Journal of Molecular Graphics and Modelling 48 (2014) 70–79

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Multivariate SAR and QSAR of cucurbitacin derivatives as cytotoxic compounds in a human lung adenocarcinoma cell line

Karen L. Lang[a], Izabella T. Silva[b], Vanessa R. Machado[b], Lara A. Zimmermann[b], Miguel S.B. Caro[c], Cláudia M.O. Simões[b], Eloir P. Schenkel[b], Fernando J. Durán[d], Lílian S.C. Bernardes[b], Eduardo B. de Melo[e,*]

[a] Department of Pharmacy, Federal University of Juiz de Fora (UFJF), Campus Governador Valadares, MG, Brazil
[b] Department of Pharmaceutical Sciences, Federal University of Santa Catarina (UFSC), Florianópolis, SC, Brazil
[c] Department of Chemistry, Federal University of Santa Catarina (UFSC), Florianópolis, SC, Brazil
[d] UMYMFOR-CONICET, Department of Organic Chemistry, University of Buenos Aires, Buenos Aires, Argentina
[e] Theoretical Medicinal and Environmental Chemistry Laboratory (LQMAT), Department of Pharmacy, Western Paraná State University (Unioeste), Cascavel, PR, Brazil

ARTICLE INFO

Article history:
Accepted 3 December 2013
Available online 12 December 2013

Keywords:
Cucurbitacins
Lung cancer
QSAR
PLS
PLS-DA
OPS

ABSTRACT

This article describes structure–activity relationship (SAR/QSAR) studies on the cytotoxic activity in a human lung adenocarcinoma cell line (A549) of 43 cucurbitacin derivatives. Modeling was performed using the methods partial least squares with discriminant analysis (PLS-DA) and PLS. For both studies, the variables were selected using the ordered predictor selection (OPS) algorithm. The SAR study demonstrated that the presence or absence of cytotoxic activity of the cucurbitacins could be described using information derived from their chemical structures. The QSAR study displayed suitable internal and external predictivity, and the selected descriptors indicated that the observed activity might be related to electrophilic attack on cellular structures or genetic material. This study provides improves the understanding of the cytotoxic activity of cucurbitacins and could be used to propose new cytotoxic agents.

1. Introduction

In recent years, more than 70 antineoplastic drugs have been introduced in therapeutics. Nevertheless, cancer treatment remains a challenge, mainly due to the wide variety of cancers and the similarity between normal and tumor cells. Currently, research on new drugs has been based on the evaluation of chemical structures of natural and synthetic compounds and the development of similar derivatives. These new chemical entities have been investigated for the treatment of several disorders, primarily, cancer and infectious diseases. Reviews have described the importance of Natural Products (NPs) as a source of potential chemotherapeutic agents, and it is interesting to note that more than 25% of the anticancer drugs approved in the last 30 years are semisynthetic molecules directly derived from NPs [1,2].

In this context, the cucurbitacins is a class of highly oxygenated triterpenic compounds derived from a cucurbitane skeleton (Fig. 1). They are predominantly found in different species of the Cucurbitaceae family and have been distinguished according to features in ring A, side chain modifications, and stereochemical characteristics. These NP scaffolds are known for their bitter taste and pharmacological properties, including their purgative, anti-inflammatory, and anti-fertility activities, as well as their cytotoxicity and anti-cancer activity [3–7].

Cucurbitacins induce both morphological and physiological changes in tumor cells, causing drastic changes in cell shape, such as rounding, swelling, pinocytic blebbing, submembranous inclusions, and blisters [7]. Some of these changes can be explained by dysregulation of cytoskeletal homeostasis [8–10]. Many studies have shown that cucurbitacins induce cell cycle arrest, particularly in the G2/M phase [11–14], and in the S phase [14]. The cell cycle arrest in the G2/M phase occurs immediately after exposure to cucurbitacins and results in apoptosis of tumor cells [11]. Recent works reported that cucurbitacins B, D, E, and I and related compounds are active against different tumor cell lines and that some of them are STAT3 (Signal Transducers and Activators of Transcription-3) inhibitors, and affect other signaling pathways that are important for cancer cell proliferation and survival, such as the mitogen-activated protein kinase (MAPK) pathway [7,13,15–19].

* Corresponding author at: Department of Pharmacy, Unioeste, 2069 Universitária St, 85819110 Cascavel, PR, Brazil. Tel.: +55 45 3220 3256.
E-mail address: eduardo.b.de.melo@gmail.com (E.B. de Melo).

1093-3263/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2013.12.004

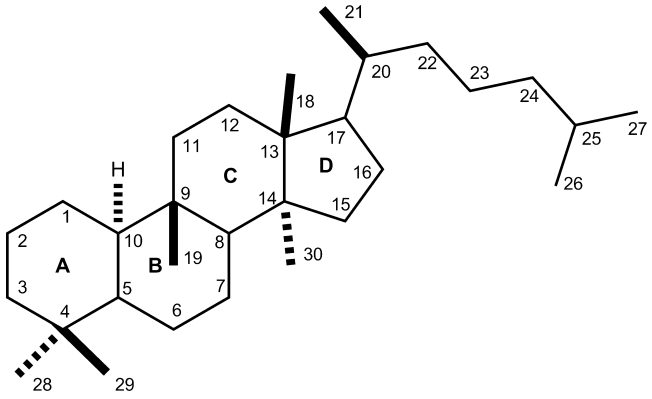

Fig. 1. The curcubitan (19-(10 → 9β)-abeo-10α-lanost-5-ene) skeleton, the general structure of cucurbitacins used in this study.

Structure–activity relationship (SAR) studies are helpful tools of Computer Aided Drug Design (CADD) [20,21], including the development of anticancer agents [22–24] and in environmental sciences [25] that have been used to describe how a given biological activity or property may vary as a function of molecular descriptors derived from the chemical structure of a set of molecules. However, despite the great potential of cucurbitacins, there are only two QSAR (quantitative structure-activity relationship) reports in the literature concerning this class of compounds [26,27]. Thus, considering the potential of cucurbitacins as new lead compounds, the main goal of this investigation was to provide a SAR and QSAR studies of a selected data set of 43 cucurbitacin derivatives assayed to their in vitro cytotoxic activity in human lung adenocarcinoma epithelial cell line (A549).

2. Materials and methods

2.1. Data set

The set of 40 cucurbitacins of interest were previously published by Lang et al. [3,28], Machado [29] and Farias et al. [30]. The basic structure of the dataset is available in Fig. 1 and all structures are available in the Fig. 2. The cytotoxic activity against lung adenocarcinoma (A549 cell line) was evaluated using the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) colorimetric assay [31]. The IC_{50} value (concentration that inhibited cell proliferation by 50% when compared to untreated controls) of each compound are available in Table 1. Additionally, data about three new derivatives (**14**, **17** and **27**) recently synthesized and assayed are also available in the Supplementary Material.

Exact values of their respective IC_{50} (range: 0.04×10^{-6} to 174.37×10^{-6} M; $-\log IC_{50}$: 7.398–3.759) were available obtained for twenty-three compounds and these were used in the QSAR step. Therefore, the complete data set was used in a qualitative study using partial least squares with discriminant analysis (PLS-DA) [32], and each was designated as active (class 1, IC_{50} exactly determined) or inactive (class 2, inactive). The active compounds were used in a QSAR study employing the PLS regression [33].

2.2. Molecular descriptors

Three-dimensional structures were built using ChemOffice [34] and optimized in the MM2 force field. The output files were converted into input files for Gaussian 09 [35] with Open Babel [36]. Thus, calculations at the AM1 theory level, followed by Hartree Fock level (HF/6-31G(d,p)) and the Density Functional Theory level (B3LYP/6-311G(d,p)) were performed. The B3LYP functional was chosen to obtain the electronic descriptors and the final geometries because provides quite satisfactory results for the analysis of these molecular characteristics [21].

Table 1
Inhibitory effects of cucurbitacins derivatives on proliferation of A549 cells.

Compound	IC_{50} (μM)	$-\log IC_{50}$
1	12.09	4.918
2	0.04	7.398
3	1.88	5.726
4	21.61	4.665
5	4.93	5.307
6	4.10	5.387
7	NM[a]	Inactive
8	NM	Inactive
9	26.49	4.577
10	72.52	4.140
11	NM	Inactive
12	NM	Inactive
13	NM	Inactive
14	112.4	3.949
15	NM	Inactive
16	174.37	3.759
17	NM	Inactive
18	2.64	5.578
19	NM	Inactive
20	42.60	4.371
21	110.34	3.957
22	55.19	4.258
23	NM	Inactive
24	NM	Inactive
25	6.90	5.161
26	11.47	4.940
27	NM	Inactive
28	NM	Inactive
29	66.74	4.176
30	77.74	4.109
31	NM	Inactive
32	0.12	6.921
33	29.09	4.536
34	0.12	6.921
35	NM	Inactive
36	NM	Inactive
37	NM	Inactive
38	11.52	4.939
39	NM	Inactive
40	NM	Inactive
41	48.55	4.314
42	53.60	4.271
43	NM	Inactive

[a] The cytotoxic evaluations of these compounds were not measurable (NM).

The partial charges (Mulliken and Natural Bond Orders) of the basic structure (Fig. 1), the orbital molecular energies (E_{HOMO-1}, E_{HOMO}, E_{LUMO} and E_{LUMO+1}), the dipole moment (D) and respective components (D_x, D_y and D_z), and the total energy (E_T) were obtained using Gauss View 5 [37]. Furthermore, the following reactivity descriptors were derived from the orbital molecular energies using the equations available in Todeschini and Consonni [38]: HOMO–LUMO energy gap (GAP), hardness (η), softness (S), ionization potential (IP), activation energy index (AEI), electronic affinity (EA), HOMO/LUMO energy fraction ($f_{(H/L)}$), molecular electronegativity (χ), electrophilicity index in the ground state (ω_{gs}), and electrophilicity index (ω).

The optimized geometries were converted in MOL2 files with Open Babel and utilized in Dragon 6 [39] to obtain 29 classes of descriptors (see http://www.talete.mi.it/help/dragon_help/index.html for more information), divided into 0D, 1D, 2D and 3D descriptors. A first step of variable reduction was performed using some options available in Dragon 6. Thus, the following descriptors were excluded: (i) the descriptors with constant values; (ii) the descriptors with constant and near-constant variables; (iii) the descriptors with a standard deviation of less than 0.0001; (iv) the

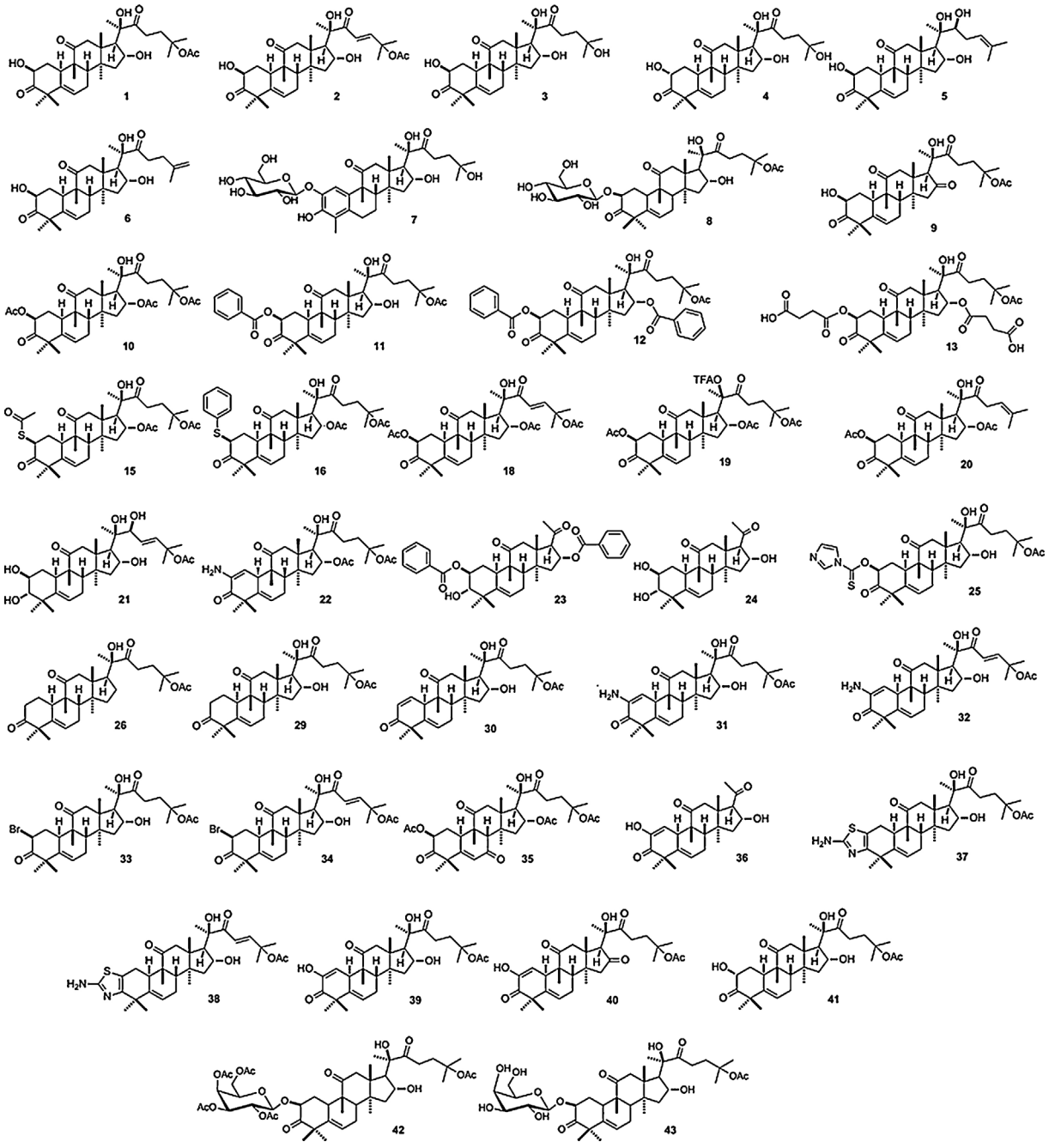

Fig. 2. Data set of forty-three curcubitacin derivatives.

descriptors with at least one missing value; and (v) the descriptors with a pair correlation larger than or equal to 0.95. Thus, a total of 1004 molecular descriptors derived with Dragon 6 were used for the qualitative study while 952 were used for the quantitative study.

The matrices with electronic descriptors and descriptors derived from Dragon were grouped into a single matrix and subjected to a second stage of reduction of variables using the free software QSAR Modeling [40]. Only the descriptors with absolute Pearson's correlation coefficient ($|R|$) values with a vector **y** value of greater than 0.2 were maintained. Thus, the SAR study used a final matrix of 239 descriptors, and the QSAR study used a final matrix with 526 descriptors.

2.3. Variable selection

The step of variable selection in a QSAR study is a way to identify reduced subsets of descriptors that reproduce the observed values of a biological activity, i.e., those that are the most useful for obtaining a more accurate prediction model [41]. In this study, the ordered predictor selection (OPS) algorithm [42], an approach available in QSAR Modeling [40], was used to sort the most important descriptors. Three informative vectors available in the QSAR Modeling were used simultaneously: the correlation vector, the regression vector and an element-wise product of both. The best models were classified in descending order of statistical quality according to their root mean square error of cross-validation (*RMSECV*)

in the first step of selection, and according to their coefficient of determination of leave-one-out cross-validation (Q^2_{LOO}) in the other steps. Because OPS use PLS to select the most important variables, the descriptors were pre-processed in the two studies using the autoscaling scheme. This procedure consists in subtracting the value of each descriptor by your average, or mean-centering the descriptor, and then dividing the result by the standard deviation of that descriptor.

2.4. SAR/QSAR analysis

Both models were obtained using PLS approaches. In this method, latent variables (LV) are obtained including the dependent variable (in this case, $-\log IC_{50}$) in the analysis, in such a way that the covariance between the projection of the samples in the new axis system (also orthogonal) and the dependent variable is maximized [43]. In classical PLS (used for regression studies), the vector $\boldsymbol{y}$ is quantitative and continuous. PLS-DA is an extension of PLS. This approach is capable of effectively separating samples into different classes based on their independent variables by finding a discriminant plane between the classes. With PLS-DA, the vector $\boldsymbol{y}$ is qualitative and encodes the class membership (i.e., inactive and active compounds in this study) as a set of dummy variables (1 for inactive and 2 for active compounds). One of the advantages of using PLS-DA compared to other classification methods is that validation tools, such as calibration models, can be used [44,45]. Another advantage for this particular study was the possibility of using the OPS method to find the most important descriptors to discriminate the two classes.

With PLS-DA, it is also necessary to determine the limit from which each sample is considered to be in a specific class and thus the threshold between zero and one should be determined. When a value above the threshold is predicted, a sample is considered to belong to a class, whereas a value below the threshold indicates that the sample does not belong to that class. In this study, the adopted threshold was 0.5, a value generally employed with PLS-DA [45]. For both approaches, the descriptors were also pre-processed using the autoscaling scheme.

2.5. Model validation

In the PLS-DA study, validation followed the procedures adopted by the majority of SAR studies that use this procedure [46,47]. Thereby, the quality of the data adjustment was assessed based on its coefficient of determination (R^2) and the root mean square error of calibration (*RMSEC*), the results obtained from cross-validation (using the values of Q^2_{LOO} and *RMSECV*), and by the visual inspection of the separation of the compounds in the two analyzed classes.

Several statistical tools are suggested in literature for validation of QSAR models. The parameters adopted to validate the internal quality were the R^2, *RMSEC*, *F*-ratio test with a 95% confidence interval ($\alpha = 0.05$), Q^2_{LOO} and *RMSECV*. The adopted limits are $R^2 > 0.6$ and $Q^2_{LOO} > 0.5$. *RMSEC* and *RMSECV* should be as low as possible [48]. The *F*-test value should be higher than the *F* value in the table ($F_{p,\, n-p-1}$, where n is the number of compounds and p is the number of LV); the higher the difference between them, the more statistically significant is the model [49]. To ensure that the models generated in this study had reliable predicted variances, the following r^2_m metrics were also adopted: r^2_m(LOO); r'^2_m(LOO); average r^2_m(LOO) and Δr^2_m(LOO). For the first three metrics, the desirable results are greater than 0.5, and for the last metric, the desirable result is less than 0.2 [44].

The robustness of the model was examined via leave-*N*-out (LNO) cross validation, using approximately 25% of the training set (i.e., $N = 1$–5). This test was repeated six times for each "*N*" value and all of the rows from the data matrix and respective $-\log IC_{50}$ values were randomized in each step of the LNO process. The average value of each Q^2_{LNO} was expected to be close to each Q^2_{LOO} value, with standard deviations close to zero [50]. The possibility of chance correlation was tested using the $\boldsymbol{y}$-randomization test, where only the $\boldsymbol{y}$ vector was scrambled 10 times. The $|R|$ between the original and the randomized vector $\boldsymbol{y}$ was used to quantify chance correlation. In this approach, two regression lines are built using $|R|$ these correlation coefficients in the *x*-axis and the R^2 and Q^2_{LOO} values in the *y*-axis. The intercepts of the equations obtained in the linear regression should be lower than 0.3 for R^2 and 0.05 for Q^2_{LOO} [51].

Once internally validated, the complete data set was split into a training set ($n = 19$) and a test set ($n = 5$) to generate the real model [50]. The test set was selected manually, in such a way that the entire range of $-\log IC_{50}$ (3.759–7.398 logarithmic units) and the structural variations of the data set were well represented. The coefficient of determination in the external validation (R^2_{pred}) and root mean square error of external prediction (*RMSEP*) were used as measures of the predictive power of the QSAR model. The recommended limit is $R^2_{pred} > 0.5$ [45], and the *RMSEP* values also should be as low as possible. However, this is not enough to guarantee that the model is actually predictive.

Therefore, the following values were also obtained: (i) the Golbraikh–Tropsha slopes of the best fit line obtained by correlating the observed and predicted values with the intercept set to zero (k) and the predicted and observed values with the intercept set to zero (k'). These slopes should be $0.85 \leq x \leq 1.15$ ($x = k$ or k'); (ii) the Golbraikh–Tropsha absolute difference of the determination coefficient of the linear relation between the observed and predicted values without an intercept (R^2_0) and the predicted and observed values without an intercept (R'^2_0) $\left(\left|R^2_0 - R'^2_0\right|\right)$. The difference should be smaller than 0.3 [52]; and (iii) the r^2_m(pred) metrics, which are parameters calculated in the same way and with the same threshold of the r^2_m(LOO) metrics, but applied to external validation [53].

Finally, considering the small size of the data set used in the QSAR study, the final validation step was the calculation of the r^2_m(overall). These metrics are calculated as are the r^2_m(LOO) and r^2_m(pred), but they are based on the predictions of both the training set and the test set. Therefore, the result is based on the prediction of a relatively large number of compounds [48].

The statistical significance of the internal and external validations were calculated in the QSAR Modeling software (R^2, *RMSEC*, Q^2_{LOO}, *RMSECV*, LNO-cross validation, and $\boldsymbol{y}$-randomization test), in an in-house Microsoft Excel spreadsheet (*F*-test, R^2_{pred}, *RMSEP*, k, k', and $\left|R^2_0 - R'^2_0\right|$), and with the RmSquare Calculator (http://aptsoftware.co.in/rmsquare). The equations for calculating all of these parameters are available in Todeschini and Consonni [38], Kiralj and Ferreira [50], Golbraikh et al. [52] and Ojha et al. [53].

3. Results and discussion

3.1. SAR analysis

The variables were selected by applying the OPS algorithm [42] to a data matrix containing 252 descriptors and a model with 8 descriptors was obtained. The PLS-DA model was built and refined using the Pirouette 4 [54]. The descriptors were reduced by preferably maintaining those with the highest regression vector [44]. The best classification was obtained using 7 descriptors (see loading plot in Fig. 3A). The number of significant LV was determined using the *RMSECV* method with the cross-validation approach. The final PLS-DA classification model (Eq. (1)) has two LV that

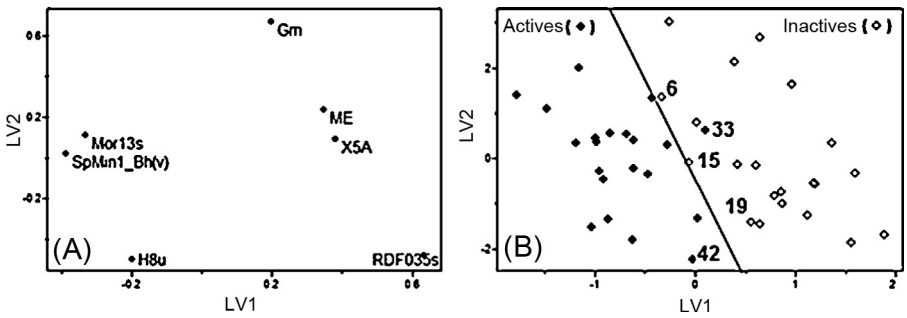

Fig. 3. Plot of the loading (A) and score (B) vectors (LV1 × LV2) for the training set.

cumulate 35.641% of the variance (LV1: 12.528%; LV2: 23.113%). The model explains 73.5% (R^2 = 0.735) of the variance and predicts 62.4% (Q^2_{LOO} = 0.624) of the variance. The separation tendency is shown in the score plot available in Fig. 3B. Five samples (**6**, **15**, **19**, **33**, **42**) did not meet the threshold (0.5) [45] adopted for the two classes, active and inactive compounds. However, inspection of the score plot (Fig. 3B) shows that this is a small error for compounds because they are located near the line that separates the classes, except for **33**, the unique active compound that was ranked among the inactive compounds.

$$\begin{aligned}\text{Class} = &-117.97 + 146.922 \times (\text{X5A}) - 61.787 \times (\text{SpMin1\_Bh(v)}) \\ &-0.040 \times (\text{Mor13s}) + 19.761 \times (\chi) - 0.335 \times (\text{H8u}) \\ &+33.307 \times (\text{Gm}) + 0.005 \times (\text{RDF035s}).\end{aligned} \quad (1)$$

$$n = 42; \quad R^2 = 0.735; \quad RMSEC = 0.257; \quad Q^2_{LOO} = 0.624; \quad RMSECV = 0.306$$

At first, the large quantity of the selected descriptors (seven) appears problematic. However, the large number of descriptors is not a real concern because the PLS method is based on the LV, which are orthogonal among themselves [32] and it prevents the classic collinearity problem among the descriptors [49]. In addition, as shown in Equation I, removing descriptors with the aim of improving the mechanistic interpretation of the model affects the classification of the samples.

Model (1) was obtained after removing an outlier compound. Detection of outliers was performed using studentized residuals (σ) versus the leverage samples plot [55]. No compound had a residuals value higher than 2.5 × σ. However, two compounds (**23** and **24**) had leverage values higher than the leverage cutoff line. Among them, compound **24** had the highest leverage value (h = 0.353) and a high studentized residual value (σ = 1.652). Although this sample's σ value is less than 2.5, it had a major influence on the quality of the model because it has the third largest σ value and the largest h value of the samples in the model. Therefore, this sample was omitted, significantly improving the sample separation. The outlier detection graph is presented in the Supplementary Material (Fig. S1).

The selected descriptors and their respective autoscaled descriptors were the following: **RDF035s** – Radial Distribution Function – 035 weighted by the I-state (0.487); **H8u** – H autocorrelation of lag 8, unweighted (−0.328); **Mor13s** – 3D-MoRSE descriptor signal 13 weighted by the I-state (−0.300); **Gm** – total symmetry index weighted by mass (0.476); **X5A** – average connectivity index of order 5 (0.454); **SpMin1_Bh(v)** – smallest eigenvalue *n*. 1 of the Burden matrix weighted by the van der Waals volume (−0.319); and χ – Molecular Electronegativity (0.431). All of the values are available in the Supplementary Material (Table S1). Regression coefficients larger than approximately half (0.244) of the maximum regression coefficient value obtained indicated that all of the descriptors were significant for the model [56].

3.2. QSAR analysis

Selecting the variables from a data matrix of 528 descriptors using the OPS algorithm provided the best model based on 8 descriptors. This model was refined using Pirouette 4 software [54] in the way that the PLS-DA model was obtained [44]. The final model (Equation II) has six descriptors and is based on two LV that cumulate 87.916% of the variance (LV1: 57.350%; LV2: 30.566%) and that explain 76.7% (R^2 = 0.767) and predict 65.7% (Q^2_{LOO} = 0.657) of the variance. The F value (26.313) was higher than the corresponding tabled value (3.634, for p = 2 and $n - p - 1$ = 16) with a 95% confidence interval (α = 0.05). The predicted values from the cross-validation step and the residuals are available in the Supplementary Material (Table S2). The difference between the values of the R^2 and the Q^2_{LOO} was 0.110 units; this difference indicates that the model does not suffer from overfitting [50]. Finally, the results from the modified squared correlation coefficient of LOO cross-validation (r^2_m(LOO)) metrics are consistent with the proposed limits [53].

$$\begin{aligned}-\log \text{IC}_{50} = &-1.924 - 11.262 \times (\text{E}_{\text{LUMO}}) + 13.016 \times (\omega) \\ &+0.069 \times (\text{Mor09s}) + 9.475 \times (\text{Ds})\end{aligned}$$

$$\begin{aligned}&n = 19; \quad R^2 = 0.767; \quad RMSEC = 0.462; \quad F_{(2,16)} = 26.313; \\ &Q^2_{LOO} = 0.657; \quad RMSECV = 0.514; \quad r^2_m(LOO) = 0.644; \\ &r'^2_m(\text{LOO}) = 0.581; \quad \text{Average } r^2_m(\text{LOO}) = 0.613; \\ &\Delta r^2_m(\text{LOO}) = 0.063.\end{aligned} \quad (2)$$

The selected descriptors and their respective autoscaled descriptors were $\mathbf{E_{LUMO}}$ – the energy of the lowest unoccupied molecular orbital (−0.229), ω – the electrophilicity index (0.284), **Mor09s** – 3D-MoRSE descriptor signal 13/weighted by the I-state (0.358), and **Ds** – *D* total accessibility index/weighted by the atomic electrotopological states (0.327). In this model, the reference value is 0.179 [56], and all descriptors can be considered important for it. The studentized residuals (σ) versus the leverage samples plot was verified, and no compound presented residuals higher than 2.5 × σ. The values of all of the descriptors are available in the Supplementary Material (Table S3).

The results of ***y***-randomization analysis and LNO cross-validation are available in Fig. 4. The ***y***-randomization analysis (Fig. 4A) helps clarify whether the explained and predicted variances are due to chance correlation [51]. It is obvious that the results obtained for all of the randomized models are of relatively poor quality compared to those of the original model because the intercepts are within the acceptable values recommended in literature, i.e., between 0.3 and 0.05. These results indicate that the variance explained by the model was not due to chance correlation.

The LNO cross-validation (Fig. 4B) employs smaller training sets than the LOO cross-validation, and it can be repeated several times due to the large number of combinations that arise when more than one compound is left out of the training set, one at a time. A QSAR model is considered robust when the average value of Q^2_{LNO} is relatively high and close to that of Q^2_{LOO} [56]. The model

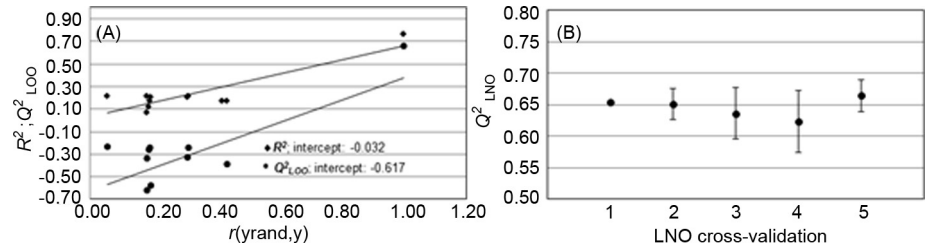

Fig. 4. Results of **y**-randomization test (A) and LNO cross-validation (B).

obtained in this study has an average Q^2_{LNO} of 0.646, only 0.008 units lower than that of Q^2_{LOO}. The standard deviation for each "*N*" (performed in hexaplicate) value is small, with a maximum of 0.049 for Q^2_{L4O}.

Only externally validated models can be considered realistic and applicable for drug design [57]. The results obtained for this step (Table 2) demonstrate that the model has a high external predictive power considering the proposed limits. The R^2_{pred}, which is usually used to measure the external predictive power, was higher than the adopted threshold value ($R^2_{pred} > 0.5$), and the associated *RMSEP* was low. The Golbraikh-Tropsha [38,52,53] and the r^2_m [48] metrics help confirm the predictive power of the model. The Golbraikh–Tropsha slopes k and k' and the absolute difference between R^2_0 and R'^2_0 $\left(\left|R^2_0 - R'^2_0\right|\right)$ are within acceptable ranges. The modified squared correlation coefficient of external validation r^2_m(pred) metrics were also suitable.

The overall modified squared correlation coefficient (r^2_m(overall)) was also evaluated, resulting in r^2_m(overall) = 0.743, r'^2_m(overall) = 0.604, average r^2_m(overall) = 0.673, and Δr^2_m(overall) = 0.139. These results are consistent with the proposed limits. This metric is not based only on the limited number of the test set compounds (in this case, five compounds) because it also considers the observed −log IC_{50} and that predicted in the LOO cross-validation. Thus, the results for this metric are more reliable for the purpose of prediction [48].

3.3. Model discussion

The models obtained in this study are of reasonable statistical quality. However, it is always desirable to obtain an interpretative model that is able to relate the physicochemical properties represented by molecular descriptor, with the mechanism of action of the system under study. However, interpreting a QSAR model in terms of the contribution of molecular descriptors to the modeled activity is always a difficult task [58]. Furthermore, the mechanism by which the set of active cucurbitacins exert their cytotoxic effect on the tested cell lines is not yet known. Therefore, the interpretation of the models for this specific set of compounds is based only on the potential encoded information in the selected descriptors per se, in studies from the literature, and in general information about cytotoxic mechanisms.

About the cytotoxic activity, Leão [59] remarked that carcinogenic substances, or their metabolites, are electrophilic substances that would tend to participate in reaction as electron receptors. When these chemicals react with protein structures or DNA, they can induce apoptosis. A number of SAR/QSAR studies have described the cellular apoptosis mechanism as related to the ability of compounds to receive electrons. Afantitis et al. [60] and Takano et al. [61] selected the descriptor E_{LUMO} (lowest unoccupied molecular orbital energy), which is clearly related to the capacity of compounds to receive electrons. The first study concerns predicting the mechanism of apoptotic induction by 4-aryl-4H-chromenes, whereas the second study assessed the cytotoxic effects of a series of naphtho[2,3-b]furan-4,9-diones in oral squamous cell carcinoma cell lines types 2, 3 and 4 (HSC-2, HSC-3 and HSC-4). In contrast, Qin et al. [22] investigated a series of chloroethylnitrosoureas that act as alkylating agents and selected GAP (the difference between the E_{HOMO}, the energy of the highest occupied molecular orbital, and E_{LUMO}) and the partial atomic charge as descriptors, both of which have been related to the alkylation of DNA in leukemia cells. Hansch et al. [62,63] related the classic descriptors LogP (logarithm of 1-octanol/water partition coefficient) and MR (molar refractivity) with the ability of compounds to interact with DNA and induce cell apoptosis in different tumor cell lines. De Souza et al. [64], in a classification obtained using principal component analysis (PCA), were able to classify a set of mono- and di-substituted tetrazole and oxadiazole derivatives assayed using tumor cells as active or inactive using the descriptors of the electrophilic reactivity index for a carbon atom (f^{elec}_{max}) and the relative negative charged surface area (RNCSA). Finally, Arantes et al. [65] selected E_{HOMO} and three partial charges when conducting a PCA study that indicated that a set of sesquiterpene lactones are prone to react with nucleophiles via a Michael addition reaction.

In the literature, some studies reports that cucurbitacins may inhibit the Signal Transducer and Activator of Transcription type 3 (STAT3) to induce apoptosis in some cancer cell lines [16]. STAT proteins are latent transcription factors. Ligand-dependent activation of the STATs is often associated with differentiation and/or growth regulation, whereas constitutive activation is often associated with growth deregulation. A growing number of tumors (multiple myeloma, leukemia, lymphoma, breast, head, neck, melanoma, ovarian, pancreatic, prostate and lung) are reported to have constitutively activated STAT3. It is possible that constitutively active STAT3 protects cell lines that have become growth factor-independent against apoptosis [66]. Members of our group recently described that a new natural cucurbitacin identified in *Wilbrandia ebracteata* roots induces apoptosis in the A549 cell line that has constitutively activated STAT3 [67]. Similar results were observed by Sun et al. [16] for the same tumor cell line, when cucurbitacin Q was evaluated.

Table 2
Results from external validation step (Model II).

Compound	−log IC_{50} observed	−log IC_{50} predicted	Residuals
2	7.398	6.822	0.576
14	3.949	4.320	−0.371
18	5.578	5.374	0.204
20	4.371	4.216	0.155
38	4.939	5.215	−0.276
R^2_{pred}	0.924		
RMSEP	0.350		
r^2_m(pred)	0.777		
r'^2_m(pred)	0.685		
Average r^2_m(pred)	0.731		
Δr^2_m(pred)	0.091		
$\left\|R^2_0 - R'^2_0\right\|$	0.045		
k	1.019		
k'	0.978		

3.3.1. SAR model: mechanistic interpretation

The radial distribution function (RDF) descriptor measures the probability that an atom is in a spherical volume of radius *R* [39]. The calculation considers the number of atoms, one atomic property and the distance between the atoms [68]. For some active compounds, this radius (3.5 Å from the center of the molecule) is located mainly in the region of the D ring. As for inactive compounds, this radius is displaced from the D ring (Supplementary Material, Fig. S2). This descriptor is directly related to the intrinsic state (*s*), which is related to the number of valence electrons. Thus, the higher the value of this parameter, the greater is the RDF035s value. The negative sign of the coefficient of the descriptor shows that activity is favored by a smaller number of valence electrons. This may occur due to the molecules capturing electrons (i.e., to perform electrophilic attack) to display cytotoxicity. The outline of the radius encompasses mainly the D ring and part of the side chain, which may contain atoms capable of performing an electrophilic attack.

The descriptor χ corresponds to the ability of an element to receive electrons and therefore, form negative ions. The greater the ionic character of an element, the greater is its electronegativity [38]. The negative sign of this descriptor suggests that the lower the ionic character of compound, the higher is the probability of it being active, which can be related to its ability to make an electrophilic attack.

The descriptor Gm is a global WHIM (Weighted Holistic Invariant Molecular) geometric descriptor. Calculating Gm is based on the atomic projections as a function of the principal axes (*x*, *y*, *z*) and it is obtained to acquire significant three-dimensional molecular data related to size, shape, symmetry and the distribution of atoms with respect to invariable reference criteria [39]. Because the coefficient is negative, the less branched and symmetrical the molecule, the greater is its tendency to be active [69]. Moreover, because the atomic mass property (*m*) is used in the weighting, the presence of high molecular weight atoms is not conducive to biological activity.

The connectivity index X5A is related to the number of edges and the number of atoms in a molecule. The higher the number of atoms (*A*), the greater is the value of X5A [38]. An increase in this value tends to reduce the potential for activity. According to Gupta et al. [70], increases in the average connectivity indices are related to branching. Thus, multi-branched molecules will have difficulty producing cytotoxicity. This tendency is related to the interpretations obtained using the Gm descriptor. Nakhjiri et al. [71] presented similar results regarding cytotoxic effects against the MCF-7 cell line: the greater the value of the average connectivity index of order 5 (X4A), the less cytotoxic were the compounds that they analyzed. The first order valence molecular connectivity index descriptor (X1v) was also used to describe the toxicity of chlorophenols in the L929 cell line (fibrosarcoma) [72].

H8u (H autocorrelation of lag 8/unweighted) is a GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptor. In a model proposed by Zhou et al. [73], the values for this descriptor displayed a negative coefficient for growth inhibition of P388 tumor cells (leukemia). The H8*w* descriptor (where *w* are different properties used for weighting) was selected in several studies on QSAR modeling of carcinogenic activity [74,75]. Thus, the characteristics encoded in this type of descriptor appear to be related to the carcinogenic properties.

Mor13s is a 3D-molecule representation of structures based on electron diffraction (3D-MoRSE) descriptor that uses a generalized scattering function called "The molecular transform". This function takes into account information about the 3D arrangement of atoms without the ambiguities that occur when using chemical graphs [76,77]. This descriptor is calculated by summing the atomic weights viewed by angular scattering functions (13 Å) and weighted by (*s*). In this case, the coefficient is positive and because (*s*) is directly proportional to the value of this descriptor, it contradicts the interpretation of the corresponding RDF035s value. Moreover, Ramírez-Galicia et al. [77] stated that 3D-MoRSE descriptors are difficult to interpret in medicinal chemistry. Thus, it is possible that this descriptor encodes different information than that represented by RDF035s. The low Pearson's correlation between those descriptors ($R = 0.18$) strengthens this hypothesis.

Table 3
Contribution of selected descriptors in each LV (Model II).

Descriptor	LV1	LV2
Mor09s	0.384	0.850
Ds	0.541	0.095
E_{LUMO}	−0.506	0.425
ω	0.551	−0.297

The topological descriptor SpMin1_Bh(v) is a Burden eigenvalue; the smallest eigenvalues were proposed as molecular descriptors with a high discrimination power that can be applied to recognizing and ordering molecular structures. The lowest eigenvalues contain contributions from all of the atoms and thus reflect the topology of the entire molecule [38]. As Fig. 3B shows, this descriptor has more influence in the first LV on classifying active compounds and therefore, it is very important for discriminating between the two classes (active or inactive compounds). Furthermore, it is weighted by van der Waals volume (*v*), suggesting that an increase in volume is related to potential for cytotoxic activity.

3.3.2. QSAR model: mechanistic interpretation

The QSAR model, constructed using only four descriptors that encode two LVs, is simpler to interpret. The OPS algorithm [42] selected descriptors easily relatable to the likely toxicity mechanism (electrophilic attacks). Previous studies indicated that this method of selecting variables is likely to select molecular descriptors associated with biological phenomena [78,79], thereby providing results that improve the understanding of certain mechanisms.

The E_{LUMO} descriptor refers to the lowest energy level in the molecule that does not contain electrons. Molecules with low E_{LUMO} values are more likely to accept electrons than molecules whose values are high [38]. The data in the literature corroborate this correlation [60,61]. Table S3 (Supplementary Material) shows that the most active compounds (**32** and **34**) possess the lowest E_{LUMO} values (−0.087 and −0.096 hartress), whereas the least active compounds (**21** and **29**) possess greater E_{LUMO} values (−0.027 and −0.035 hartress), which explains the negative sign described in this model. The descriptor ω has an equivalent meaning, and the higher its value, the easier it is for molecule to receive electrons [38]. The same interpretation for both descriptors suggests that they encode the same information in the model, which is enhanced by the high Pearson's correlation between them ($R = 0.944$). However, it is important to note that the PLS regression generates LVs that are mutually orthogonal [80]. Consulting the weight values of the descriptors in each latent variable (Table 3) confirmed that the contributions of the LV2 of E_{LUMO} and ω (which accumulates 30.161% of the variance) are different.

Mor09s is also a 3D-MoRSE descriptor and Ds is also a global WHIM descriptor. Like Mor13s, they are weighted by (*s*), which is directly proportional to their values, and their coefficients in the model are positive. This can lead to misinterpretation of their roles in the model. Nevertheless, as for RDF035s and Mor13s, the Pearson's correlation between Mor09s and Ds is low ($R = 0.48$), and their individual contributions to the LV obtained are quite different (Table 3).

As previously mentioned, the interpretation of the 3D-MoRSE descriptors can be complex, but the WHIM descriptors, such as D*w*,

are related to the global molecular density. Thus, the higher the value of D*w*, the greater is the trend toward the compounds being active. Considering that the density of a molecule is inversely proportional its volume, it is possible to propose that the small size of denser molecules allows them to penetrate the plasma membrane of the cells studied more rapidly than do larger molecules, and thus exert their cytotoxic effect more easily. One of the factors determining the higher global molecular density is the number of atoms in the molecule (nAT) [39]. These variables are directly proportional to each other. Although this idea is inverse the proposal to volume reduction, it is interesting to note that the greater accessibility of an atom to the external environment (i.e., how much less branched a molecule is), the higher its value (*s*) [81], which may contribute to the increase of global molecular density. This interpretation may explain the relationship between this property and those encoded by the Gm and X5A descriptors in model (1). Finally, because Ds is also a geometric descriptor, its contribution to the model indicates that the form a molecule adopts also influences its activity.

3.4. Interpretation remarks

It is interesting to note that STAT3 is activated by phosphorylation of a single tyrosine, typically in response to extracellular ligands [66]. In this type of reaction, the phosphate performs an electrophilic attack on the hydroxyl side chain of the amino acid. Thus, it is possible that the cytotoxic effect of the compounds described here occurs due to their reaction with this amino acid, which would block its phosphorylation. Silva et al. [67] and Sun et al. [16] reported that cucurbitacins inhibit the activation pathway of STAT3 in the A549 cells by this mechanism.

This hypothesis is reinforced by the interpretation of the values for the RDF035s and χ descriptors in model (1) and the interpretation of the E_{LUMO} and ω values in model (2). However, cytotoxic activity, as well as other biological activities, is most likely a multivariate phenomenon. Thus, the interpretation of the values of the Gm, X5A, SpMin1_Bh(v), Mor13s, Mor09s e Ds descriptors indicates that other processes, such as transport across cell membranes, are also relevant to the activity under study. Finally, the selection of the descriptor H8u is supported by reports in the literature.

Considering that there are no experimental results to suggest that the specific data set evaluated in this study reflects known mechanisms of action, the indications obtained and the results acquired in both the SAR and QSAR studies should be considered acceptable.

4. Conclusions

The results indicate that it is possible to explain the cytotoxic activity of cucurbitacins using molecular descriptors that encode informations on the structural variations of these derivatives. In the SAR study, the results indicated that the classification of a cucurbitacin as active and inactive in terms of cytotoxic activity against A549 cells is associated mostly with the possibility of electrophilic attack, though the structural symmetry and molecular volume also influence the classification. The model constructed in the QSAR study had good internal and external predictive power, is robust, and free of chance correlation, demonstrating that the model is significant and that it can be used for predictive purposes. The selected descriptors suggest that the cytotoxic potency varies mainly according to the capacity of the molecule to perform electrophilic attacks. However, the molecule's size and shape are also relevant, which indicates that cytotoxic potency is a biological phenomenon with multivariate characteristics. Therefore, even though the mechanism of action by which these compounds exert their cytotoxic activity is not known, the interpretation of the models strengthens the possibility that it involves an electrophilic attack mechanism, including the interaction with STAT3. In conclusion, this study provides some insight into the characteristics that endow certain cucurbitacin derivatives with cytotoxic activity against A549 cells (human lung adenocarcinoma). Thus, this study may enhance the understanding of the activity of this class of compounds and may be useful for designing new potent cucurbitacin derivatives that would be prototypes for lung cancer therapeutics.

Acknowledgements

The Brazilian authors are grateful for financial support: Fundação Araucária (grant 2010/7354); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, MCTI, grant 472979/2011-6); Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC/PRONEX, grant 2671/2012-9); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/MEC) and CNPq/MCTI (for granting their research fellowships and resources PROAP). The Argentinian authors thank to Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT) and UBA (Univerdad de Buenos Aires) for financial support and research fellowships.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2013.12.004.

References

[1] A. Ganesan, The impact of natural products upon modern drug discovery, Curr. Opin. Chem. Biol. 12 (2008) 306–317.
[2] D.J. Newman, G.M. Cragg, Natural products as sources of new drugs over the 30 years from 1981 to 2010, J. Nat. Prod. 75 (2012) 311–335.
[3] K.L. Lang, I.T. Silva, L.A. Zimmermann, V.R. Machado, M.R. Teixeira, M.I. Lapuh, M.A. Galetti, J.A. Palermo, G.M. Cabrera, L.S.C. Bernardes, C.M.O. Simões, E.P. Schenkel, M.S.B. Caro, F.J. Durán, Synthesis and cytotoxic activity evaluation of dihydrocucurbitacin B and cucurbitacin B derivatives, Bioorg. Med. Chem. 20 (2012) 3016–3030.
[4] M. Miró, Cucurbitacins and their pharmacological effects, Phytother. Res. 9 (1995) 159–168.
[5] J.C. Chen, M.H. Chiu, R.L. Nie, G.A. Cordel, S.X. Qiu, Cucurbitacins and cucurbitane glycosides: structures and biological activities, Nat. Prod. Rep. 22 (2005) 386–399.
[6] J.L. Ríos, J.M. Escandell, M.C. Recio, R. Atta Ur, New insights into the bioactivity of cucurbitacins, Stud. Nat. Prod. Chem. 32 (2005) 429–469.
[7] D.H. Lee, G.B. Iwanski, N.H. Thoennissen, Cucurbitac: ancient compound shedding new light on cancer treatment, Sci. World J. 10 (2010) 413–418.
[8] K.L.K. Duncan, M.D. Duncan, M.C. Alley, E.A. Sausville, Cucurbitacin E-induced disruption of the actin and vimentin cytoskeleton in prostate carcinoma cells, Biochem. Pharmacol. 52 (1996) 1553–1560.
[9] J.M. Escandell, P. Kaler, M.C. Recio, T. Sasazuki, S. Shirasawa, L. Augenlicht, J.L. Ríos, L. Klampfer, Activated kRas protects colon cancer cells from cucurbitacin-induced apoptosis: the role of p53 and p21, Biochem. Pharmacol. 76 (2008) 198–207.
[10] N. Wakimoto, D. Yin, J. O'Kelly, T. Haritunians, B. Karlan, J. Said, H. Xing, H.P. Koeffler, Cucurbitacin B has a potent antiproliferative effect on breast cancer cells in vitro and in vivo, Cancer Sci. 99 (2008) 1793–1797.
[11] X. Shi, B. Franko, C. Frantz, H.M. Amin, R. Lai, JSI-124 (cucurbitacin I) inhibits Janus kinase-3/signal transducer and activator of transcription-3 signalling, downregulates nucleophosmin-anaplastic lymphoma kinase (ALK), and induces apoptosis in ALK-positive anaplastic large cell lymphoma cells, Br. J. Haematol. 135 (2006) 26–32.
[12] T. Tannin-Spitz, S. Grossman, S. Dovrat, H.E. Gottlieb, M. Bergman, Growth inhibitory activity of cucurbitacin glucosides isolated from *Citrullus colocynthis* on human breast cancer cells, Biochem. Pharmacol. 73 (2007) 56–67.
[13] N.H. Thoennissen, G.B. Iwanski, N.B. Doan, R. Okamoto, P. Lin, S. Abbassi, J.H. Song, D. Yin, M. Toh, W.D. Xie, J.W. Said, H.P. Koeffler, Cucurbitacin B induces apoptosis by inhibition of the JAK/STAT pathway and potentiates antiproliferative effects of gemcitabine on pancreatic cancer cells, Cancer Res. 69 (2009) 5876–5884.
[14] K.T. Chan, F.Y. Meng, Q. Li, C.Y. Ho, T.S. Lam, Y. To, W.H. Lee, M. Li, K.H. Chu, M. Toh, Cucurbitacin B induces apoptosis and S phase cell cycle arrest in BEL-7402

human hepatocellular carcinoma cells and is effective via oral administration, Cancer Lett. 294 (2010) 118–124.
[15] L. Yang, S. Wu, Q. Zhang, F. Liu, P. Wu, 23,24-Dihydrocucurbitacin B induces G2/M cell-cycle arrest and mitochondria-dependent apoptosis in human breast cancer cells (Bcap37), Cancer Lett. 256 (2007) 267–278.
[16] J. Sun, M.A. Blaskovich, R. Jove, S.K. Livingston, D. Coppola, S.M. Sebti, Cucurbitacin Q: a selective STAT3 activation inhibitor with potent antitumor activity, Oncogene 27 (2008) 1344.
[17] M.A. Blaskovich, J. Sun, A. Cantor, J. Turkson, R. Jove, S. Sebti, Discovery of JSI-124 (Cucurbitacin I), a selective Janus kinase/signal transducer and activator of transcription 3 signaling pathway inhibitor with potent antitumor activity against human and murine cancer cells in mice, Cancer Res. 63 (2003) 1270–1279.
[18] B. Jayaprakasam, N.P. Seeram, M.G. Nair, Anticancer and antiinflammatory activities of cucurbitacins from *Cucurbita andreana*, Cancer Lett. 189 (2003) 11–16.
[19] M.D. Duncan, K.L.K. Duncan, Cucurbitacin E targets proliferating endothelia, J. Surg. Res. 69 (1997) 55–60.
[20] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, A practical overview of quantitative structure-activity relationship, EXCLI J. 8 (2009) 74–88.
[21] F.A. Molfetta, A.T. Bruni, F.P. Rosseli, A.B.F. Silva, A partial least squares and principal component regression study of quinone compounds with trypanocidal activity, Struct. Chem. 18 (2007) 49–57.
[22] Y. Qin, H. Deng, H. Yan, R. Zhong, An accurate nonlinear QSAR model for the antitumor activities of chloroethylnitrosoureas using neural networks, J. Mol. Graph. Model. 29 (2011) 826–833.
[23] A. Speck-Planche, V.V. Kleandrova, F. Luan, M.N.D.S. Cordeiro, Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents, Anticancer Agents Med. Chem. 12 (2012) 678–685.
[24] A. Speck-Planche, V.V. Kleandrova, F. Luan, M.N.D.S. Cordeiro, Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents, Eur. J. Pharm. Sci. 47 (2012) 273–279.
[25] U. Lahl, U. Gundert-Remy, The use of (Q)SAR methods in the context of REACH, Toxicol. Mech. Meth. 18 (2008) 149–158.
[26] G.V. Dang, B.M. Rode, H. Stuppner, Quantitative electronic structure-activity relationship (QESAR) of natural cytotoxic compounds: maytansinoids, quassinoids and cucurbitacins, Eur. J. Pharm. Sci. 2 (1994) 331–350.
[27] J. Bartalis, F.T. Halaweish, In vitro and QSAR studies of cucurbitacins on HepG2 and HSC-T6 liver cell lines, Bioorg. Med. Chem. 19 (2011) 2757–2766.
[28] K.L. Lang, T.R. Guimarães, V.R. Machado, L.A. Zimmermann, I.T. Silva, M.R. Teixeira, F.J. Durán, J.A. Palermo, C.M.O. Simões, M.S.B. Caro, E.P. Schenkel, New cytotoxic cucurbitacins from *Wilbrandia ebracteata* Cogn., Planta Med. 77 (2011) 1648–1651.
[29] V.R. Machado, Semi-síntese de derivados glicosilados de dihidrocucurbitacina B, isolada de *Wilbrandia ebracteata* Cogn, Dissertation, Pharmacy School, Universidade Federal de Santa Catarina, 2012.
[30] M.R. Farias, E.P. Schenkel, R. Mayer, G. Rucker, Cucurbitacins as constituents of *Wilbrandia ebracteata*, Planta Med. 59 (1993) 272–275.
[31] T. Mosmann, Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays, J. Immunol. Methods 65 (1983) 55–63.
[32] M. Barker, W. Rayens, Partial least squares for determination, J. Chemom. 17 (2003) 166–173.
[33] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[34] ChemOffice Ultra, version 8.0, Cambridge Soft Corporation, Cambridge, USA, 2004.
[35] Gaussian 09W, version 7.0, Gaussian Inc., Wallingford, USA, 2009.
[36] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminform. 3 (2011) 33.
[37] Gauss View, version 5.0, Gaussian Inc., Wallingford, USA, 2009.
[38] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, 2nd ed., Wiley-VCH, Weinheim, 2009.
[39] Dragon User Guide, version 6.0, Talete srl., Italy, 2011.
[40] J.P.A. Martins, M.M.C. Ferreira, QSAR modeling: a new open source computational package to generate and validate QSAR models, Quim. Nova 26 (2013) 554–560.
[41] M.P. González, C. Terán, L. Saíz-Urra, M. Teijeira, Variable selection methods in QSAR: an overview, Curr. Top. Med. Chem. 8 (2008) 1606–1627.
[42] R.F. Teófilo, J.P. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48.
[43] P.P. Roy, J.T. Leonard, K. Roy, Exploring the impact of size of training sets for the development of predictive QSAR models, Chemom. Intell. Lab. Syst. 90 (2008) 31–42.
[44] P.P. Roy, K. Roy, On some aspects of variable selection for partial least squares regression models, QSAR Comb. Sci. 27 (2008) 302–313.
[45] L.T.F.M. Camargo, M.M. Sena, A.J. Camargo, A quantum chemical and chemometrical study of indolo[2,1-b]quinazoline and their analogues with cytotoxic activity against breast cancer cells, SAR QSAR Environ. Res. 20 (2009) 537–549.
[46] B. Dejaegher, L. Dhooghe, M. Goodarzi, S. Apers, L. Pieters, Y.V. Heyden, Classification models for neocryptolepine derivatives as inhibitors of the beta-haematin formation, Anal. Chim. Acta 705 (2011) 98–110.
[47] S. Askjaer, M. Langgård, Combining pharmacophore fingerprints and PLS-discriminant analysis for virtual screening and SAR elucidation, J. Chem. Inf. Model. 48 (2008) 476–488.
[48] I. Mitra, A. Saha, K. Roy, Chemometric QSAR modeling and *in silico* design of antioxidant NO donor phenols, Sci. Pharm. 79 (2011) 31–57.
[49] A.C. Gaudio, E. Zandonade, Proposition, validation and analysis of QSAR models, Quim. Nova 24 (2001) 658–671.
[50] R. Kiralj, M.M.C. Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: theory and application, J. Braz. Chem. Soc. 20 (2009) 770–787.
[51] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs, Environ. Health Perspect. 111 (2003) 1361–1375.
[52] A. Golbraikh, M. Shen, Z. Xiao, Y. Xiao, K. Lee, A. Tropsha, Rational selection of training and test set for the development of validated QSAR models, QSAR Comb. Chem. 17 (2003) 241–253.
[53] P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Further exploring rm2 metrics for validation of QSPR models, Chemom. Intell. Lab. Syst. 107 (2011) 194–205.
[54] Pirouette, version 4, Informetrix Inc., USA, 2007.
[55] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Chem. 26 (2007) 694–701.
[56] S. Wold, L. Eriksson, S. Clementi, Statistical validation of QSAR results, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, Wiley-VCH, Weinheim, 1998, pp. 309–318.
[57] V. Consonni, D. Ballabio, R. Todeschini, Evaluation of model predictive ability by external validation techniques, J. Chemom. 24 (2010) 194–201.
[58] F. Luan, A. Melo, F. Borges, M.N.L.D.S. Cordeiro, Affinity prediction on A3 adenosine receptor antagonists: the chemometric approach, Bioorg. Med. Chem. 19 (2011) 6853–6859.
[59] M.B.C. Leão, A.C. Pavão, V.A.A. Espinoza, C.A. Taft, E.P. Bulnes, A multivariate model of chemical carcinogenesis, J. Mol. Struct. Theochem. 719 (2005) 129–135.
[60] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromenes, Bioorg. Med. Chem. 14 (2006) 6686–6694.
[61] A. Takano, K.E.N. Hashimoto, M. Ogawa, J. Koyanagi, T. Kurihara, H. Wakabayashi, H. Kikuchi, Y. Nakamura, N. Motohashi, H. Sakagami, K. Yamamoto, A. Tanaka, Tumor-specific cytotoxicity and type of cell death induced by naphtho[2,3-b]furan-4,9-diones and related compounds in human tumor cell lines: relationship to electronic structure, Anticancer Res. 29 (2009) 455–464.
[62] C. Hansch, B. Bonavida, A. Jazirehi, J.J. Cohen, C. Milliron, A. Kurup, Quantitative structure–activity relationships of phenolic compounds causing apoptosis, Bioorg. Med. Chem. 11 (2003) 617–620.
[63] C. Hansch, A. Jazirehi, S.B. Mekapati, R. Garga, B. Bonavida, QSAR of apoptosis induction in various cancer cells, Bioorg. Med. Chem. 11 (2003) 3015–3019.
[64] A.O. de Souza, M.T.C. Pedrosa, J.B. Alderete, A.F. Cruz, M.A.F. Prado, R.B. Alves, C.L. Silva, Cytotoxicity, antitumoral and antimycobacterial activity of tetrazole and oxadiazole derivatives, Pharmazie 60 (2005) 396–397.
[65] F.F.P. Arantes, L.C.A. Barbosa, C.R.A. Maltha, A.J. Demunera, P.H. Fidêncio, J.W.M. Carneiro, A quantum chemical and chemometric study of sesquiterpene lactones with cytotoxicity against tumor cells, J. Chemom. 25 (2011) 401–407.
[66] J.F. Bromberg, M.H. Wrzeszczynska, G. Devgan, Y. Zhao, R.G. Pestell, C. Albanese, J.E. Darnell Jr., STAT3 as an oncogene, Cell 98 (1999) 295–303.
[67] I.R. Silva, M.R. Teixeira, K.L. Lang, T.R. Guimarães, S.E. Dudek, F.J. Durán, S. Ludwig, M.S.B. Caro, E.P. Schenkel, C.M.O. Simões, Proliferative inhibition and apoptotic mechanism on human non-small-cell lung cancer (A549 cells) of a novel cucurbitacin from Wilbrandia ebracteata Cogn., Int. J. Cancer Res. 9 (2013) 54–68.
[68] R.M.V. Abreu, C.F.R. Ferreira, M.J.R.P. Queiroz, QSAR model for predicting radical scavenging activity of di(hetero)arylamines derivatives of benzo[b]thiophenes, Eur. J. Med. Chem. 44 (2009) 1952–1958.
[69] P. Gramatica, E. Papa, A. Marrocchi, L. Minuti, A. Taticchi, Quantitative structure–activity relationship modeling of polycyclic aromatic hydrocarbon mutagenicity by classification methods based on holistic theoretical molecular descriptors, Ecotoxicol. Environ. Saf. 66 (2007) 353–361.
[70] R.A. Gupta, A.K. Gupta, L.K. Soni, S.G. Kaskhedikar, Study of physicochemical properties-antitubercular activity relationship of naphthalene-1,4-dione analogs: a QSAR approach, Chem. Pap. 63 (2009) 723.
[71] M. Nakhjiri, M. Safavi, E. Alipour, S. Emami, A.F. Atash, M. Jafari-Zavareh, S.K. Ardestani, M. Khoshneviszadeh, A. Foroumadi, A. Shafiee, Asymmetrical 2,6-bis(benzylidene)cyclohexanones: synthesis, cytotoxic activity and QSAR study, Eur. J. Med. Chem. 50 (2012) 113–123.
[72] X. Liu, J. Chen, H. Yu, J. Zhao, J.P. Giesy, X. Wang, Quantitative structure activity relationship (QSAR) for toxicity of chlorophenols on L929 cells in vitro, Chemosphere 64 (2006) 1619–1626.
[73] W. Zhou, Z. Dai, Y. Chen, Z. Yuan, Computational QSAR models with high-dimensional descriptor selection improve antitumor activity design of ARC-111 analogues, Med. Chem. Res. 22 (2013) 278–286.
[74] B. Bhhatarai, P. Gramatica, Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse, Mol. Divers. 15 (2011) 467–476.
[75] A.M. Helguera, M. Natália, D.S. Cordeiro, M.A.C. Pérez, R.D. Combes, M.P. González, QSAR modeling of the rodent carcinogenicity of nitrocompounds, Bioorg. Med. Chem. 16 (2008) 3395–3407.

[76] B.F. Rasulev, N.D. Abdullaev, V.N. Syrov, J. Leszczynski, A Qqantitative structure-activity relationship (QSAR) study of the antioxidant activity of flavonoids, QSAR Comb. Sci. 24 (2005) 1056–1065.
[77] G. Ramírez-Galicia, H. Martínez-Pacheco, R. Garduño-Juárez, O. Deeb, Exploring QSAR of antiamoebic agents of isolated natural products by MLR, ANN, and RTO, Med. Chem. Res. 21 (2012) 2501–2516.
[78] A.P. Hartmann, D.H. Jornada, E.B. Melo, A new fully validated and interpreted quantitative structure–activity relationship model of p-aminosalicylic acid derivatives as neuraminidase inhibitors, Chem. Pap. 67 (2013) 556–567.
[79] E.B. Melo, M.M.C. Ferreira, A 4D structure–activity relationship model to predict HIV-1 integrase strand transfer inhibition using the LQTA-QSAR methodology, J. Chem. Inf. Mod. 52 (2012) 1722–1732.
[80] S. Wold, PLS for multivariate linear modeling, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, Wiley-VCH, Weinheim, 1998, pp. 195–218.
[81] L.B. Kier, L.H. Hall, The electrotopological state: structure modeling for QSAR and database analysis, in: J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach, United Kingdom, 1999, pp. 491–562.