European Journal of Medicinal Chemistry 44 (2009) 4826–4840

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Prediction of telomerase inhibitory activity for acridinic derivatives based on chemical structure

Daimel Castillo-González [a,b], Miguel Ángel Cabrera-Pérez [b,*], Maykel Pérez-González [b], Aliuska Morales Helguera [b,c], Alexander Durán-Martínez [c]

[a] Department of Pharmacy, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba
[b] Molecular Simulation and Drug Design Group, Centre of Chemical Bioactive, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba
[c] Department of Chemistry, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba

ARTICLE INFO

Article history:
Received 26 January 2009
Received in revised form
2 June 2009
Accepted 23 July 2009
Available online 6 August 2009

Keywords:
Telomerase
G-quartet
QSAR
Drug design
Acridines
Acridones

ABSTRACT

Telomerase is a reverse transcriptase enzyme that activates in more than 85% of cancer cells and it is associated with the acquisition of a malignant phenotype. Some experimental strategies have been suggested in order to avoid the enzyme effect on unstopped telomere elongation. One of them, the stabilization of the G-quartet structure, has been widely studied. Nevertheless, no QSAR studies to predict this activity have been developed. In the present study a classification model was carried out to identify, through molecular descriptors with structural fragments and groups information, those acridinic derivatives with better inhibitory concentration on telomerase enzyme. A linear discriminant model was developed to classify a data set of 90 acridinic derivatives (48 more potent derivatives with $IC_{50} < 1\ \mu M$ and 42 less potent with $IC_{50} \geq 1\ \mu M$). The final model fit the data with sensitivity of 87.50% and specificity of 82.85%, for a final accuracy of 85.33%. The predictive ability of the model was assessed by a prediction set (15 compounds of 90% and 82.29% of prediction accuracy); a tenfold full cross-validation procedure (removing 15 compounds in each cycle, 84.80% of good prediction) and the prediction of inhibitory concentration on telomerase enzyme for external data of 10 novel acridines (90% of good prediction). The results of this study suggest that the established model has a strong predictive ability and can be prospectively used in the molecular design and action mechanism analysis of this kind of compounds with anticancer activity.

1. Introduction

Telomere, a complex of guanine-rich repeat sequences and associated proteins, caps and protects every eukaryotic chromosome end against chromosomal fusion, recombination, and terminal DNA degradation [1]. Telomeric DNA consists of short guanine-rich repeat sequences in all eukaryotes with linear chromosomes, and its length in human somatic cells is remarkably heterogeneous among individuals ranging from 5 to 20 kb, according to age, organ, and the proliferative history of each cell [2]. During the process of DNA synthesis and cell division, telomeres shorten as a result of the incomplete replication of linear chromosomes, the so-called 'end-replication problem'. This progressive telomere shortening is one of the molecular mechanisms underlying ageing, as critically short telomeres trigger chromosome senescence and loss of cell viability [2–4]. To prevent degradation by exonucleases or processing as damaged DNA, the telomere 3′ single-strand overhang folds back into the D-loop of duplex telomeric DNA to form a protective 'T-loop', which is reinforced with TRF2 and other telomeric DNA-binding proteins named Shelterin [5].

Telomerase is a ribonucleoprotein composed of human telomerase reverse transcriptase (hTERT), the catalytic subunit, and its template RNA (hTERC) [6–9]. As direct evidence that telomere erosion plays a major role in cellular senescence, ectopic expression of hTERT in normal human cells with endogenous hTERC resulted in activation of telomerase, stabilization of telomere lengths and extension of cellular life span. The catalytic subunit (hTERT) is the rate-limiting factor for telomerase activity both biologically and biochemically. In 85–90% of cancer cells and in parasites, telomerase is activated [10,11]. This enzyme is inactive in most normal somatic cells, providing a potentially specific therapeutic target. In order to extend telomere DNA the telomerase enzyme requires the telomere primer to be single-stranded. The formation of higher ordered structures such as G-quadruplexes prevents hybridization

* Corresponding author. Tel.: +53 42 281 473/192; fax: +53 42 281130.
E-mail addresses: migue@gammu.com, macabrera@uclv.edu.cu, macabreraster@gmail.com (M.Á. Cabrera-Pérez).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.07.029

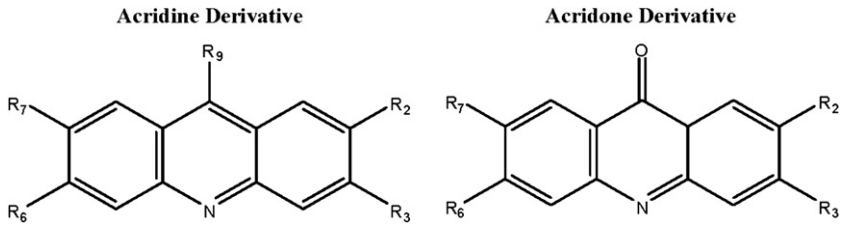


Fig. 1. General chemical structures for the acridinic derivatives used in the QSAR study.

of the telomerase RNA template onto the primer and thus inhibits telomerase activity through an indirect topological mechanism [12].

Stabilization of the quadruplex conformation of telomeres, such as by binding small molecules, has been shown to be an effective method to inhibit telomerase activity [13–16]. The development of small molecules that can selectively bind to and stabilize the G-quadruplex conformation of the telomere is therefore a current area of interest in anticancer as well as antiparasitic drug design. Compounds that have been shown to bind to quadruplex DNA have traditionally been planar, aromatic compounds that bind via external end-stacking to the G-quartet on either one end or both ends of the quadruplex [17–30]. These compounds, which include anthraquinones, cationic porphyrins, acridines, macrocyclic compounds and analogs, have planar aromatic surface areas that mimic the large planar surface of the G-tetrads in quadruplex DNA [18,19,31]. Since essentially all known quadruplex DNA binders are based on, or derived from duplex intercalators, many of them exhibit little selectivity for quadruplex over duplex structures and this can result in nonspecific cytotoxicity. Increasing the selectivity of telomerase inhibitors for their quadruplex targets is an important focus of research.

QSAR (Quantitative Structure–Activity Relationship) studies have demonstrated to be a very useful tool in the design and development of novel compounds with different biological activities. One of the main fields where this kind of studies have been carried out is in the prediction of antitumoral activity [32,33]. Nevertheless, the telomerase inhibitory concentration of potential candidates have not been developed using the QSAR methodology. For this reason, in the present paper we intend to develop a QSAR study of a family of acridinic derivatives with demonstrated telomerase inhibitory activity based on stabilization of G-quadruplex. The main goal of the present research is to develop a classification model in order to identify the more potent acridinic derivatives that permit the design of novel candidates taking into consideration the structural information contained in simple molecular descriptors.

2. Methods

2.1. Data set

A data set of 90 acridinic derivatives was carefully assembled from literature [30,34–39]. The group was composed by mono, di, and tri-substituted acridines and acridones. The general chemical structures of these derivatives are described in Fig. 1.

All these compounds have a proved inhibitory potency against telomerase enzyme [40]. This activity has been mainly attributed to the planarity of these aromatic structures, which could intercalate within the double-stranded DNA structure, thus interfering with the cellular machinery. The inhibition of telomerase, by stabilization of G-quartet structure, was determined using the cell-free telomeric repeat amplification protocol (TRAP) assay [41]. Prior to the evaluation in the TRAP telomerase assays, the compounds were tested for their ability to inhibit Taq polymerase, in order to ensure that no false positives were obtained in the subsequent TRAP assays. TRAP assays were then performed at concentrations lower

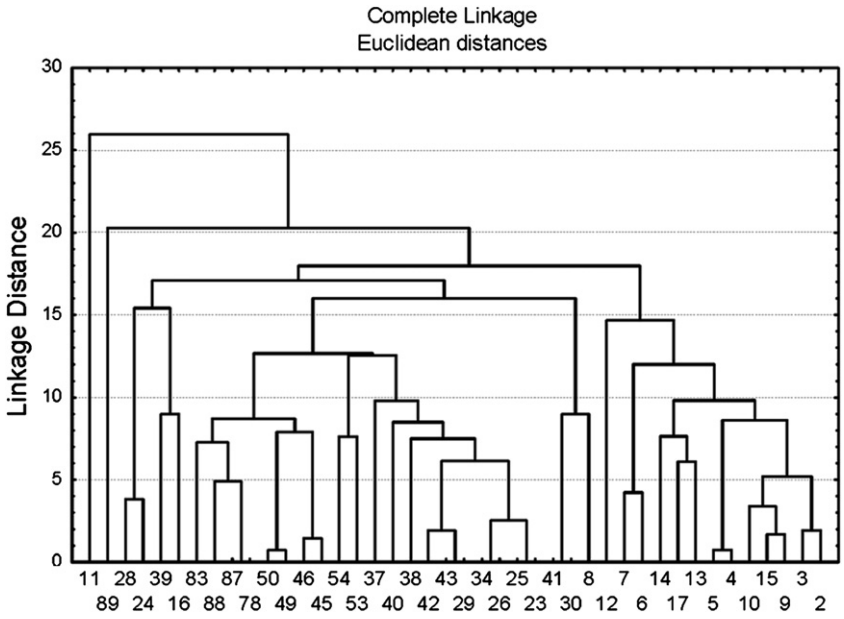


Fig. 2. A dendrogram illustrating the results of the hierarchical cluster analysis for the less potent acridinic derivatives.

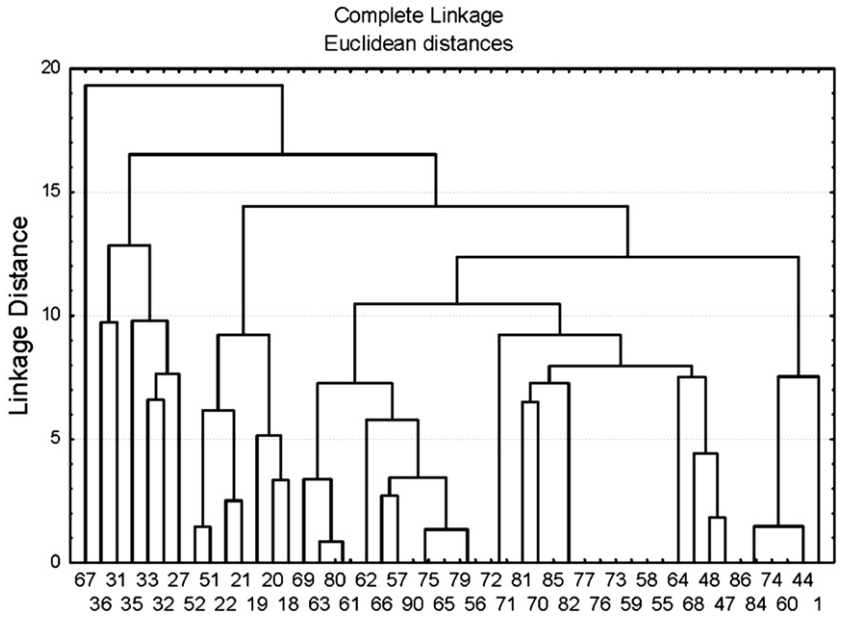


Fig. 3. A dendrogram illustrating the results of the hierarchical cluster analysis for the more potent acridinic derivatives.

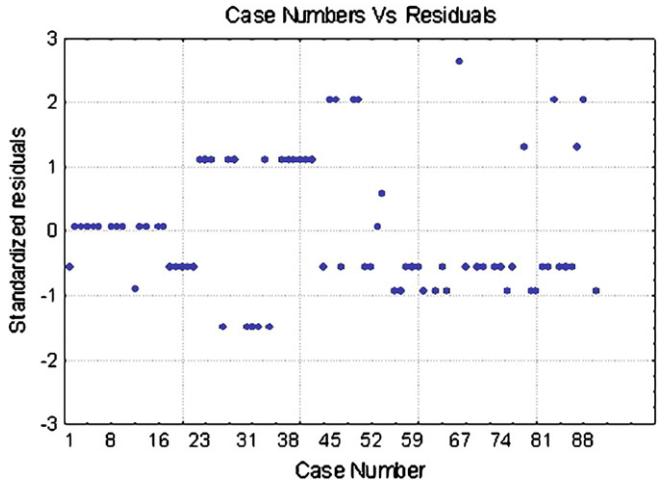


Fig. 4. Distribution of the standardized residuals for all acridinic derivatives used in the study.

than which Taq polymerase inhibition was observed, using extracts from the A2780 ovarian carcinoma cell line as a source of telomerase.

This experimental method has certain variability and the quantitative value of telomerase inhibition fluctuates from one experiment to the next; however, the fluctuation within the same experiment is ±10% [42]. For the complete data set was reported the molecular structure and the telomerase inhibitory concentration, by stabilization of G-quartet structure (see Table 4 for details).

The compounds were first clustered in two groups according to their mean telomerase inhibitory concentration values (IC_{50}). The first group, more potent compounds, includes all chemicals with $IC_{50} < 1$ μM, while the second one includes the less potent derivatives with $IC_{50} \geq 1$ μM [43,44]. This classification criterion was adopted not only because this value have been reported as a cut-off value but also to get a reasonable ratio between both chemical groups in the data set.

In order to obtain a validated and predicted QSAR model, the available data set was divided into training and prediction sets, using a k-means cluster analysis (k-MCA). Firstly, we carried out a k-MCA1 with the more potent compounds and later another, k-MCA2, using the less potent compounds. The k-MCA was carried out with the STATISTICA software 6.0 [45]. For acceptable statistical quality of data clusters, 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). We also inspected the standard deviation between and within clusters, the respective Fisher ratio and their p-level of significance considered to be lower than 0.05. Compounds for the training and test sets were randomly collected from the previous clusters. This procedure permitted us to select, in a representative way and in all level of the linking distance (*Y*-axis), compounds for the training and prediction sets. In Figs. 2 and 3 are shown the results of the developed cluster analysis.

Table 1
Inter-correlation among the three descriptors selected as statistically significant by GLDA analysis.

	nCar	*nCbH*	*nArNHR*
nCar	1.00	0.97	0.12
nCbH		1.00	0.18
nArNHR			1.00

Table 2
Box's test of equality of covariance matrices.

Log determinants		
Disc	Rank	Log determinant
−1	3	0.101
1	3	−39.065
Pooled within-groups	3	−0.491
Test results		
	Box's M	1484.262
	F Approx.	236.254
	df1	6
	df2	36748.004
	Sig.	0.000

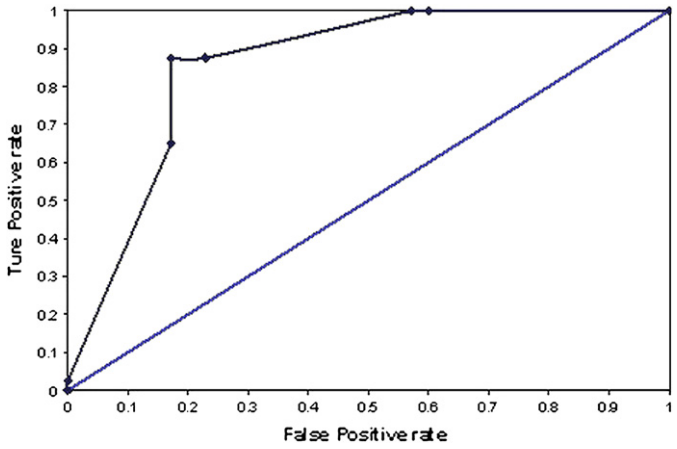


Fig. 6. Receiver Operating Characteristic (ROC) curve for the classification model (Eq (3)).

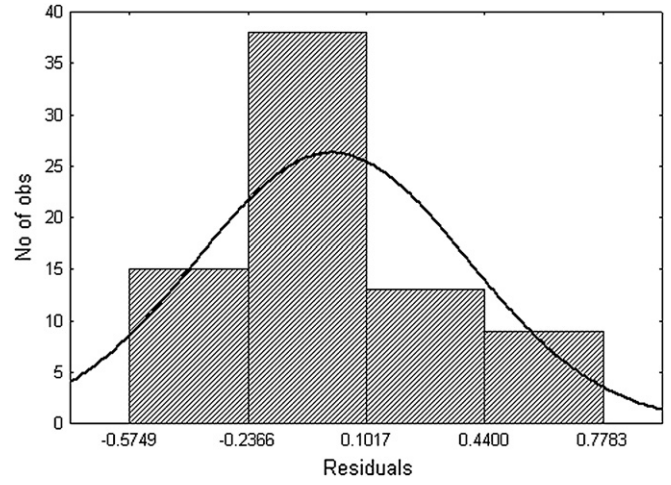


Fig. 5. Histogram for the frequency distribution of residuals within the classification model Eq. (3).

Finally, the training and prediction sets were composed of 75 compounds and 15 compounds, respectively. The compounds belonging to the prediction set were never used in the development of the discriminant function and they were reserved to assess the discriminant model obtained.

2.2. Molecular descriptors

Our study is based on the different sets of 1D-descriptors available in the DRAGON software, which in turn have a long history of structure–activity and structure–property correlations [46]. Two main family of descriptors were used, the atom-centred fragments and the functional group counts. The input of the software consists of SMILES (simplified molecular input line entry specification) codes for each compound [47].

Count descriptors directly encode particular features of molecular structure and are simply obtained from the chemical structure of molecules by counting defined elements such as atoms (nAT), bonds (nBT), rings (nCIC), H-bond acceptor (nHA) and H-bond donor (nHD) atoms, path counts, walk counts and so on. If atom-types are considered, atom-type counts are obtained such as number of carbon atoms (nC), number of halogens (nX), number of oxygen atoms (nO), etc. Analogously, functional group counts and fragment counts are calculated such as number of hydroxyl-groups (nOH), number of nitro-groups (nNO), number of amino-groups (nNH_2), etc.

Atom-centred fragments are also very simple molecular descriptors that describe each atom by its own atom-type and the bond-types and atom-types of its first neighbor. Functionalities in a molecule can be represented by two to five atoms, which consist of a central atom and its neighboring bonded atoms. Each fragment is represented by a single descriptor. Their use greatly increases the specific chemical information regarding different functional groups.

2.3. Model search

The discriminant function (Eq. (1)), which best describe telomerase inhibitory concentration values as a linear combination of the predictor X-variables (1D-descriptors) and weighted by the a_n

Table 3
Results of the inter-relation of molecular descriptors in the prediction of different range of telomerase inhibitory activity.

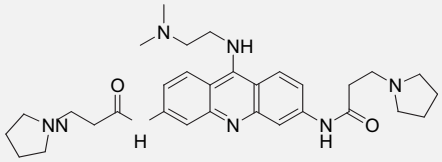
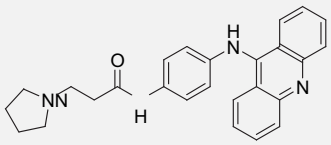
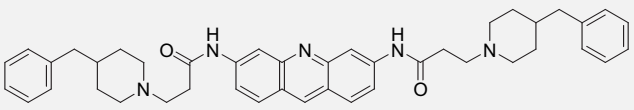
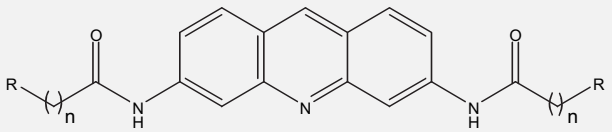
Molecular structure	No.	IC_{50} (μM)	*nArNHR*	*nCar*	*nCbH*
	65	0.018	1	13	6
	54	>20	1	19	12
	12	>50	0	25	16

Table 4
Telomerase inhibition data set for acridinic derivatives. Results of the classification for the training and test sets compounds.



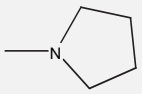
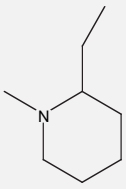
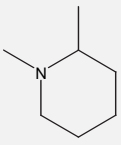
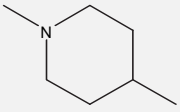
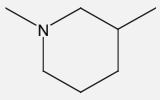
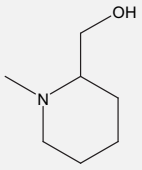
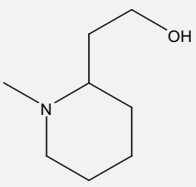
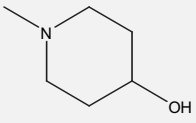
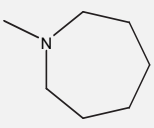
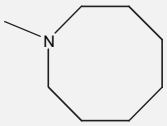
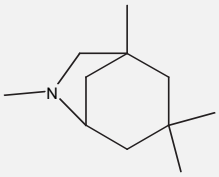
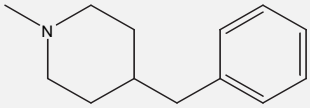
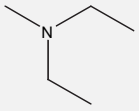
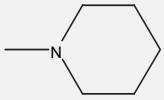
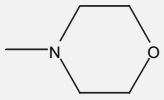
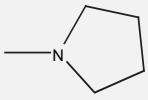
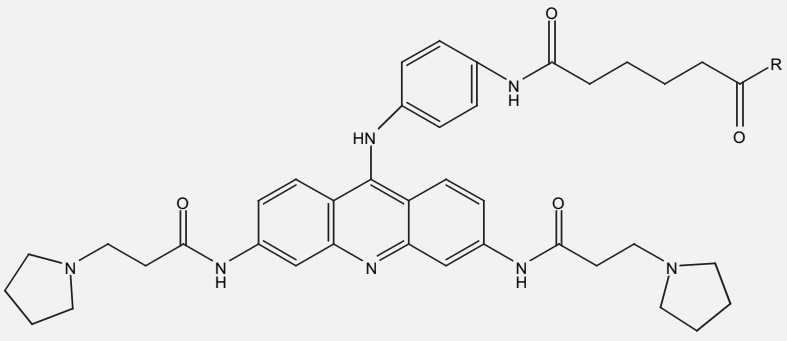
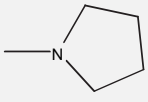
No.	R	*n*	IC_{50} (μM)	Experim. class	Pred. class	Reference
1(**53**)		1	5.2	–	–	[39]
2(**2**)		2	2.7	–	–	[35]
3(**3**)		2	2.6	–	–	[35]
4(**4**)		2	1.35	–	–	[35]
5(**5**)		2	4.4	–	–	[35]
6(**6**)		2	5.4	–	–	[35]
7(**7**)[a]		2	4.1	–	–	[35]
8(**8**)		2	8.0	–	–	[35]
9(**9**)		2	3.1	–	–	[35]

Table 4 (*continued*)

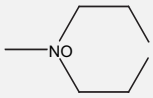
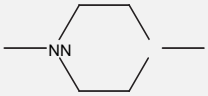
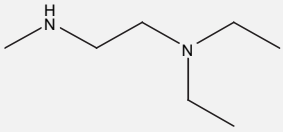
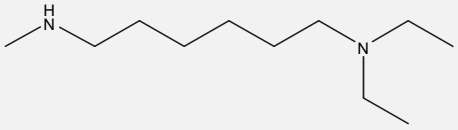
No.	R	*n*	IC_{50} (μM)	Experim. class	Pred. class	Reference
10(**10**)		2	>50	–	–	[35]
11(**11**)[a]		2	>50	–	–	[35]
12(**12**)		2	>50	–	–	[35]
13(**13**)		2	5.8	–	–	[35]
14(**14**)		2	8.2	–	–	[35]
15(**15**)[a]		2	2.8	–	–	[35]
16(**16**)		2	>50	–	–	[35]
17(**17**)		2	5.2	–	–	[35]

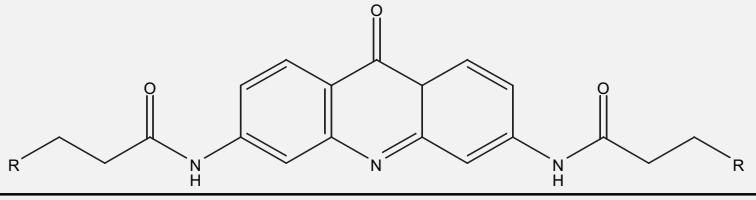


No.	R	IC_{50} (μM)	Experim. class	Pred. class	Reference
18(**18**)		0.318	+	+	[36]

(*continued on next page*)

Table 4 (*continued*)

No.	R	IC_{50} (μM)	Experim. class	Pred. class	Reference
19(**19**)		0.267	+	+	[36]
20(**20**)		0.165	+	+	[36]
21(**21**)		0.098	+	+	[36]
22(**22**)		0.08	+	+	[36]



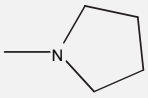
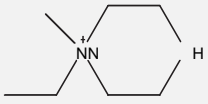
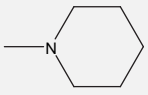
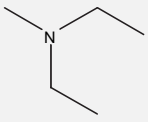
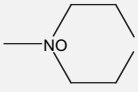
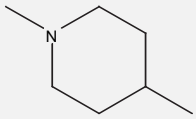
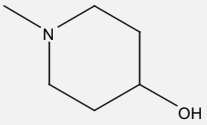
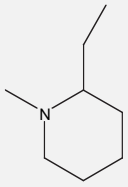
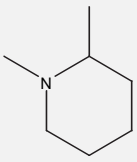
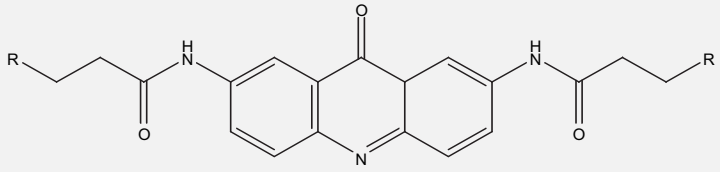
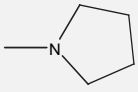
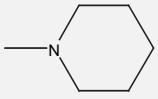
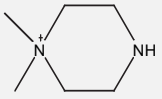
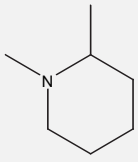
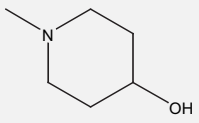
No	R	IC_{50} (μM)	Experim. class	Pred. class	Reference
23(**23**)		8.1	–	–	[30]
24(**24**)		1.9	–	–	[30]
25(**34**)		5.9	–	–	[30]
26(**37**)		4.3	–	–	[30]
27(**38**)		2.7	–	–	[30]
28(**39**)		49	–	–	[30]

Table 4 (*continued*)

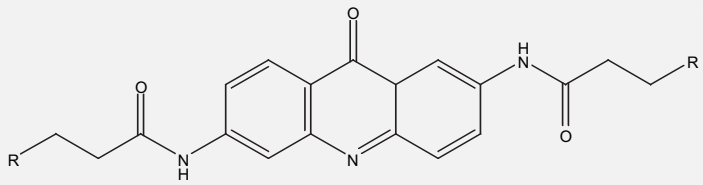
No	R	IC_{50} (μM)	Experim. class	Pred. class	Reference
29(**40**)		1.7	–	–	[30]
30(**41**)		1.7	–	–	[30]
31(**42**)		2.3	–	–	[30]
32(**43**)[a]		2.3	–	–	[30]

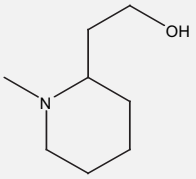
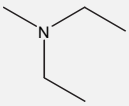
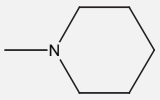
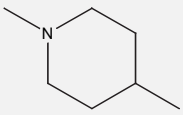
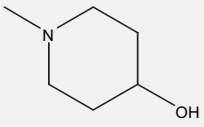


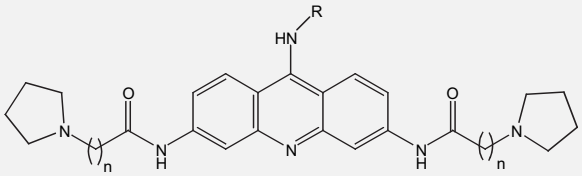
No	R	IC_{50} (μM)	Experim. class	Pred. class	Reference
33(**25**)		5.8	–	–	[30]
34(**26**)[a]		1.9	–	–	[30]
35(**27**)		0.6	+	–	[30]
36(**28**)		1.9	–	–	[30]
37(**29**)		1.5	–	–	[30]
38(**30**)[a]		2.3	–	–	[30]

(*continued on next page*)

Table 4 (*continued*)



No	R	IC_{50} (μM)	Experim. class	Pred. class	Reference
39(**31**)		0.2	+	–	[30]
40(**32**)		0.7	+	–	[30]
41(**33**)		0.4	+	–	[30]
42(**35**)		0.2	+	–	[30]
43(**36**)[a]		0.2	+	–	[30]



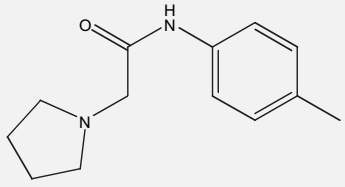
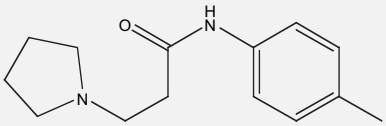
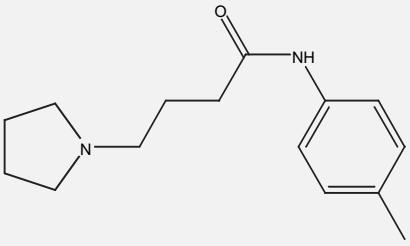
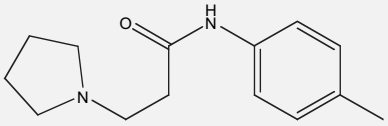
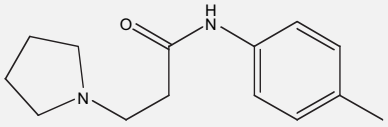
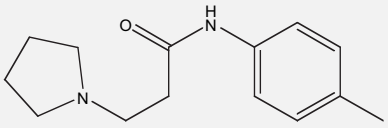
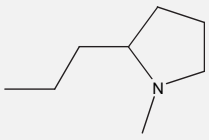
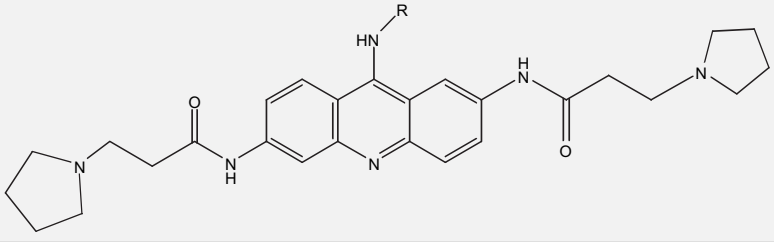
No	R	*n*	IC_{50} (μM)	Experim. class	Pred. class	Reference
44(**1**)	$-C_6H_4N(CH_3)_2$ (p)	2	0.1	+	+	[36]
45(**44**)	$-C_6H_4N(CH_3)_2$ (p)	3	0.099	+	+	[38]
46(**45**)	$-C_6H_4N(CH_3)_2$ (p)	4	1.93	–	+	[38]
47(**46**)	$-C_6H_4N(CH_3)_2$ (p)	5	6.91	–	+	[38]
48(**47**)		2	0.167	+	+	[38]
49(**48**)[a]		2	0.067	+	+	[38]

Table 4 (*continued*)

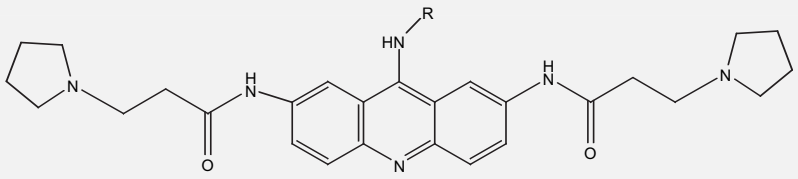
No	R	*n*	IC_{50} (μM)	Experim. class	Pred. class	Reference
50(**49**)		2	0.117	+	+	[38]
51(**50**)		3	3.26	−	+	[38]
52(**51**)		4	0.255	+	+	[38]
53(**52**)		5	0.146	+	+	[38]
54(**55**)[a]	$-C_6H_4NH_2$ (p)	2	0.074	+	+	[42]
55(**56**)	$-CH_2CH_2CH_2N(CH_3)_2$	2	0.060	+	+	[42]
56(**57**)	$-CH_2CH_2C_5H_{10}N$	2	0.05	+	+	[42]
57(**58**)	$-C_6H_4NH_2$ (m)	2	0.06	+	+	[42]
58(**59**)	$-C_6H_4NH_2$ (o)	2	0.02	+	+	[42]
59(**60**)	$-C_6H_4N(CH_3)_2$ (m)	2	0.1	+	+	[42]
60(**61**)	$-C_6H_{11}$ (c)	2	0.09	+	+	[42]
61(**62**)[a]	$-CH_2CH_2OCH_3$	2	0.14	+	+	[42]
62(**63**)	$-C_7H_{13}$ (c)	2	0.21	+	+	[42]
63(**64**)	$-C_6H_4COCH_3$ (p)	2	0.04	+	+	[42]
64(**65**)	$-CH_2CH_2N(CH_3)_2$	2	0.018	+	+	[42]
65(**66**)[a]		2	0.018	+	+	[42]
66(**67**)	$-CH_2C_5H_4N$ (m) (c)	2	0.066	+	+	[42]
67(**68**)	$-C_6H_4NHCOCH_3$ (m)	2	0.1	+	+	[42]
68(**69**)[a]	$-C_3H_5$ (c)	2	0.05	+	+	[42]
69(**70**)	$-C_6H_4F$ (p)	2	0.07	+	+	[42]
70(**71**)	$-C_6H_4SCH_3$ (o)	2	0.15	+	+	[42]
71(**72**)[a]	$-C_6H_4SCH_3$ (m)	2	0.1	+	+	[42]



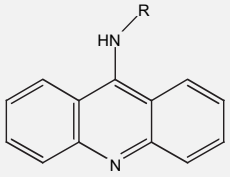
No	R	IC_{50} (μM)	Experim. class	Pred. class	Reference
72(**73**)	$-C_6H_4NH_2$ (p)	0.08	+	+	[42]
73(**74**)	$-C_6H_4N(CH_3)_2$ (p)	0.17	+	+	[42]
74(**75**)	$-CH_2CH_2N(CH_3)_2$	0.27	+	+	[42]
75(**76**)	$-C_6H_4NH_2$ (m)	0.21	+	+	[42]
76(**77**)[a]	$-C_6H_4NH_2$ (o)	0.11	+	+	[42]
77(**78**)	$-C_6H_5$	1.33	−	−	[42]
78(**79**)	$-CH_2CH_2CH_2N(CH_3)_2$	0.08	+	+	[42]
79(**80**)	$-C_6H_{11}$ (c)	0.21	+	+	[42]

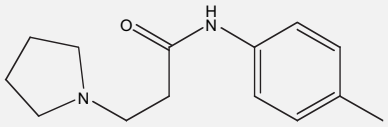
(*continued on next page*)

Table 4 (*continued*)



No	R	IC_{50} (μM)	Experim. class	Pred. class	Reference
80(**81**)	$-C_6H_4OCH_3$ (p)	0.46	+	+	[42]
81(**82**)	$-C_6H_4NH_2$ (o)	0.17	+	+	[42]
82(**83**)	$-C_6H_4NH_2$ (m)	1.09	−	+	[42]
83(**84**)	$-C_6H_4N(CH_3)_2$ (m)	0.6	+	+	[42]
84(**85**)	$-C_6H_4NH_2$ (p)	0.2	+	+	[42]
85(**86**)	$-C_6H_4N(CH_3)_2$ (p)	0.5	+	+	[42]
86(**87**)	$-C_6H_5$	1.29	−	−	[42]
87(**88**)	$-C_6H_4OCH_3$ (m)	2.73	−	+	[42]
88(**89**)[a]	$-C_6H_4OH$ (o)	1.03	−	+	[42]
89(**90**)	$-CH_2CH_2N(CH_3)_2$	0.57	+	+	[42]



No	R	IC_{50} (μM)	Experim. class	Pred. class	Reference
90(**54**)		>20	−	−	[38]

[a] Compounds used as a prediction set; c: cyclic, m: meta, p: para, o: ortho.

coefficients, was obtained with the General Linear Discriminant Analysis Module (GLDA) implemented in STATISTICA 6.0 [45].

$$IC_{50} = a_1A_1 + a_2A_2 + a_3A_3 \ldots + a_nA_n + a_0 \quad (1)$$

In developing these classification functions, IC_{50} values of +1 and −1 were assigned to more potent and less potent compounds, respectively. In addition, the compounds were considered unclassified (U) by the model when the differences in the percentage of classification between two groups did not differ by more than 5%.

The "*best subset*" technique, using the Wilks Lambda value as the criterion for choosing the best subset of predictor effects, was applied to select the molecular descriptors (X-variables) with the highest influence on the telomerase inhibition.

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 more 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 [48,49]. One very useful and informative approach of avoiding multicollinearity is the orthogonal-descriptor technique suggested by Randić some years ago [48,49]. In the Randić approach, after choosing a starting descriptor, subsequent descriptors are added only as their orthogonal complements to the descriptor already present. This approach has the advantages that the equation coefficients are stable, and the new information supplied by each additional descriptor is clearly distinguished in the final equation statistics. In this paper, to tackle the multicollinearity problem, we have applied the Randić approach. The resulting orthogonal-descriptor model was standardized afterward.

Several diagnostic statistical tools were used for evaluating the final model equation, in terms of goodness-of-fit and goodness-of-prediction. The goodness-of-fit was estimated by standard statistics such as the λ value, the square of Mahalanobis distance (D^2), the Fisher ratio (F), the corresponding p-level (p) as well as the proportion between the cases and variables in the equation. The validity of the preadopted assumptions such as normality, homocedasticity, noncollinearity and linearity of the model was also checked. The goodness-of-prediction of the final model was assessed by an internal cross-validation (CV), specifically by the leave-group-out (LGO) technique [50,51]. Basically, CV-LGO consists of forming several subsets from the complete data set, each missing a small group of k cases ($k = 15\%$ of the data set). 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 (10-fold LGO-CV was carried out).

A more rigorous procedure to assess the "realistic" predictive power of the model and the generalizability of the QSAR model for new chemical compounds was evaluated using and external data. This type of model validation is important, if we take into consideration that the predictive ability of a QSAR model can be estimated using only a set of compounds that was not used for building the model [50,51]. Therefore, it is important to ensure that the prediction algorithms are able to perform well on novel data from the same data domain. In this study a more demanding evaluation is provided by an external validation, where the predictive power of the model was assessed with an

external data of 10 tri-substituted acridines reported in the literature [38].

3. Results and discussion

The best discriminant function to classify the more potent compounds with telomerase inhibitory activity, by stabilization of G-quadruplex, is given below together with the statistical parameters of the GLDA:

$$IC_{50} = -10.40 + 4.11nArNHR + 1.42nCar - 1.90nCbH$$
$$N = 75,\ \lambda = 0.58,\ D_2 = 2.75,\ F(3,71) = 16.62,\ p < 0.001 \qquad (2)$$

As can be appreciated, the large F index and the small p value are indicative of the model statistical significance. In addition, the values of the Wilks λ statictic (λ can take values from cero, perfect discrimination, to one, no discrimination) and the Mahalanobis distance (a measure of the separation between the more potent and less potent groups) show that the model displays an adequate discriminatory capacity.

The parametrical assumptions (normality, homocedasticity and noncollinearity) are very important factors in the application of linear multivariate statistic techniques to carry out any QSAR model. Nevertheless, the validity and statistical significance of any QSAR model is strongly conditioned by the above mentioned factors [52].

Further analysis of the previous classification model (Eq. (2)) should only be carried out after checking the reliability of the parametrical assumptions. In this study the model was obtained by GLDA, being the simplest mathematical form that might be envisaged for the model in absence of any *a priori* information. However, a simple observation of the distribution of standardized residual (observed minus predicted divided by the square root of the residual mean square) for all compounds studied (see Fig. 4) no systematic pattern was observed. This result suggests 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. As can was explain previously multicollinearity does not affect the predictability of the model, but the physical meaning of the resultant model, as well as the influence of the predictive variables may be misinterpreted and consequently erroneous inferences about the activity under study could be obtained. In order to study the collinearity of the variables in our model the cross-correlation matrix was analyzed. As can be seen in Table 1, the pairs of descriptors *nCar* and *nCbH* are strongly correlated, but we have to take care to discard one of them because of they have different chemical information.

Following the Randić approach, we determined orthogonal complements for all variables in Eq. (2), which in turn were further standardized, to then find the best equation (see Eq. (3))

$$IC_{50} = 0.091 + 1.71^1\Omega(nArNHR) + 0.66^2\Omega(nCar) - 1.11^3\Omega(nCbH)$$
$$N = 75,\ \lambda = 0.58, D^2 = 2.75, F(3,71) = 16.62, p < 0.001 \qquad (3)$$

The relative contributions of the variables in the orthogonalized model are quite similar to those in the non-orthogonalized model. For example, the variables *nCbH* and *nArNHR* have similar absolute contributions in Eq. (3), while in Eq. (2) the contribution of *nArNHR* is twofold the contribution of *nCbH*. These variations are due to the fact that the coefficients of the orthogonal model are more stable than the model without orthogonalization. In the case of variables *nCar* and *nCbH* both have different contributions to the prediction of telomerase inhibition, being the absolute contribution of *nCbH* almost twofold the contribution of *nCar*.

We tested the normal distribution of residual plotting the frequency histogram (see Fig. 5) and using different distribution fitting tests such as Kolmogorov–Smirnov and Shapiro–Wilk.

At first sight, the results of the Kolmogorov–Smirnov ($D = 0.251$ with $p < 0.01$) and Shapiro–Wilk ($W = 0.88$ with $p < 0.001$) tests indicate that the hypothesis of normality should be rejected. However, the deviations from normality seem to be not severe as can be seen in the frequency histogram shown in Fig. 5.

After orthogonalization of descriptors the verification of homocedasticity was determined considering the homogeneity of the (co)variances suggested by the Box M statistical test ($p < 0.01$). The results of this test appear in Table 2.

As can be appreciated the statistically significant F ratio ($F = 236.25$; $p < 0.000$) suggested that the population covariance matrices were not equal, which reject the null hypothesis of equal population covariance matrices. Nevertheless, this violation is not very problematic for discriminant analysis, and we should be able to evaluate this likelihood of errors by examining the determinants of the pooled within-groups covariance.

A better threshold for *a priori* classification probability can be estimated by means of the Receiver Operating Characteristics (ROC) curve [53]. This is a useful technique not only for obtaining the best thresholds but also for organizing classifiers [54,55]. As Fig. 6 shows, the optimal threshold for predicting the more potent chemicals with the present QSAR model is 0.87. Further, one can see that the model is not a random since the area under the ROC curve is significantly higher than the area under the random classifier curve (diagonal line) with a value of $0.87u^2$.

As can be seen in Eq. (3) only three variables are present in the discriminant model. The positive contribution to the telomerase

Table 5
General performance of the final model (Eq. (3)) in terms of its Cooper statistic.

		Predicted class		
		More potent (T/P)[a]	Less potent (T/P)[a]	Total
Experimental class	More potent	35/7 (a)	5/1 (b)	40/8 ($a+b$)
	Less potent	6/1 (c)	29/6 (d)	35/7 ($c+d$)
	Total	41/8 ($a+c$)	34/7 ($b+d$)	75/15 ($a+b+c+d$)

Statistic	Formula	Results (T/P)[a]
Sensitivity (true positive rate)	$a/(a+b)$	87.50/87.50
Specificity (true negative rate)	$d/(c+d)$	82.85/85.71
Accuracy	$(a+d)/(a+b+c+d)$	85.33/86.67
Positive predictivity	$a/(a+c)$	85.36/87.50
Negative predictivity	$d/(b+d)$	85.29/85.71
False positive	$c/(c+d)$	17.14/14.28
False negative	$b/(a+b)$	12.50/12.50

[a] (T/P): for training and prediction sets.

Table 6
Results of the 10-fold full cross-validation procedure for the acridinic data set.

	% Classif. (training set)	% Classif. (test set)	λ	F	D^2
1	82.67	100.00	0.565	20.818	3.108
2	85.33	86.67	0.620	14.492	2.388
3	85.33	86.67	0.577	17.318	2.874
4	85.33	86.67	0.586	16.743	2.767
5	84.00	93.33	0.573	17.640	2.907
6	84.00	93.33	0.613	14.960	2.465
7	85.33	86.67	0.595	16.089	2.659
8	86.67	80.00	0.521	21.801	3.587
9	85.33	86.67	0.588	16.567	2.738
10	84.00	93.33	0.602	15.660	2.599
Mean	**84.80**	**89.33**	**0.584**	**17.209**	**2.809**

inhibitory activity for variables such as *nArNHR* (number of secondary aromatic amines) it is in correspondence with reported in other studies, where the secondary amine (–NH– group) plays an important role in the ligand binding to quadruplex substituents and with the water molecules present in the binding site [56]. As can be seen in the acridinic framework (see Fig. 1 and Table 4), all substituents in positions 3 and 6 are secondary amines, while in position 9, the substituents are secondary amines and carbonyl groups. For this reason, this molecular descriptor also explains the less telomerase inhibitory potency of acridones with regard to acridines derivatives.

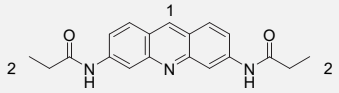
The *nCbH* descriptor (the number of unsubstituted benzene C(sp^2)) has a negative contribution to the telomerase inhibitory potency. The substitution on the acridinic skeleton is very important due to the electrostatic interactions produced among the lateral chains (generally present N atoms) with the G-tetrad. The tri-substituted acridinic derivatives could have a better stacking with the groups of the wide grooves located in the opposite sides of the G-tetrad face, increasing the stability of the complex [56]. Compounds with high values of this descriptor would have a weak interaction with the G-quartet. For instance, compound **90** is a mono-substituted acridine and it has a great number of unsubstituted C(sp^2) and the IC_{50} value is greater than 20 μM. If this compound is compared with the bi-substituted and tri-substituted acridines (compounds **17** and **49**), the IC_{50} values decreases in the same order as the number of unsubstituted benzene C(sp^2). This descriptor can also explain the variability of telomerase inhibitory concentration, by substitutions on the phenyl ring of the side chain at ninth position of the acridinic framework. For example, compounds **72**, **73**, **75** and **76** have lower IC_{50} values than the unsubstituted derivative (**77**) and the same occur for compounds **80**–**85** compared with compound **86**. Compounds from **1** to **17** have lower telomerase inhibitory potency compared with similar derivatives with a side chain at ninth position of the acridine ring.

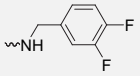
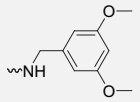
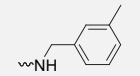
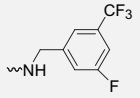
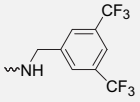
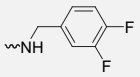
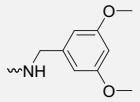
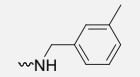
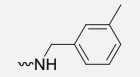
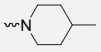
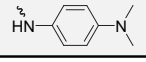
The positive sign of *nCar* indicates that acridinic derivatives with large number of aromatic carbons could have a strong binding to the G-quartet. A good example of this general contribution can be seen in compounds **23**–**43** (acridones), where with the addition of a carbonyl group the aromaticity of the ring is lost, affecting the hydrophobic interactions. In these conditions the nitrogen atom is also unable to support a positive charge so the electrostatic interactions could not play their effect. This also explain why acridones are less potent than their respective acridines [30]. The combination of both variables (*nCbH* and *nCar*) explains the differences in the telomerase inhibitory concentration of compounds with aromatic rings or cycloalcanes substituents at terminal position of C9 side chain. Compound **77** has six aromatic carbons more than compound **79**, but five of them are of unsubstituted benzene with a negative contribution to telomerase inhibitory activity.

As was previously explained, each one of the analyzed fragments has a direct effect in the prediction of telomerase inhibitory activity but only their inter-relation will bring a global interpretation of the molecule role in the interaction with G-quartet structure. In Table 3 can be appreciated a clear example of this descriptor inter-relation in three acridinic derivatives with different values of IC_{50}.

These results evidenced that the present model is able to identify those acridinic derivatives with good telomerase inhibition and could be used in the design of potential selective telomerase inhibitors with high affinity for the G-quartet, bringing also a good

Table 7
Results of the classification of an external test set of 10 acridines.



Mol	1	2	IC_{50} (μM)	Real	Class
1			0.03	+	+
2			0.35	+	+
3			1	–	+
4			0.24	+	+
5			0.86	+	+
6			0.88	+	+
7			0.44	+	+
8			0.23	+	+
9			0.39	+	+
10			0.36	+	+

"window of opportunity" for the design of library of novel acridinic derivatives.

The classification results for the training set are shown in Table 4.

The general performance of the model (goodness-of-fit) was assessed in terms of its Cooper Statistic (see Table 5).

The model classified correctly the 87.50% of compounds belongs to the more potent compounds (sensitivity) and with 82.85% those belongs to the less potent compounds group (specificity), for a global accuracy of 85.33%. Unclassified compounds were not obtained.

The most important criterion for the quality of the discriminant model is based on the statistics for the prediction set. The classification results for the prediction set are also shown in Table 4.

The results achieved shown an 87.50% of good classification for more potent compounds and 85.71% for compounds with IC_{50} values greater than 1 μM. The global classification was 86.66%.

The statistical results of the 10-fold cross-validation procedure are depicted in Table 6. The overall classification performed for the training and test sets were 84.80% and 89.33%, respectively. This model showed a high Matthews correlation coefficient (MCC) of 70.51. This statistical parameter quantifies the strength of the linear relation between the molecular descriptors and the classifications. The prediction results for the 15% full cross-validation test evidenced the quality (robustness, stability, and predictive power) of the obtained models.

In this study, an external data of 10 novel acridine derivatives was also evaluated [38]. The results for this validation process are summarized in Table 7. The equation showed a 90% of good classification.

The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability [57]. In this paper, we developed a reliable QSAR model for predicting the antitelomerase activity by G-quadruplex stabilization of acridinic compounds, considering TRAP assay. In order to describe the applicability domain of the previous QSAR model we used the leverage approach [59,60]. This approach provides a measure of the distance between the descriptor values for a chemical and the mean of descriptor values for all chemicals. So, we can plot leverage values *vs* standard residuals for each compound of the training set (see Fig. 7). From this plot (called William's plot), the applicability domain is established inside a squared area within ±3 standard deviations and a leverage threshold h^* of 0.226 ($h^* = 3p'/N$, where p' is the number of model parameters and N the number of chemicals). As can be seen in Fig. 7, the majority of compounds of the training set are inside this area. However, two acridinic compounds have a leverage greater than h^* but show standard deviation values within the limit, which implies that they are not to be considered outliers but instead are influential chemicals. These chemicals greatly influence the linear discriminate analysis; in fact, the GLDA model is forced near the observed value and their residuals (observed–predicted value) are small; that is, they are well-predicted. For future predictions, if a chemical belonging to the test set has a leverage value greater than h^*, this means that the prediction is the result of substantial extrapolation and therefore may not be reliable [58]. As can be noted in Fig. 7, all compounds used as external data set lie within this area, evidencing no extrapolation for their predicted values of telomerase inhibitory concentration.

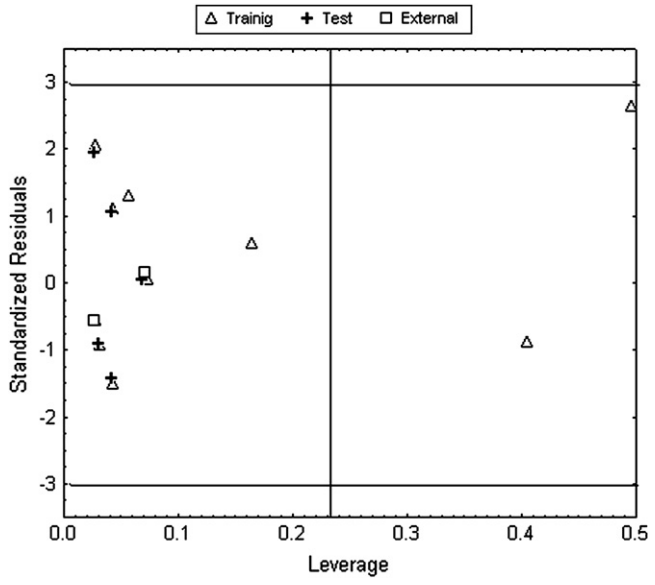


Fig. 7. Williams plot based on classification model (Eq. (3)).

4. Conclusions

The inhibition of the enzyme telomerase is a promissory target for the treatment of cancer. The results of this research are considered by these authors as the first QSAR study that allows a structural interpretation of the interactions of a novel family of compounds (acridines and acridones derivatives), with probed properties as telomerase inhibitors by stabilization of the quartet-G structure, with the telomeric structure. The results show how simple molecular descriptors, with substructural information, are able to differentiate the more potent acridines from the rest of compounds of this family with greater values of telomerase inhibitory activity. This kind of computational model permits to identify functional fragments and groups that can be used in the design of libraries of acridinic derivatives with better predicted properties to interact and stabilize the G-quartet. The present paper is a first step in the prediction of this important biological activity, where a good computational tool is necessary in order to increase the successful in the obtaining of novel and potent antitumoral compounds belongs to this heterocyclic family.

Acknowledgment

The authors would like to dedicate the present paper to the memory of Maykel Pérez González, one of the authors who died before finishing this work. This research was sponsored by the Cuban Higher Education Ministry (R&D project number 6.181-2006).

References

[1] E.H. Blackburn, Cell 106 (2001) 661–673.
[2] W.E. Wright, J.W. Shay, J. Am. Geriatr. Soc. 53 (2005) S292–S294.
[3] K. Collins, J.R. Mitchell, Oncogene 21 (2002) 564–579.
[4] M.A. Blasco, Embo. J. 24 (2005) 1095–1103.
[5] G.H. Lang, Y. Wang, N. Nomura, M. Matsumura, Mar. Biotechnol. (NY) 6 (2004) 347–354.
[6] L.A. Harrington, C.W. Greider, Nature 353 (1991) 451–454.
[7] A.A. Avilion, L.A. Harrington, C.W. Greider, Dev. Genet. 13 (1992) 80–86.
[8] K. Collins, C.W. Greider, Genes Dev. 7 (1993) 1364–1376.
[9] L. Maes, E. Lippens, J.P. Kalala, L. de Ridder, Cell Prolif. 38 (2005) 3–12.
[10] J.W. Shay, J. Cell Physiol. 173 (1997) 266–270.
[11] J.W. Shay, W.E. Wright, Carcinogenesis 26 (2005) 867–874.
[12] D.M. Shcherbakova, M.E. Zvereva, O.V. Shpanchenko, O.A. Dontsova, Mol. Biol. (Mosk) 40 (2006) 580–594.
[13] T. Mashimo, H. Sugiyama, Nucleic Acids Symp Ser (Oxf) (2007) 239–240.
[14] G. Lixia, Y. Fei, J. Jiajia, L. Jianhui, Biotechnol. Lett. 30 (2008) 47–53.
[15] L.H. Hurley, R.T. Wheelhouse, D. Sun, S.M. Kerwin, M. Salazar, O.Y. Fedoroff, F.X. Han, H. Han, E. Izbicka, D.D. Von Hoff, Pharmacol. Ther. 85 (2000) 141–158.
[16] F. Koeppel, J.F. Riou, A. Laoui, P. Mailliet, P.B. Arimondo, D. Labit, O. Petitgenet, C. Helene, J.L. Mergny, Nucleic Acids Res. 29 (2001) 1087–1096.

[17] L. Oganesian, M.E. Graham, P.J. Robinson, T.M. Bryan, Biochemistry 46 (2007) 11279–11290.
[18] W.J. Zhang, T.M. Ou, Y.J. Lu, Y.Y. Huang, W.B. Wu, Z.S. Huang, J.L. Zhou, K.Y. Wong, L.Q. Gu, Bioorg. Med. Chem. 15 (2007) 5493–5501.
[19] P. Phatak, J.C. Cookson, F. Dai, V. Smith, R.B. Gartenhaus, M.F. Stevens, A.M. Burger, Br. J. Cancer 96 (2007) 1223–1233.
[20] S. Neidle, M.A. Read, Biopolymers 56 (2000) 195–208.
[21] S.M. Gowan, R. Heald, M.F. Stevens, L.R. Kelland, Mol. Pharmacol. 60 (2001) 981–988.
[22] L. Rossetti, M. Franceschin, A. Bianco, G. Ortaggi, M. Savino, Bioorg. Med. Chem. Lett. 12 (2002) 2527–2533.
[23] D. Gomez, J.L. Mergny, J.F. Riou, Cancer Res. 62 (2002) 3365–3368.
[24] S.M. Gowan, J.R. Harrison, L. Patterson, M. Valenti, M.A. Read, S. Neidle, L.R. Kelland, Mol. Pharmacol. 61 (2002) 1154–1162.
[25] Y. Ishikawa, T. Yamashita, Y. Tomisugi, T. Uno, Nucleic Acids Res. Suppl. (2001) 107–108.
[26] M.Y. Kim, M. Gleason-Guzman, E. Izbicka, D. Nishioka, L.H. Hurley, Cancer Res. 63 (2003) 3247–3256.
[27] C.P. Li, J.H. Huang, A.C. Chang, Y.M. Hung, C.H. Lin, Y. Chao, S.D. Lee, J. Whang-Peng, T.S. Huang, Pharm. Res. 21 (2004) 93–100.
[28] T. Tauchi, K. Shin-Ya, G. Sashida, M. Sumi, A. Nakajima, T. Shimamoto, J.H. Ohyashiki, K. Ohyashiki, Oncogene 22 (2003) 5338–5347.
[29] B. Guyen, C.M. Schultes, P. Hazel, J. Mann, S. Neidle, Org. Biomol. Chem. 2 (2004) 981–988.
[30] R.J. Harrison, A.P. Reszka, S.M. Haider, B. Romagnoli, J. Morrell, M.A. Read, S.M. Gowan, C.M. Incles, L.R. Kelland, S. Neidle, Bioorg. Med. Chem. Lett. 14 (2004) 5845–5849.
[31] R. Kieltyka, J. Fakhoury, N. Moitessier, H.F. Sleiman, Chemistry (2007).
[32] L.G. Valerio Jr., K.B. Arvidson, R.F. Chanderbhan, J.F. Contrera, Toxicol. Appl. Pharmacol. (2007).
[33] B. Li, M.P. Lyle, G. Chen, J. Li, K. Hu, L. Tang, M.A. Alaoui-Jamali, J. Webster, Bioorg. Med. Chem. (2007).
[34] C.M. Incles, C.M. Schultes, H. Kempski, H. Koehler, L.R. Kelland, S. Neidle, Mol. Cancer Ther. 3 (2004) 1201–1206.
[35] R.J. Harrison, S.M. Gowan, L.R. Kelland, S. Neidle, Bioorg. Med. Chem. Lett. 9 (1999) 2463–2468.
[36] C.M. Schultes, B. Guyen, J. Cuesta, S. Neidle, Bioorg. Med. Chem. Lett. 14 (2004) 4347–4351.
[37] R.J. Harrison, J. Cuesta, G. Chessari, M.A. Read, S.K. Basra, A.P. Reszka, J. Morrell, S.M. Gowan, C.M. Incles, F.A. Tanious, W.D. Wilson, L.R. Kelland, S. Neidle, J. Med. Chem. 46 (2003) 4463–4476.
[38] C. Martins, M. Gunaratnam, J. Stuart, V. Makwana, O. Greciano, A.P. Reszka, L.R. Kelland, S. Neidle, Bioorg. Med. Chem. Lett. 17 (2007) 2293–2298.
[39] M.J. Moore, C.M. Schultes, J. Cuesta, F. Cuenca, M. Gunaratnam, F.A. Tanious, W.D. Wilson, S. Neidle, J. Med. Chem. 49 (2006) 582–599.
[40] P. Belmont, J. Bosson, T. Godet, M. Tiano, Anticancer Agents Med. Chem. 7 (2007) 139–169.
[41] N.W. Kim, M.A. Piatyszek, K.R. Prowse, C.B. Harley, M.D. West, P.L. Ho, G.M. Coviello, W.E. Wright, S.L. Weinrich, J.W. Shay, Science 266 (1994) 2011–2015.
[42] D. Shi, R.T. Wheelhouse, D. Sun, L. Hurley, J. Med. Chem. 44 (2001) 4509–4523.
[43] P. Mailliet, A. Laoui, J.F. Riou, G. Doerflinger, J.L. Mergny, F. Hamy, T. Caulfield, Triazine Derivatives and Their Applications as Antitelomerase Agents, vol. US6887873 B2, Aventis Pharma S.A., France, 2005.
[44] P. Mailliet, J.F. Riou, M. Alasia, G. Doerflinger, J.L. Mergny, A. Laoui, O. Petitgenet, E. Renou, Chemical Derivatives and Their Applications as Anti-telomerase agents, vol. US 6858608 B2, Aventis Pharma S.A., France, 2005, 26.
[45] I. StatSoft, STATISTICA, 2001.
[46] R. Todeschini, V. Consonni, A. Maui, M. Pavan, Dragon (2005).
[47] D. Weininger, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[48] D.J. Klein, M. Randić, D. Babić, B. Lučić, S. Nikolić, N. Trinajstić, Int. J. Quant. Chem. 63 (1991) 215–222.
[49] M. Randi&cacute, New J. Chem. 15 (1991) 517–525.
[50] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269–276.
[51] K. Rose, L.H. Hall, L.B. Kier, J. Chem. Inf. Comput. Sci. 42 (2002) 651–666.
[52] J. Stewart, L. Gill, Econometrics, second ed. Prentice Hall, London, 1998.
[53] F. Provost, T. Fawcett, Analysis and visualization of classifier performance comparison under class and cost distributions, in: A.A.f.A. Intelligence (Ed.), Third International Conference on Knowledge Discovery and Data Mining (KDD-97), American Association for Artificial Intelligence, 1997.
[54] H. Toivonen, A. Srinivasan, R.D. King, S. Kramer, C. Helma, Bioinformatics 19 (2003) 1183–1193.
[55] R. Benigni, A. Giuliani, Bioinformatics 19 (2003) 1194–1200.
[56] N.H. Campbell, G.N. Parkinson, A.P. Reszka, S. Neidle, J. Am. Chem. Soc. 130 (2008) 6722–6724.
[57] S. Raic-Malic, D. Svedruzic, T. Gazivoda, A. Marunovic, A. Hergold-Brundic, A. Nagl, J. Balzarini, E. De Clercq, M. Mintas, J. Med. Chem. 43 (2000) 4806–4811.
[58] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W. Tong, G. Veith, C. Yang, ATLA 33 (2005) 155–173.
[59] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Environ. Health Perspect. 111 (2003) 1361–1375.
[60] P. Gramatica, P. Pilutti, E. Papa, Atmos. Environ. 37 (2003) 3115–3124.