1964−1979.

(17) Verma, R. P. Understanding apoptosis in terms of QSAR. *Anticancer Agents Med. Chem.* **2006**, *6* (1), 41−52.

(18) Clare, B. W.; Supuran, C. T. A perspective on quantitative structure-activity relationships and carbonic anhydrase inhibitors. *Expert Opin. Drug. Metab. Toxicol.* **2006**, *2*, 113−137.

(19) Glenn, M. P.; Kahnberg, P.; Boyle, G. M.; Hansford, K. A.; Hans, D.; Martyn, A. C.; Parsons, P. G.; Fairlie, D. P. Antiproliferative and phenotype-transforming antitumor agents derived from cysteine. *J. Med. Chem.* **2004**, *47*, 2984−2994.

(20) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley VCH: Weinheim, Germany, 2000.

(21) Adeel Malik, A.; Singh, H.; Andrabi, M.; Husain, S. H.; Shandar, A. Databases and QSAR for cancer research. *Cancer Inf.* **2006**, *2*, 99−111.

(22) Balo, M. C.; Fernández, F.; Lens, E.; López, C.; De Clercq, E.; Andrei, G.; Snoeck, R.; Balzarini, J. Synthesis and antiviral activities of some novel carbocyclic nucleosides. *Nucleosides Nucleotides* **1996**, *15*, (7), 1335−1346.

(23) Blanco, J. M.; Caamaño, O.; Fernández, F.; Gómez, G.; Nieto, M. I.; Balzarini, J.; Padalko, E.; De Clercq, E. Synthesis and antiviral and cytostatic activities of carbocyclic nucleosides incorporating a modified cyclopentane ring. 1: guanosine analogues. *Nucleosides Nucleotides* **1997**, *16* (1, 2), 159−171.

(24) Nieto, M. I.; Blanco, J. M.; Caamaño, O.; Fernández, F.; García-Mera, X.; Balzarini, J.; Padalko, E.; Neytsb, J.; De Clercq, E. Synthesis, antiviral and cytostatic activities of carbocyclic nucleosides incorporating a modified cyclopentane ring. Part 2: adenosine and uridine analogues. *Nucleosides Nucleotides* **1998**, *17*, (7), 1255−1.

(25) Blanco, J. M.; Caamaño, O.; Fernandez, F.; Garcia-Mera, X.; Hergueta, A. R.; Lopez, C.; Rodriguez-Borges, J. E.; Balzarini, J.; De Clercq, E. Synthesis and antiviral and antineoplastic activities of some novel carbocyclic guanosine analogues with a cyclobutane ring. *Chem. Pharm. Bull. (Tokyo)* **1999**, *47*, (9), 1314−7.

(26) Figueira, M. J.; Blanco, J. M.; Caamaño, O.; Fernández, F.; García-Mera, X.; López, C.; Andrei, G.; Snoeck, R.; Padalko, E.; Neyts, J.; Balzarini, J.; De Clercq, E. Synthesis and antiviral and cytostatic activities of carbocyclic nucleosides incorporating a modified cyclobutane ring. Part 1. Guanosine analogues. *Archiv der Pharmazie* **1999**, *332*, 348−352.

(27) López, C.; Balo, C.; Blanco, J. M.; Fernández, F.; De Clercq, E.; Balzarini, J. A cyclobutane carbonucleoside with marked selectivity against tk+ and tk- varicella zoster virus *Nucleosides Nucleotides* **2001**, *20* (4, 7), 1133−1135.

(28) Figueira, M. J.; Caamaño, O.; Fernández, F.; Blanco, J. M. Synthesis of (6)-c-4-amino-r-1,t-2,c-3-cyclopentanetrimethanol and higher homologues of 8-azapurine arabino-carbocyclic nucleosides. *Tetrahedron* **2002**, *58*, 7233−7240.

(29) Yao, S. W.; Lopes, V. H.; Fernandez, F.; Garcia-Mera, X.; Morales, M.; Rodriguez-Borges, J. E.; Cordeiro, M. N. Synthesis and QSAR study of the anticancer activity of some novel indane carbocyclic nucleosides. *Bioorg. Med. Chem.* **2003**, *11* (23), 4999−5006.

(30) Fernández, F.; García-Mera, X.; Morales, M.; Vilariño, L.; Caamaño, O.; De Clercq, E. Synthesis of new 6-substituted purinyl-5′-nor-1′-homocarbanucleosides based on indanol. *Tetrahedron* **2004**, *60*, 9245−9253.

(31) Caamaño, O.; Gómez, G.; Fernández, F.; García, M. D.; García-Mera, X.; De Clercq, E. A Convenient Synthesis of New Purinyl-homo-carbonucleosides on a Cyclopentane Ring Fused with Pyridazine. *Synthesis* **2004**, *17*, 2855−2862.

(32) Fernández, F.; García-Mera, X.; López, C.; Morales, M.; Rodríguez-Borges, J. E. A Convenient Synthesis of Novel Pyrimidinyl-5′-nor-1′-homocarbanucleosides Based on Indanol. *Synthesis* **2005**, *20*, 3549−3554.

(33) Nieto, M.; Caamaño, O.; Fernández, F.; Gómez, M.; Balzarini, J.; De Clercq, E. Synthesis, antiviral and cytostatic activities, of carbocyclic nucleosides incorporating a modified cyclopentane ring. Part 4: adenosine and uridine analogues. *Nucleosides, Nucleotides Nucleic Acids* **2002**, *21* (3), 243−255.

(34) Figueira, M. J.; Caamaño, O.; Fernández, F.; Rodríguez-Borges, J. E.; Balzarini, J.; De Clercq, E. Synthesis and Biological Evaluation of Carbocyclic Nucleosides with 2′,3′-Dihomo-xylo-carbocyclic or Carbocyclic Fused to a Tetrahydrofuran Ring. *Synthesis* **2004**, *12*, 1991−1995.

(35) Hergueta, A. R.; Fernandez, F.; Lopez, C.; Balzarini, J.; De Clercq, E. Novel carbocyclic nucleosides containing a cyclobutyl ring: adenosine analogues. *Chem. Pharm. Bull. (Tokyo)* **2001**, *49* (9), 1174−7.

(36) Balo, C.; Fernández, F.; Lens, E.; López, C.; Andrei, G.; Snoeck, R.; Balzarini, J.; De Clercq, E. Novel carbocyclic nucleosides containing a cyclopentyl ring. Adenosine and uridine analogues. *Archiv der Pharmazie* **1997**, *330*, 265−267.

(37) Blanco, J. M.; Caamaño, O.; Fernandez, F.; Rodriguez-Borges, J. E.; Balzarini, J.; de Clercq, E. Carbocyclic analogues of nucleosides from bis-(Hydroxymethyl)-cyclopentane: synthesis, antiviral and cytostatic activities of adenosine, inosine and uridine analogues. *Chem. Pharm. Bull. (Tokyo)* **2003**, *51*, (9), 1060−3.

(38) Wnuk, S. F.; Yuan, C. S.; Borchardt, R. T.; Balzarini, J.; De Clercq, E.; Robins, M. J. Anticancer and antiviral effects and inactivation of *S*-adenosyl-L-homocysteine hydrolase with 5′-carboxaldehydes and oximes synthesized from adenosine and sugar-modified analogues. *J. Med. Chem.* **1997**, *40* (11), 1608−18.

(39) Gonzalez-Diaz, H.; Viña, D.; Santana, L.; de Clercq, E.; Uriarte, E. Stochastic entropy QSAR for the in silico discovery of anticancer compounds: prediction, synthesis, and in vitro assay of new purine carbanucleosides. *Bioorg. Med. Chem.* **2006**, *14* (4), 1095−107.

(40) Prekupec, S.; Svedruzic, D.; Gazivoda, T.; Mrvos-Sermek, D.; Nagl, A.; Grdisa, M.; Pavelic, K.; Balzarini, J.; De Clercq, E.; Folkers, G.; Scapozza, L.; Mintas, M.; Raic-Malic, S. Synthesis and biological evaluation of iodinated and fluorinated 9-(2-hydroxypropyl) and 9-(2-hydroxyethoxy)methyl purine nucleoside analogues. *J. Med. Chem.* **2003**, *46* (26), 5763−72.

(41) Moosavi-Movahedi, A. A.; Hakimelahi, S.; Chamani, J.; Khodarahmi, G. A.; Hassanzadeh, F.; Luo, F. T.; Ly, T. W.; Shia, K. S.; Yen, C. F.; Jain, M. L.; Kulatheeswaran, R.; Xue, C.; Pasdar, M.; Hakimelahi, G. H. Design, synthesis, and anticancer activity of phosphonic acid diphosphate derivative of adenine-containing butenolide and its water-soluble derivatives of paclitaxel with high antitumor activity. *Bioorg. Med. Chem.* **2003**, *11* (20), 4303−13.

(42) Raic-Malic, S.; Svedruzic, D.; Gazivoda, T.; Marunovic, A.; Hergold-Brundic, A.; Nagl, A.; Balzarini, J.; De Clercq, E.; Mintas, M. Synthesis and antitumor activities of novel pyrimidine derivatives of 2,3-*O*,*O*-dibenzyl-6-deoxy-L-ascorbic acid and 4,5-didehydro-5,6-dideoxy-L-ascorbic acid. *J. Med. Chem.* **2000**, *43*, (25), 4806−11.

(43) Santana, L.; Teijeira, M.; Uriarte, E.; Balzarini, J.; De Clercq, E. Synthesis, conformational analysis and antiviral and antitumoral activity of new 1,2-disubstituted carbocyclic nucleosides. *Eur. J. Med. Chem.* **2002**, *37* (9), 755−60.

(44) Hakimelahi, G. H.; Mei, N. W.; Moosavi-Movahedi, A. A.; Davari, H.; Hakimelahi, S.; King, K. Y.; Hwu, J. R.; Wen, Y. S. Synthesis and biological evaluation of purine-containing butenolides. *J. Med. Chem.* **2001**, *44* (11), 1749−57.

(45) Hakimelahi, G. H.; Moosavi-Movahedi, A. A.; Sambaiah, T.; Zhu, J. L.; Ethiraj, K. S.; Pasdar, M.; Hakimelahi, S. Reactions of purines-containing butenolides with L-cysteine or N-acetyl-L-cysteine as model biological nucleophiles: a potent mechanism-based inhibitor of ribonucleotide reductase caused apoptosis in breast carcinoma MCF7 cells. *Eur. J. Med. Chem.* **2002**, *37* (3), 207−17.

(46) Hatse, S.; Naesens, L.; De Clercq, E.; Balzarini, J. N6-cyclopropyl-PMEDAP: a novel derivative of 9-(2-phosphonylmethoxyethyl)-2,6-diaminopurine (PMEDAP) with distinct metabolic, antiproliferative, and differentiation-inducing properties. *Biochem. Pharmacol.* **1999**, *58* (2), 311−23.

(47) Teran, C.; Santana, L.; Teijeira, M.; Uriarte, E.; De Clercq, E. Design, synthesis, conformational analysis and biological activities of purine-based 1,2-di-substituted carbocyclic nucleosides. *Chem. Pharm. Bull.* (Tokyo) **2000**, *48* (2), 293−5.

(48) Wang, Z. X.; Wiebe, L. I.; Balzarini, J.; De Clercq, E.; Knaus, E. E. Chiral synthesis of 4-[1-(2-deoxy-beta-L-ribofuranosyl)] derivatives of 2-substituted 5-fluoroaniline: "cytosine replacement" analogues of deoxy-beta-L-cytidine. *J. Org. Chem.* **2000**, *65* (26), 9214−9.

(49) Wnuk, S. F.; Yuan, C. S.; Borchardt, R. T.; Balzarini, J.; De Clercq, E.; Robins, M. J. Nucleic acid related compounds. 84. Synthesis of 6′-(*E* and *Z*)-halohomovinyl derivatives of adenosine, inactivation of *S*-adenosyl-L-homocysteine hydrolase, and correlation of anticancer and antiviral potencies with enzyme inhibition. *J. Med. Chem.* **1994**, *37* (21), 3579−87.

(50) Kundu, N. G.; Das, P.; Balzarini, J.; De Clercq, E. Palladium-catalyzed synthesis of [E]-6-(2-acylvinyl)uracils and [E]-6-(2-acylvinyl)-1-[(2-hydroxyethoxy)methyl]uracils−their antiviral and cytotoxic activities. *Bioorg. Med. Chem.* **1997**, *5* (11), 2011−8.

(51) Mikhailopulo, I. A.; Poopeiko, N. E.; Pricota, T. I.; Sivets, G. G.; Kvasyuk, E. I.; Balzarini, J.; De Clercq, E. Synthesis and antiviral and cytostatic properties of 3′-deoxy-3′-fluoro- and 2′-azido-3′-fluoro-2′,3′-dideoxy-D-ribofuranosides of natural heterocyclic bases. *J. Med. Chem.* **1991**, *34* (7), 2195−202.

(52) Raic-Malic, S.; Hergold-Brundic, A.; Nagl, A.; Grdisa, M.; Pavelic, K.; De Clercq, E.; Mintas, M. Novel pyrimidine and purine derivatives of L-ascorbic acid: synthesis and biological evaluation. *J. Med. Chem.* **1999**, *42* (14), 2673−8.

(53) De Clercq, E.; Balzarini, J.; Torrence, P. F.; Mertes, M. P.; Schmidt, C. L.; Shugar, D.; Barr, P. J.; Jones, A. S.; Verhelst, G.; Walker, R. T. Thymidylate synthetase as target enzyme for the inhibitory activity of 5-substituted 2′-deoxyuridines on mouse leukemia L1210 cell growth. *Mol. Pharm.* **1981**, *19* (2), 321−30.

(54) De Clercq, E.; Descamps, J.; Balzarini, J.; Giziewicz, J.; Barr, P. J.; Robins, M. J. Nucleic acid related compounds. 40. Synthesis and biological activities of 5-alkynyluracil nucleosides. *J. Med. Chem.* **1983**, *26* (5), 661−6.

(55) Robins, M. J.; Hatfield, P. W.; Balzarini, J.; De Clercq, E. Nucleic acid related compounds. 47. Synthesis and biological activities of pyrimidine and purine "acyclic" nucleoside analogues. *J. Med. Chem.* **1984**, *27* (11), 1486−92.

(56) Hunston, R. N.; Jones, A. S.; McGuigan, C.; Walker, R. T.; Balzarini, J.; De Clercq, E. Synthesis and biological properties of some cyclic phosphotriesters derived from 2′-deoxy-5-fluorouridine. *J. Med. Chem.* **1984**, *27* (4), 440−4.

(57) Al-Razzak, L. A.; Schwepler, D.; Decedue, C. J.; Balzarini, J.; De Clercq, E.; Mertes, M. P. 5-Quinone derivatives of 2′-deoxyuridine 5′-phosphate: inhibition and inactivation of thymidylate synthase, antitumor cell, and antiviral studies. *J. Med. Chem.* **1987**, *30* (2), 409−19.

(58) Bobek, M.; An, S. H.; Skrincosky, D.; De Clercq, E.; Bernacki, R. J. 2′-Fluorinated isonucleosides. 1. Synthesis and biological activity of some methyl 2′-deoxy-2′-fluoro-2′-pyrimidinyl-D-arabinopyranosides. *J. Med. Chem.* **1989**, *32* (4), 799−807.

(59) Pestana, M.; Gageiro, J., Análise de dados para Ciências Sociais. A Complementaridade do SPSS; 2nd ed.; Ediç˜oes Sílabo: Lisboa, 2000; p 570.

(60) 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-97), 1997; American Association for Artificial Intelligence Press: 1997.

(61) Toivonen, H.; Srinivasan, A.; King, R. D.; Kramer, S.; Helma, C. Statistical evaluation of the Predictive Toxicology Challenge 2000−2001. *Bioinformatics* **2003**, *19* (10), 1183−93.

(62) Benigni, R.; Giuliani, A. Putting the Predictive Toxicology Challenge into perspective: reflections on the results. *Bioinformatics* **2003**, *19* (10), 1194−200.

(63) Randić, M.; Wilkins, C. L. Graph Theoretical Approach to Recognition of Structural Similarity in Molecules. *J. Chem. Inf. Comput. Sci.* **1979**, *19* (1), 31−37.

(64) Randić, M. On Characterization of Cyclic Structures. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1063−1071.

(65) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20* (4), 269−76.

(66) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17* (2−4), 241−53.

(67) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des.* **2002**, *16* (5−6), 357−69.

(68) Todeschini, R.; Consonni, V.; Pavan, M. *Dragon Software*, version 2.1; 2002.

(69) Statsoft, I. *STATISTICA (data analysis software system)*, version 6.0; 2002.

(70) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Pandu, K. S. Structure-activity relationships of antifilarial antimycin analogues: a multivariate pattern recognition study. *J. Med. Chem.* **1990**, *33*, 136−142.

(71) Randić, M. Correlation of enthalphy of octanes with orthogonal connectivity indices. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45−59.

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

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

(74) Peterangelo, S. C.; Seybold, P. G. Synergistic interactions among QSAR descriptors. *Int. J. Quant. Chem.* **2004**, *96*, 1−9.

(75) Osborne, J.; Waters, E. Four assumptions of multiple regression that researchers should always test. *Pract. Assess. Res. Eval.* **2002**, *8*, 2.

(76) Van Waterbeemd, H. Chemometric Methods in Molecular Design. In *Method and Principles in Medicinal Chemistry*; Manhnhold, R.; Krogsgaard-Larsen, R.; Timmerman, H., Eds.; VCH: Weinheim, 1995; Vol. 2, pp 265−282.

JM061445M