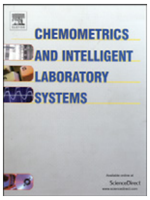
Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents

Laverne M.A. Mullen, Pablo R. Duchowicz *, Eduardo A. Castro

Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, C.C. 16, Suc.4, (1900) La Plata, Argentina

ARTICLE INFO

Article history:
Received 1 October 2010
Received in revised form 17 April 2011
Accepted 21 April 2011
Available online 1 May 2011

Keywords:
QSAR Theory
Apoptosis
Melanoma cells
Molecular descriptors
Anticancer drug

ABSTRACT

We establish predictive Quantitative Structure–Activity Relationships for exploring the relationship between the structures of a new emerging family of small triphenylmethyl-containing molecules and their cell anti-proliferative activities against the human melanoma cell lines SK-MEL-5 and UACC-62 in cell culture. In this way, we provide descriptive models of the increasing number of reported structure–activity studies, in order to assist the further optimization and development of anticancer agents posing the triphenylmethyl pharmacophore. We appropriately represent the molecular structures by means of descriptors derived from Dragon, Recon, and CORAL programs, while the predictive power of the QSAR are checked through the Cross Validation technique and also by leaving some compounds as part of an external test set.

1. Introduction

Apoptosis is a programmed cell death in multi-cellular organisms, considered as a series of biochemical events which is the organism's natural way of removing damaged or degraded cells. It is known that the process of apoptosis can be disrupted, and so a number of cancers are correlated to a lack of apoptosis which leads to cells multiplying to form tumors. The cell-division cycle involve four distinct phases: G1 (Gap1), S (DNA Synthesis), G2 (Gap2), and M (Mitosis), and therefore many anticancer drugs operate by interrupting cell proliferation during one of such phases.

It has been shown that various compounds carrying the triphenylmethyl (TPM) functional group arrest cells in the G1-phase of the cell cycle and induce apoptotic death of multiple cancer cell lines in culture [1]. An example of this is the antifungal agent clotrimazole, which inhibits the growth of cancer cells in culture and arrests cells in the G1-phase of the cell cycle. Also, S-trityl-L-cysteine (STLC) induces M-phase arrest with potent antitumor activity against the NCI-60 cell line panel [2] whereas the diphenyl derivative of STLC is inactive. Hergenrother et al. [1] have indicated that despite the molecular size and hydrophobic nature of the triphenylmethyl-containing compounds, these can be placed into at least four distinct categories, each with a different mechanism of action including causing cell cycle arrest, inhibiting tubulin polymerization, dissociating mitochondrial-bound hexokinase in cancer cells, and inhibiting calcium-dependent potassium ion channels.

The triphenylmethyl group is fundamental for the activity displayed by the triphenylmethylamides (TPMAs), which impede growth of human melanoma cells SK-MEL-5 and UACC-62 in the G1-phase of the cell cycle. It has been shown that diphenylmethyl-containing derivatives are essentially inactive against cultured cancer cells and so, show effectively no cell-cycle interruption [2]. The TPMAs also exhibit reduced toxicity to healthy bone marrow cells of humans, but display powerful anti-proliferative activity against lymphoma, leukemia and breast cancer cell lines in addition to melanoma [3]. It would appear that the substituent of the amide nitrogen of the TPMAs has little influence on the compound's anticancer properties as many different amide derivatives have death inducing properties in cancer cells in culture.

All these findings motivate us to establish Quantitative Structure–Activity Relationships (QSAR) analysis on a new identified set of small TPM compounds that could serve as a rational guide for designing further potent and selective therapeutic agents. Previous efforts have tried to identify a unified picture for the mechanism of cell death induction for some of these compounds, thus evaluating them in a battery of biological essays [1,3].

The very known basis of the QSAR Theory relies on the hypothesis that the biological activity manifested by a chemical compound completely results from its own molecular structure [4]. It is an approach of thermodynamical nature, in the sense that QSAR is only interested on the initial and final states (molecular structure and biological activity, respectively), but does not offer specific details on the usually complex mechanism/path of action involved. However, it is possible to get insight on the underlying mechanism by means of the QSAR-based predicted activities.

* Corresponding author. Tel.: +54 221 425 7430, +54 221 425 7291; fax: +54 221 425 4642.
E-mail addresses: pabloducho@gmail.com, prduchowicz@yahoo.com.ar (P.R. Duchowicz).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.04.011

The chemical structure is represented by numerical entities called molecular descriptors, which are used to describe different characteristics/attributes of a certain structure in order to yield information about the activity being studied [5–8]. Even though the relationship between the structure and the activity remains unknown for a given dataset, the QSAR technique has been based on statistically determined linear or nonlinear models relating the chemical behavior of compounds with descriptors, in order to find out useful parallelisms. It can be argued that the current state of art of the QSAR field involves addressing the following crucial factors: (a) the chosen set of descriptors employed, carrying suitable information of the molecular structure; (b) the modeling method employed; (c) the number of descriptors to be included in the model; (d) the composition of the training and test sets; and (e) the choice of the validation techniques to be applied.

A bibliographical search performed by our group reveals that no attempts have been carried out in past years to quantify the relationship between the anticancer activity and the structure of TPMs, as previous reported studies have searched for better activities by defining structure–activity relationships (SAR) for the functional groups present in the leading structure in terms of their responses. In this way, present work constitutes an interesting alternative to this trial-error based procedure, and proposes several models based on theory which would enable to acceptably describe the experimental tendencies of the anticancer activities.

2. Materials and methods

2.1. Experimental data set

The observed cell anti-proliferative activities of the triphenylmethyl-containing compounds inducing death in the human melanoma cell lines SK-MEL-5 and UACC-62 are extracted from recent studies [1,3]. These data are provided in the Supplementary Tables 1S and 2S for each chemical structure as log $_{10}IC_{50}[\mu M]$ values. We establish QSAR on two different sets of TPM derivatives: i) phosphonate and phosphonochloridate TPM, and ii) phenethylamine and D-phenylalanine TPM.

2.2. Geometry optimization and molecular descriptors calculation

The initial conformations of the compounds are drawn by means of the “Model Build” modulus of the HyperChem 6.03 program for Windows [9]. We pre-optimize the molecular structures with the Molecular Mechanics Force Field (MM+) procedure included in the HyperChem, and refine the resulting geometries by means of the Semiempirical Method PM3 from the Molecular Orbitals Theory using the Polak-Ribiere algorithm and a gradient norm limit of 0.01 kcal Å^{-1}.

Afterwards, we compute 1497 molecular descriptors using the Dragon program [10], including descriptors of all types such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, Atom-Centred Fragments, Empirical and Properties [7].

In addition, we calculate atomic charge density-based descriptors encoding electronic and structural information relevant to the chemistry of intermolecular interactions, by means of the Recon 5.5 software [11]. This sort of computed descriptors are not provided by Dragon software, while the robustness of Recon has previously been demonstrated elsewhere [12,13]. Recon is an algorithm for the reconstruction of molecular charge densities and charge density-based electronic properties of molecules, using atomic charge density fragments precomputed from *ab initio* wavefunctions. The method is based on the Quantum Theory of Atoms in Molecules [14]. A library of atomic charge density fragments has been built in a form that allows for the rapid retrieval of the fragments and molecular assembly. In present case, the smiles chemical notation is employed as input for the generation of 248 Transferable Atom Equivalent (TAE) descriptors, developed by Breneman and co-workers [15].

In this way, the total number of calculated structural descriptors for the molecular set under analysis results in 1745 variables.

2.3. Model development

The QSAR established in this work are obtained via two different modeling approaches with the purpose of comparing the consistency of our results: a) the search of molecular descriptors via multivariable linear regressions; and b) the calculation of flexible descriptors with the CORAL (CORrelation and Logic) program.

2.3.1. Linear descriptors search

In recent years theoretical and experimental researchers have focused an increasing attention on finding the most efficient tools for selecting molecular descriptors in QSAR studies. There is a great number of available feature selection methods to search the best structural descriptors from a pool of variables, and the Replacement Method (RM) [16,17], employed here, has been successfully applied elsewhere [18–22]. In brief, the RM is an efficient optimization tool which generates multi-parametric linear regression QSAR models on a training (calibration) molecular set by searching the set **D** of D descriptors for an optimal subset **d** of $d<<D$ ones with minimum model's standard deviation (S). The quality of the results achieved with this technique approaches that obtained by performing an exact (combinatorial) full search of molecular descriptors although, of course, requires much less computational work. Finally, RM results consider the Variance Inflation Factor (VIF), a method of detecting the severity of multicollinearity which represents a high degree of correlation (linear dependency) among several independent variables [23,24]. The VIF for a given descriptor can be easily calculated if we know the correlation coefficient between that descriptor and the remaining ones of the model (R_{ij}):

$$\mathrm{VIF} = \frac{1}{1-R_{ij}^2} \quad (1)$$

In practice, when VIF>10 then this would indicate that there exists significant multicollinearity in the chosen subset of descriptors.

2.3.2. The CORAL method

CHEMPREDICT/CORAL (CORrelation And Logic) version 1.4 [25] is a freeware for Windows. Each molecular structure must be represented by SMILES (Simplified Molecular Input Line Entry System) notation, calculated with ACD/ChemSketch software [26]. CORAL approach is based on the presence of certain SMILES attributes occurring in the molecule which can be associated to the activity of the molecule under evaluation [27–30]. We use as SMILES attributes the symbols representing the chemical elements, cycles, branching of molecular skeleton, charges, etc. The CORAL modeling process considers not only the presence of individual elements SMILES attributes (s_k), but also clusters of two (ss_k) and three (sss_k) elements. For example, SMILES = Clc1ccccc1 then s_k = (Cl, c, 1, c, c, c, c, c, 1); ss_k = (Clc, c1, cc, cc, cc, cc, cc, c1); sss_k = (Clc1, c1c, ccc, ccc, ccc, ccc, cc1).

The model is a one-variable correlation between the activity values and the flexible descriptor (DCW) that is defined as:

$$\mathrm{DCW(threshold)} = \alpha \sum_k \mathrm{CW}(s_k) + \beta \sum_k CW(ss_k) + \gamma \sum_k CW(sss_k) \quad (2)$$

where α, β, γ are 1 or 0, and CW is the correlation weight for the element/s of the SMILES. The threshold is the parameter to define rare (noise) SMILES attributes. The rare SMILES attributes can lead to overtraining: excellent correlation for the training set accompanied by poor correlation for the validation set. Thus they can bring “noise”. The threshold can be defined as 0, 1, 2, …N, with N being the number of compounds in training set. If threshold is defined 5, all SMILES attributes that take place in less than 5 SMILES notations of the training set will be classified as rare. In present study, numerical data for CW can be calculated by the Monte Carlo simulation by maximizing R parameter, the correlation coefficient between the activity values and the DCW descriptor defined in Eq. (2) for the training molecular set. The quality of the prediction is dependent on the selected options/parameters in the algorithm, such as the number of epochs used during the Monte Carlo optimization procedure, D_{start}, $d_{\text{precision}}$, dR_{weight}, dC_{weight}, threshold range and others, which should be correctly specified in order to calculate the *DCW* values. More specific details on the CORAL algorithm can be found in the recent literature [27–30].

2.3.3. Analysis of the happenstance of the model

Another simple way of proving that the structure–activity relationships established in this study do not result from happenstance involves checking their robustness by means of the so-called *y*-randomization [31]. This technique consists of scrambling the experimental property values in such a way that they do not correspond to the respective compounds. After analyzing 10000 cases of *y*-randomization for each developed QSAR, the smallest standard deviation value obtained using this procedure (S^{Rand}) turned out to be a poorer value when compared to the one found when considering the true calibration (S). Therefore, the correlations found are not fortuitous and result in real structure–activity relationships.

2.3.4. Model validation

Every QSAR research has to demonstrate the appropriate validation of the proposed models, in order to verify that these relationships behave predictive and are not only limited to work correlatively on the training set. The validation process is the most import step during the model design, and it is considered the basis of the QSAR hypothesis. If the validation fails, there is no way to employ the QSAR formalism to create new chemical information on the biological activity from the known one in the available molecular training set.

The theoretical validation practiced over each linear regression developed is based on the Leave-More-Out Cross Validation procedure ($l-n\%-o$) [32], $n\%$ being the percentile of molecules removed successively from the training set. Statistical parameters $R_{l-n\%-o}$ and $S_{l-n\%-o}$ measure the stability of the developed QSAR upon inclusion/exclusion of compounds, and according to the specialized literature, R^2_{loo} should be greater than 0.7 for obtaining a validated model [33]. However, from our own experience in establishing QSAR models, $R^2_{loo}<0.7$ could also lead to satisfactory models, as the Leave-More-Out technique provides the predictive power of the model by defect, in the sense that no-compound should be excluded and all training compounds should be present during the cross-validation evaluation. Therefore, we consider that R^2_{loo} should be greater than the value 0.7 but this is not an exclusive rule. The number of cases for random data removal analyzed here in this study is $l-n\%-o$ is 100,000.

We also apply a rigorous and more realistic validation that consists on omitting from the complete molecular set presented in Tables 1S and 2S some compounds which constitute the “test set,” denoted here as “test.” The main purpose of performing such as splitting is to assess whether the QSAR found have predictive capability for estimating the activity on the “fresh” test set compounds (never seen by the model). We select the TPM molecules composing the training and test series as a previous step to the model search, and this is done in such a way that both sets share similar qualitative structure–activity characteristics.

3. Results and discussion

We use the Matlab 7.0 program in all our calculations [34]. In every reported QSAR, N is the number of training set molecules, *range* is the experimental range of activities covered by the model, d the number of descriptors of the model, R is the correlation coefficient, S the model's standard deviation, F is the Fisher parameter, *res* the residual for a given molecule (difference between the experimental and predicted activity), outliers>x.S indicates the number of molecules predicted to have *res* greater than x times S, Corr^{max} represents the maximum intercorrelation coefficient between two given descriptors of the model, VIF is the variance inflation factor, *loo* and $l-20\%-o$ subindices belong to the Leave-One-Out and Leave-20%-Out Cross Validation results, respectively, and Rand supraindex stands for *y*-randomization. A brief description for each molecular descriptor appearing in the developed QSARs is provided in Table 1. All the models obey the semiempirical “Rule of Thumb,” stating that at least five or six data points should be present per descriptor [35]. Dispersion plots of residuals (residuals as function of predicted activities) for each QSAR are provided as Supplementary Figures. Correlation matrices together with the numerical values for each descriptor appearing in the established models are also provided as part of the supplementary material. The predicted activities for each QSAR are supplied in Tables 2 and 3.

3.1. QSAR on phosphonate and phosphonochloridate TPM derivatives

We apply the RM approach on the molecules **1**–**25** reported in reference [1] (Table 1S), which involve both active and inactive structures. The following are the most predictive structure–activity relationships that can be found on each cell line, obtained from the simultaneous analysis of 1745 descriptors provided by Dragon and Recon programs.

SK-MEL-5 cell line:

$$log_{10}\text{IC}_{50} = 2.514(\pm 0.3) + 11.299(\pm 2)\cdot \text{MATS1}v - 19.563(\pm 5)\cdot R1v^{+} \tag{3}$$

$N=19$, $range=0.342\text{-}2.000$, $d=2$, $N/d=9.5$, $R=0.884$, $S=0.29$, $F=28.6$, outliers$>2.S=0$, $\text{Corr}^{\text{max}}=0.325$, $R_{loo}=0.837$, $S_{loo}=0.35$, $R_{l\text{-}20\%\text{-}o}=0.748$, $S_{l\text{-}20\%\text{-}o}=0.47$, $S^{Rand}=0.32$.

Table 1
List of molecular descriptors involved in QSAR equations.

Descriptor	Category	Type	Brief description
*JGI*9	2D	Galvez Topological Charge Indices	Mean topological charge index of order 9
*MATS*1*v*		2D-Autocorrelations	Moran Autocorrelation-lag 1/weighted by atomic van der Waals volumes
*MATS*3*e*			Moran Autocorrelation-lag 3/weighted by atomic Sanderson electronegativities
*BELe*2		BCUT	Lowest eigenvalue no. 2 of Burden matrix/weighted by atomic Sanderson electronegativities
*BEHm*4			Highest eigenvalue no. 4 of Burden matrix/weighted by atomic masses
Ds	3D	WHIM	D total accessibility index/weighted by atomic electrotopological states
*P*1*p*			1st component shape directional WHIM index/weighted by atomic polarizabilities
*R*1*v*$^{+}$			R maximal autocorrelation of lag 1/weighted by atomic van der Waals volumes
*Mor*22*m*			signal 22/weighted by atomic masses

Table 2
Experimental and predicted log $_{10}IC_{50}[\mu M]$ activities for human melanoma cell line SK-MEL-5.

ID	Exp.	Eq. (3)	Eq. (5)	Eq. (7)
1	>2(2.176)[a]	2.400	2.176	–
2^	0.732	1.164	1.196	–
3^	>2(2.000)	2.234	2.385	–
4	1.155	1.182	1.155	–
5	1.090	1.021	0.978	–
6^	1.176	1.175	1.284	–
7	1.179	1.043	1.179	–
8	0.929	1.220	1.026	–
9	0.820	1.274	0.768	–
10	1.017	1.243	1.026	–
11	0.996	0.744	0.996	–
12	1.580	1.192	1.583	–
13	>2(2.176)	1.776	2.175	–
14	0.519	0.938	0.707	–
15^	0.924	1.372	1.001	–
16^	1.093	0.758	0.814	–
17	0.342	0.526	0.650	–
18	1.009	0.753	0.955	–
19	0.591	0.766	0.485	–
20^	1.013	0.868	0.791	–
21	1.182	1.006	1.182	–
22	>2(2.176)	1.757	1.900	–
23	>2(2.176)	2.261	2.360	–
24	0.732	0.550	0.543	–
25	0.732	0.925	0.732	–
26	0.763	–	–	0.925
27^	1.462	–	–	0.821
28^	1.146	–	–	1.078
29^	1.301	–	–	1.743
30	0.908	–	–	0.880
31	0.623	–	–	0.824
32	0.934	–	–	0.971
33	1.556	–	–	1.395
34	0.954	–	–	0.984
35	1.230	–	–	0.889
36	0.740	–	–	0.779
37	0.732	–	–	0.595
38^	0.756	–	–	0.866
39	0.785	–	–	0.891
40	1.380	–	–	1.152
41	1.690	–	–	1.771
42^	0.908	–	–	0.719
43	0.826	–	–	0.858
44	1.114	–	–	1.252
45	1.079	–	–	1.024
46	0.708	–	–	0.835
47^	0.398	–	–	0.728
48	0.763	–	–	0.764

^ test set compound.

[a] modeled to the value in parentheses.

Table 3
Experimental and predicted log $_{10}IC_{50}[\mu M]$ activities for human melanoma cell line UACC-62.

ID	Exp.	Eq. (4)	Eq. (6)	Eq. (8)
1	>2.000	2.121	1.995	–
2	0.519	0.909	0.519	–
3	>2.000	1.993	2.004	–
4^	1.140	1.099	0.800	–
5	1.204	1.156	1.048	–
6	1.130	0.943	1.298	–
7	1.430	1.426	1.430	–
8^	1.185	0.878	1.031	–
9	0.763	1.013	0.764	–
10	1.083	1.041	1.083	–
11	1.107	1.017	1.116	–
12	1.681	1.193	1.674	–
13	>2.000	1.774	2.000	–
14	0.633	1.130	0.615	–
15	0.875	0.735	0.894	–
16	1.146	0.605	0.924	–
17	0.398	0.641	0.801	–
18^	1.100	0.886	1.051	–
19	0.699	0.804	0.677	–
20	1.100	1.110	0.927	–
21	0.914	1.144	0.914	–
22	>2.000	1.926	2.001	–
23^	>2.000	2.428	2.478	–
24^	0.519	1.161	0.492	–
25^	0.763	1.564	1.277	–
26	0.919	–	–	0.880
27	1.114	–	–	1.068
28	1.301	–	–	1.483
29^	2.000	–	–	1.707
30^	1.114	–	–	1.246
31	0.322	–	–	0.436
32	1.672	–	–	1.461
33	1.851	–	–	1.608
34	1.114	–	–	1.125
35	1.699	–	–	1.659
36	1.699	–	–	1.828
37	1.415	–	–	1.461
38	1.146	–	–	1.178
39	1.431	–	–	1.512
40	2.000	–	–	1.730
41	2.000	–	–	2.186
42	1.322	–	–	1.357
43^	1.301	–	–	1.574
44	1.431	–	–	1.752
45	1.322	–	–	1.316
46	2.000	–	–	1.765
47	1.602	–	–	1.558
48^	1.653	–	–	1.664

^ test set compound.

$N_{\text{test}}=6$, $R_{\text{test}}=0.796$, $S_{\text{test}}=0.38$.

UACC-62 cell line:

$$\begin{aligned} log_{10}\mathrm{IC}_{50} = {} & 3.608(\pm 0.7) + 67.597(\pm 23)\cdot \mathrm{JGI9} \\ & + 9.546(\pm 2)\cdot \mathrm{MATS1}v - 5.670(\pm 2)\cdot P1p \end{aligned} \tag{4}$$

$N=19$, $range=0.398\text{-}2.000$, $d=3$, $N/d=6.33$, $R=0.864$, $S=0.29$, $F=14.7$, outliers$>2.S=0$, $Corr^{\max}=0.407$, $R_{loo}=0.813$, $S_{loo}=0.34$, $R_{l\text{-}20\%\text{-}o}=0.680$, $S_{l\text{-}20\%\text{-}o}=0.45$, $S^{\text{Rand}}=0.31$.

$N_{\text{test}}=6$, $R_{\text{test}}=0.645$, $S_{\text{test}}=0.59$.

According to the statistical parameters of calibration and cross-validation, both QSAR are able to predict the experimental activities reasonably well. For the case of Eq. (4), it requires an additional descriptor for improving performance, especially the Leave-One-Out statistics. These two linear regressions achieve acceptable predictions on six test set compounds (see Tables 2 and 3, test data denoted with ^). Fig. 1a and b plots the predicted log $_{10}IC_{50}$ as function of experimental values for Eqs. (3) and(4), respectively, showing that there is a tendency of the data to have a straight line trend. In addition, Fig. 1Sa and b reveal that the residuals obey a random pattern around the zero line, indicating the absence of non-modeled factors.

The correlations matrices for Eqs. (3)and(4) in Table 3S also include the VIF parameters for each chosen descriptor; according to these results, descriptors participating in these models are non-collinear and include non-redundant structural information content. As it is appreciated, the models require 2D and 3D aspects from the molecular structure in order to establish the proper links with the biological activities. The *MATS1v* descriptor is the Moran 2D Autocorrelation-lag 1/weighted by atomic van der Waals volumes [36]; $R1v^+$ is the R maximal autocorrelation of lag 1/weighted by atomic van der Waals volumes, a 3D-GETAWAY descriptor [37]; the 2D Galvez Topological Charge index descriptor *JGI9* corresponds to the mean topological charge index of order 9 [38]; the 3D descriptor *P1p* is the 1st component shape directional WHIM index/weighted by atomic polarizabilities [39].

In present data set, it is possible to improve the performance of Eqs. (3)and(4) by using flexible descriptor definitions calculated with

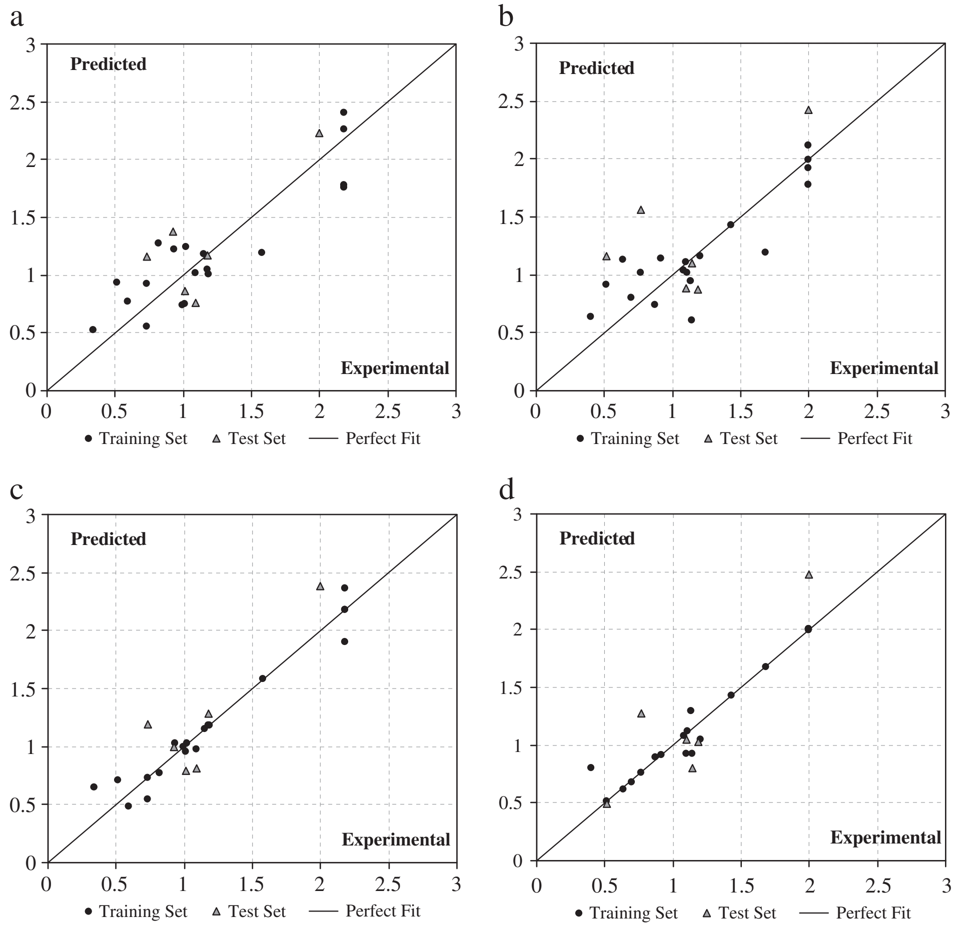
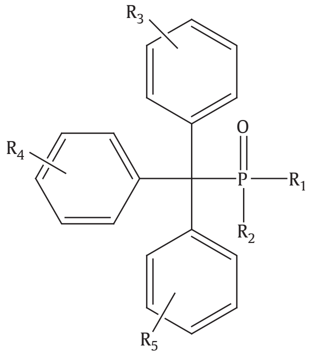


Fig. 1. QSAR on phosphonate and phosphonochloridate TPM derivatives. (a) Predicted log $_{10}IC_{50}$[μM] for human melanoma cell line SK-MEL-5 according to Eq. (3) as function of experimental values ($N=19$); (b) predicted log $_{10}IC_{50}$[μM] for UACC-62 according to Eq. (4) as function of experimental values ($N=19$); (c) Predicted log $_{10}IC_{50}$[μM] for SK-MEL-5 according to Eq. (5) as function of experimental values ($N=19$); (d) predicted log $_{10}IC_{50}$[μM] for UACC-62 according to Eq. (6) as function of experimental values ($N=19$).

Table 4

Predicted anticancer activities for melanoma cell lines according to Eqs. (5)and(6).



No.	R1	R2	R3	R4	R5	SK-MEL-5 (Eq. (5))	UACC-62 (Eq. (6))
1	n-pentyl		Hydrogen			0.510	0.497
2	3,3-dihydroxy propanyloxy		Hydrogen			2.472	2.612
3	2-hydroxy propanyloxy		Hydrogen			2.939	2.419
4	1,2-dihydroxy propanyloxy		Hydrogen			4.287	3.290
5	5-isoheptenyloxy		Hydrogen			0.510	0.549
6	Methoxy	Propyloxy	4-methoxy			0.509	0.669
7	Methoxy	Ethoxy	4-methoxy		4-n-butyloxy	0.533	0.660
8	Methoxy	Ethoxy	4-hydroxy		4-ethoxy	2.041	2.128
9	Ethoxy		4-hydroxy		4-(2-hydroxy) ethoxy	2.085	2.228
10	Chloro	2-propynyloxy	4-methyl	3-methoxy	hydrogen	0.389	0.715

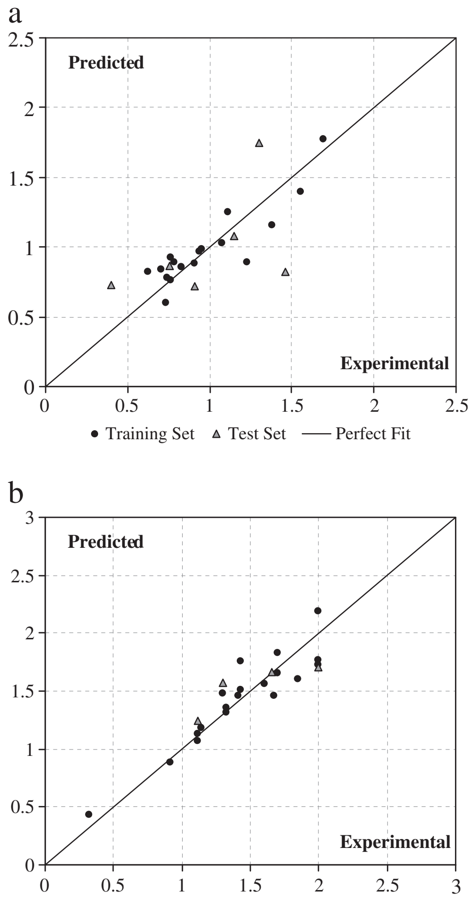


Fig. 2. QSAR on phenethylamine and D-phenylalanine TPM derivatives. (a) Predicted log $_{10}IC_{50}$[μM] for SK-MEL-5 according to Eq. (7) as function of experimental values ($N=17$); (b) Predicted log $_{10}IC_{50}$[μM] for UACC-62 according to Eq. (8) as function of experimental values ($N=19$).

the CORAL program. We run a Monte Carlo simulation for obtaining the correlation weights of Eq. (2), leading to the following QSAR.

SK-MEL-5 cell line:

$$log_{10}IC_{50} = -6.214(\pm 0.4) + 0.107(\pm 0.006)\cdot DCW_1 \qquad (5)$$

$N=19$, range $=0.342$–2.000, $d=1$, $N/d=19$, $R=0.975$, $S=0.14$, $F=323.5$, outliers$>2.5S=0$, Corr$^{max}=0$, $R_{loo}=0.966$, $S_{loo}=0.16$, $R_{l\text{-}20\%\text{-}o}=0.952$, $S_{l\text{-}20\%\text{-}o}=0.21$, $S^{Rand}=0.34$.

$N_{test}=6$, $R_{test}=0.867$, $S_{test}=0.36$.

UACC-62 cell line:

$$log_{10}IC_{50} = -4.743(\pm 0.4) + 0.093(\pm 0.006)\cdot DCW_2 \qquad (6)$$

$N=19$, range $=0.398$–2.000, $d=1$, $N/d=19$, $R=0.970$, $S=0.13$, $F=267.5$, outliers$>3.S=1$, Corr$^{max}=0$, $R_{loo}=0.964$, $S_{loo}=0.14$, $R_{l\text{-}20\%\text{-}o}=0.950$, $S_{l\text{-}20\%\text{-}o}=0.16$, $S^{Rand}=0.30$.

$N_{test}=6$, $R_{test}=0.872$, $S_{test}=0.40$.

The numerical parameters used in the CORAL calculation are: number of epochs: 33, number of probes: 3, range of threshold values: 0–2, $D_{start}=0.5$, $d_{precision}=0.01$, $dR_{weight}=0$, $dC_{weight}=0$, threshold range = 0-2, and $\alpha=\beta=0$ (refer to Eq. (2)). For the case of DCW_1, preferable threshold value is 2, while for DCW_2 is 0. Fig. 1c and d plots the predicted activities as function of the experimental data.

Now, one may adopt the most predictive QSAR models from Eqs. (5)and(6) to predict unknown TPM compounds and classify them as being either active (low predicted log $_{10}IC_{50}$) or inactive (high predicted log $_{10}IC_{50}$). With the purpose of deriving new chemical information from the designed QSAR, we draw and optimize 107 phosphonate and phosphonochloridate TPM derivatives which do not have experimentally assigned anticancer activities, and which have structural characteristics resembling the training molecular set. We include in Table 4 the top-five most active and most inactive predicted compounds. These QSAR based findings are in line with previous reported structure–activity observations [1], in the sense that an increase in activity is seen in the cell lines with compounds that have methyl ether substitutions on the aromatic rings (**6**, **7**, **10**), and phosphonochloridate derivatives are quite effective in killing melanoma cell lines (**10**). Compounds that increase aqueous solubility are not very potent (**2**, **3**, **4**, **8**, **9**). The most active compound in both cell lines is **1**.

3.2. QSAR on Phenethylamine and D-Phenylalanine TPM Derivatives

As a next analysis we establish QSAR on the anticancer activities reported in reference [3], compounds **26**–**48** (Table 2S), also involving active and inactive structures.

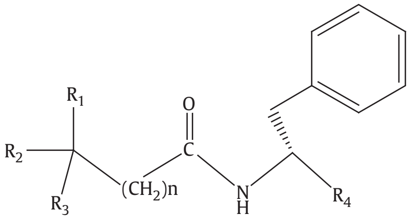
SK-MEL-5 cell line:

$$log_{10}IC_{50} = -67.026(\pm 9.671) + 18.627(\pm 3)\cdot BEHm4 \\ -2.821(\pm 0.9)\cdot Ds \qquad (7)$$

$N=17$, range $=0.623$–1.690, $d=2$, $N/d=8.5$, $R=0.884$, $S=0.16$, $F=25.1$, outliers$>2.5S=0$, Corr$^{max}=0.336$, $R_{loo}=0.836$, $S_{loo}=0.19$, $R_{l\text{-}20\%\text{-}o}=0.823$, $S_{l\text{-}20\%\text{-}o}=0.29$, $S^{Rand}=0.16$.

$N_{test}=6$, $R_{test}=0.493$, $S_{test}=0.44$.

Table 5
Predicted anticancer activities for melanoma cell lines according to Eqs. (7)and (8).



No.	N	R₁	R₂	R₃	R₄	SK-MEL-5 (Eq. (7))	UACC-62 (Eq. (8))
11	0	Phenyl	phenyl	4-propyloxy phenyl	hydrogen	0.804	1.261
12	2	Phenyl	phenyl	4-methoxy phenyl	hydrogen	0.843	1.609
13	0	Phenyl	phenyl	4-hydroxybenzyl	hydrogen	0.894	0.841
14	0	Phenyl	phenyl	4-propyloxy phenyl	methyl carboxylate	0.925	1.611
15	0	Phenyl	4-methyloxy phenyl	4-ethyloxy phenyl	hydrogen	1.004	1.568

UACC-62 cell line:

$$\begin{aligned} log_{10}IC_{50} &= -344.569(\pm 43) + 185.291(\pm 23)\cdot BELe2 \\ &\quad -6.603(\pm 0.8)\cdot MATS3e - 1.053(\pm 0.3)\cdot Mor22m \end{aligned} \tag{8}$$

$N=19$, $range=0.322\text{-}2.000$, $d=3$, $N/d=6.33$, $R=0.927$, $S=0.17$, $F=30.4$, outliers$>2.S=0$, $\text{Corr}^{\max}=0.435$, $R_{loo}=0.881$, $S_{loo}=0.22$, $R_{l\text{-}20\%\text{-}o}=0.761$, $S_{l\text{-}20\%\text{-}o}=0.34$, $S^{Rand}=0.24$.

$N_{\text{test}}=4$, $R_{\text{test}}=0.850$, $S_{\text{test}}=0.30$.

Here again, both 2D and 3D types of descriptors are needed to achieve predictive QSAR equations (refer to Fig. 2). The 2D-*BEHm*4 and *BELe*2 descriptors are the BCUT [40] highest eigenvalue no. 4 of Burden matrix/weighted by atomic masses, and the Lowest eigenvalue no. 2 of Burden matrix/weighted by atomic Sanderson electronegativities, respectively; the *Ds* 3D-WHIM [39] descriptor is the *D* total accessibility index/weighted by atomic electrotopological states, while *Mor*22*m* is WHIM signal 22/weighted by atomic masses; *MATS*3*e* descriptor is the Moran 2D-Autocorrelation-lag 3/weighted by atomic Sanderson electronegativities [36]. The dispersion plot of residuals for Eqs. (7)and(8) are included in Fig. 2Sa and b. After performing the analysis of present data set with CORAL flexible descriptors, it is not possible to improve the predictive capability of Eqs. (7)and(8), so we decide to discard the application of this approach on the actual set of molecules.

We report in Table 5 the predictions for the top-five most active anticancer unknown molecules, as obtained with Eqs. (7)and(8). According to this table, the most active TPM compound for both cell lines is **13**.

4. Conclusions

This work establishes Quantitative Structure–Activity Relationships that may prove useful for guiding the rational search of new therapeutic agents in cancer disease, in present case for molecules carrying the triphenylmethyl motif, which have been proven to have considerable potential as anticancer agents although the mode of action of this new molecular class still remains unknown. We consider two different approaches for establishing the QSAR models: a) selection of molecular descriptors via multivariable linear regressions; and b) the employment of flexible descriptors calculated with the CORAL program. The results achieved by the two methodologies lead to consistent predictions.

Supplementary materials related to this article can be found online at doi:10.1016/j.chemolab.2011.04.011.

Acknowledgements

L.M.A.M. would like to thank the IAESTE agency for a fellowship that made this work possible. P.R.D. and E.A.C. thank the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) and the Universidad Nacional de La Plata for supporting this work.

References

[1] R. Palchaudhuri, V. Nesterenko, P.J. Hergenrother, J. Am. Chem. Soc. 130 (2008) 10274.
[2] R.S. Dothager, K.S. Putt, B.J. Allen, B.J. Leslie, V. Nesterenko, P.J. Hergenrother, J. Am. Chem. Soc. 127 (2005) 8686.
[3] R. Palchaudhuri, P.J. Hergenrother, Bioorg. Med. Chem. Lett. 18 (2008) 5888.
[4] C. Hansch, A. Leo, Exploring QSAR. Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, D. C., 1995
[5] A.R. Katritzky, V.S. Lobanov, M. Karelson, Chem. Soc. Rev. 24 (1995) 279.
[6] M.V.E. Diudea, QSPR/QSAR Studies by Molecular Descriptors, Nova Science Publishers, New York, 2001.
[7] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, Weinheim, 2009.
[8] N. Trinajstic, Chemical Graph Theory, CRC Press, Boca Raton (FL), 1992.
[9] Hyperchem 6.03, Hypercube, Inc, Gainesville, 2007. http://www.hyper.com.
[10] Dragon, Milano Chemometrics and QSAR Research Group, http://michem.disat.unimib.it/chm.
[11] Recon Version 5.5, Rensselaer Polytechnic Institute, Troy, New York, USA, 2002. http://www.drugmining.com.
[12] B.K. Lavine, C.E. Davidson, C. Breneman, W. Katt, J. Chem. Inf. Comput. Sci. 43 (2003) 1890.
[13] A. Worachartcheewan, C. Nantasenamat, T. Naenna, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Eur. J. Med. Chem. 44 (2009) 1664.
[14] R.F.W. Bader, Atoms in Molecules—A Quantum Theory, Clarendon Press, Oxford, 1990.
[15] C.M. Breneman, L.W. Weber, in: G.A. Jeffrey, J.F. Piniella (Eds.), The Application of Charge Density Research to Chemistry and Drug Design, Plenum, New York, 1991.
[16] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. González, Chem. Phys. Lett. 412 (2005) 376.
[17] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Chemom. Intell. Lab. Syst. 92 (2008) 138.
[18] P.R. Duchowicz, M. Fernández, J. Caballero, E.A. Castro, F.M. Fernández, Bioorg. Med. Chem. 14 (2006) 5876.
[19] P.R. Duchowicz, M.G. Vitale, E.A. Castro, M. Fernandez, J. Caballero, Bioorg. Med. Chem. 15 (2007) 2680.
[20] P.R. Duchowicz, A. Talevi, L.E. Bruno-Blanch, E.A. Castro, Bioorg. Med. Chem. 16 (2008) 7944.
[21] P.R. Duchowicz, M. Goodarzi, M.A. Ocsachoque, G.P. Romanelli, E.V. Ortiz, J.C. Autino, D.O. Bennardi, D. Ruiz, E.A. Castro, Sci. Total Environ. 408 (2009) 277.
[22] M. Goodarzi, P.R. Duchowicz, C.H. Wu, F.M. Fernández, E.A. Castro, J. Chem. Inf. Model. 49 (2009) 1475.
[23] Multicollinearity.doc © 2002 Jeeshim and KUCC625. (2003-05-09).
[24] J.D. Curto, J.C. Pinto, Int. Stat. Rev. 75 (2007) 114.
[25] Coral 1.4, http://www.insilico.eu/coral.
[26] ACD/ChemSketch Freeware version 12.01, Advanced Chemistry Development, Inc, Toronto, ON, Canada, 2009. www.acdlabs.com.
[27] A.A. Toropov, E. Benfenati, Curr. Drug Discov. Technol. 4 (2007) 77.
[28] A.A. Toropov, E. Benfenati, Bioorg. Med. Chem. 26 (2008) 4801.
[29] A.A. Toropov, A.P. Toropova, E. Benfenati, Chem. Biol. Drug Des. 73 (2009) 515.
[30] A.A. Toropov, A.P. Toropova, E. Benfenati, D. Leszczynska, J. Leszczynski, Eur. J. Med. Chem. 45 (2010) 1387.
[31] S. Wold, L. Eriksson, Statistical validation of QSAR results, in: H. van de Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, VCH, Weinheim, 1995, p. 309.
[32] D.M. Hawkins, S.C. Basak, D. Mills, J. Chem. Inf. Model. 43 (2003) 579.
[33] A. Golbraikh, A. Tropsha, J. Mol. Graphics Modell. 20 (2002) 269.
[34] Matlab 7.0, The MathWorks, Inc., 2008. http://www.mathworks.com.
[35] M.S. Tute, in: N.J. Harter, A.B. Simmord (Eds.), History and Objectives of Quantitative Drug Design in Advances in Drug Research, Academic Press, London, 1971, p. 1.
[36] G. Moreau, P. Broto, Nouv. J. Chim. 4 (1980) 359.
[37] V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, J. Chem. Inf. Model. 42 (2002) 693.
[38] J. Galvez, J.V. de Julian-Ortiz, R. Garcia-Domenech, J. Mol. Graphics Modell. 20 (2001) 84.
[39] P. Gramatica, M. Corradi, V. Consonni, Chemosphere 41 (2000) 763.
[40] F.R. Burden, J. Chem. Inf. Model. 29 (1989) 225.