British Journal of Pharmaceutical Research
7(2): 122-139, 2015, Article no.BJPR.2015.097
ISSN: 2231-2919

SCIENCEDOMAIN *international*
www.sciencedomain.org

A QSAR Study of New Caffeine Derivatives with Epithelial Anticancer Activity

Luana K. S. Gonçalves[1*], Josinete B. Viera[1], Nayara S. R. Silva[1], César F. Santos[1], Francinaldo S. Braga[1], Josivan S. Costa[1,2], Williams J. C. Macêdo[2,3], Carlos H. T. P. Silva[4], Lorane I. S. Hage-Melim[1] and Cleydson Breno R. Santos[1,2,4*]

[1]Department of Biological Sciences and Health, Laboratory of Modeling and Computational Chemistry, Federal University of Amapá, Macapá, Brazil.
[2]Postgraduate Program in Biotechnology and Biodiversity-Network BIONORTE, Federal University of Amazonas, Manaus, Brazil.
[3]Laboratory of Molecular Modeling and Simulation System, Federal Rural University of Amazônia, Capanema, Pará, Brazil.
[4]Faculty of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, Brazil.

Authors' contributions

This investigation was performed in collaboration with all authors. Authors LKSG, JBV, FSB and CBRS designed the study, wrote the protocol, involved in writing the first draft, participated in data collection. Authors NSRS, CFS, JSC and WJCM managed the literature search, analyses of the study and manuscript preparation. Authors CHTPS, LISH and CBRS performed data interpretation and were actively involved in reading the manuscript. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJPR/2015/17914
Editor(s):
(1) Vasudevan Mani, Universiti Teknologi MARA (UiTM), Selangor, Malaysia.
(2) Ali Nokhodchi, Professor of Pharmaceutics and Drug Delivery, School of Life Sciences, University of Sussex, UK.
Reviewers:
(1) Gyula Oros, Plant Protection Institute HAS, Budapest, Hungary.
(2) Anonymous, Central University, New Delhi, India.
Complete Peer review History: http://www.sciencedomain.org/review-history.php?iid=1177&id=14&aid=9537

Original Research Article

Received 30th March 2015
Accepted 4th May 2015
Published 1st June 2015

ABSTRACT

Aims: To study and propose new caffeine derivatives with epithelial anticancer activity using quantum chemistry methods and multivariate analysis (PCA, HCA, PLS and PCR).
Place and Duration of Study: Laboratory of Modeling and Computational Chemistry at Federal University of Amapá (UNIFAP), Macapá, Brazil, between March 2014 and February 2015.

**Corresponding author: Email: luana.unifap@gmail.com, breno@unifap.br;*

Methodology: Caffeine and 31 derivatives with epithelial anticancer activity were selected from the literature, and modeled with the GaussView 3.0 program. The optimization was performed using the DFT method and B3LYP/6-31G** base set implemented in the Gaussian 03 program. Principal component analysis (PCA) and hierarchical cluster analysis (HCA) were employed to select the molecular descriptors related to epithelial anticancer activity. The Pearson correlation between activity and important descriptors were used for the regression partial least squares (PLS) and principal component regression (PCR) models built, and these models were used to predict the anticancer activity of fourteen new caffeine derivatives (test set) with unknown activity.
Results: The PCA results showed that the descriptors related to the compounds with anticancer activity were: area (A^2), distance radical 1 (dR1), distance radical 3 (dR3), radical partition coefficient 1 (logPR1) and radical partition coefficient 3 (logPR3). HCA showed similar results obtained with PCA, and the compounds were grouped in accordance with their biological activities. The results obtained with the PLS and PCR models were close, with variation between the models of R^2=±0.005, R^2ajust=±0.1998, s=±0.0053, $F_{(5,27)}$=±49.2261, Q^2=±0.012, SEV=±0.0117, PRESS=±0.0473 and S_{PRESS}=±0.0087. The PLS model showed that eight compounds of the test set (37 and 40-46) are predicted to be more active, and they had values of ICT_{50}<0.48mM. However, the PCR model only seven compounds of all test sets (33-36, 38, 39 and 42) were predicted as most active, which showed values of $ICT_{50}\geq$0.48mM, a total of 7 compounds proposed as less active of fourteen suggested compounds. However, compounds 37, 40, 41, 43, 44, 45 and 46 were the ones that had values of ICT_{50}<0.48mM in both PLS and PCR models, suggesting that these new compounds in the two models are more potent than caffeine may be tested for epithelial anticancer activity.
Conclusion: The PLS and PCR models showed good predictive ability. The test set showed for seven new caffeine compounds satisfactory results for epithelial anticancer activity. This strategy is fundamental for use in experimental syntheses and biological evaluation, and to understand the structural requirements for designing new ligands as anticancer agents.

*Keywords: Caffeine; epithelial cancer; molecular modeling; B3LYP/6-31G**; QSAR.*

1. INTRODUCTION

During the history man observed that plants used for therapeutic purposes. The choice of these plants for pharmaceutical applications were based more on empirical data that the scientific. Current events instigated a search for compounds that stand out from studies, by means of modifications to these structures, promoting better activity and reduced undesirable effects such as toxicity [1].

Caffeine (1,3,7-trimethylxanthine), see Fig. 1, is a chemical compound classified as alkaloid (basic character substance derived from plants containing in its formula, basically nitrogen, oxygen, hydrogen and carbon) pertaining to the group xanthine (organic substance, total nitrogenous, existing in the cardiac muscle, urine, and in various organs in some plants) [2]. Caffeine is found in some plants and used for consumption in beverages, as an infusion, as a stimulant, extremely soluble in hot water, and has no smell and has bitter taste [3].

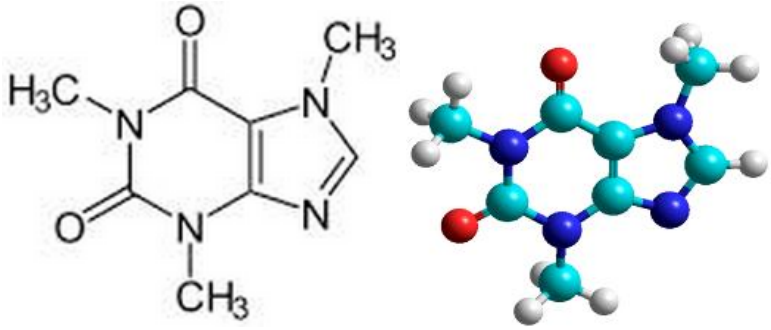

Fig. 1. Structural formula of caffeine (1,3,7-trimethylxanthine) in 2D and 3D format

Currently a variety of factors have motivated the search for new drugs of vegetable origin, among these stands out the discovery of drugs to combat cancer successfully [4]. Chaturvedi [5] relates that presently the antitumor action is the most extensively studied biological activity of plants, where studies have shown that they are able to combat tumors by selective alkylation, controlling and inhibiting cell division. These factors and cellular functions lead the cells to die by apoptosis. Cancer is called malignant neoplasm or malignant tumor, where the origin are due to genetic alterations can be oncogene activation, tumor suppressor genes inactivation, inactivation of genes responsible by apoptosis, mutations produced by chemical, physical and biological agents, and characterized by loss of function from the absence of differentiation, uncontrolled proliferation, invasiveness of adjacent tissues and metastasis [6-8].

On a worldwide there was an increase to 14.1 million new cases of different types of cancer in 2012, these 8.2 million deaths caused [9]. The types most frequently diagnosed were lung (13.0% of the total), breast (11.9%), and colon and rectum (9.7%). The most common determinants of death were lung (19.4% of total), liver (9.1%) and stomach (8.8%) cancers [10]. Through the need to eradicate this type of pathology, caffeine shows meaningful data that can come to combat the epithelial cancer. This pathology occurs in countries that possess favorable climatic conditions, where excessive sun exposure is the main risk factor for epithelial cancer. Among the countries with tropical climates stand out Brazil, followed by Australia [11,12].

Epithelial cancer better known as skin cancer is characterized by the uncontrolled growth of a group of cells of an organism, and this pathology is the result mainly of genetic alterations, environmental factors and lifestyle. This type of cancer is very incident in the world and the entire population is susceptible [13]. Epithelial tissue has several different layers and depending on where occurs at the cellular disorder cancer receives several different nomenclatures, such as: basocellular carcinoma, squamous cell carcinoma and malignant melanoma [14].

Ultraviolet radiation (UVR) is a complete carcinogen, and initiates the malignization process by mutations in the DNA and promotes the development of cancer by inflammatory process inherent to cumulative UVR exposure [15]. The UV rays contribute to the development of both forms of skin cancer: melanoma and non-melanoma. The non-melanocytic tumors are associated with cumulative sun damage, while melanocytic are intimately associated with intense episodes of excessive sun exposure, resulting in sunburn [16]. It is observed that skin manifestations present an evolutionary spectrum of appearance, in this order: burning, skin thickening, hyperchromic spots, fine lines, deep wrinkles and actinic [15].

The caffeine has stimulating activity on the central nervous system, cardiac muscle, and gastric acid secretion. Its activity is relatively beneficial when used in moderate doses, but in excess doses used, stimulates improperly the central nervous system, causing often accented frame insomnia and irritability [17,18]. For these reasons develop new studies for the use of caffeine. As well as there are new activities of other drugs on the market, such as thalidomide.

Thalidomide was first used in 1956 in Germany, through the therapeutic activity as a sedative [19]. Over the years she was gaining the market in 40 countries include Canada and the UK, but in the United States of America (USA) was not successful, the Food and Drug Administration (FDA) did not approve their entry for security criteria not many enlightened on drugs [20,21]. Due to the discovery that the drug has teratogenic characteristics in 1961, it was totally suspended from the world market [22,23]. Although the global scandal of thalidomide about their teratogenic activity, today it is being introduced in healing of erythema nodosum leprosum (ENL), and another very interesting activity to be mentioned is combat the multiple myeloma [24]. The study presented analogs that were effective against multiple myeloma, containing less side effects of drugs, and future research will be developed with nano basis for production of less toxic drugs that can combat cancer cells [25,26].

In this paper, a QSAR study of caffeine and 31 derivatives with epithelial anticancer activity (see Fig. 2) was performed. The structures were modeled and various different molecular descriptors were calculated with B3LYP/6-31G** method. The Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) were employed to select the molecular descriptors that are most related with biological activity investigated. A QSAR model was built with the Partial Least Square (PLS) and Principal

Component Regression (PCR) methods to carry out predictions of fourteen new caffeine compounds (test set) with unknown anticancer activity [27-30]. These predictions help in the interactions between molecules and their molecular targets, and to aid in future studies searching for other new anticancer drugs.

2. METHODOLOGY

The study was performed in Laboratory of Modeling and Computational Chemistry at Federal University of Amapá, Macapá, Brazil, between March 2014 and March 2015.

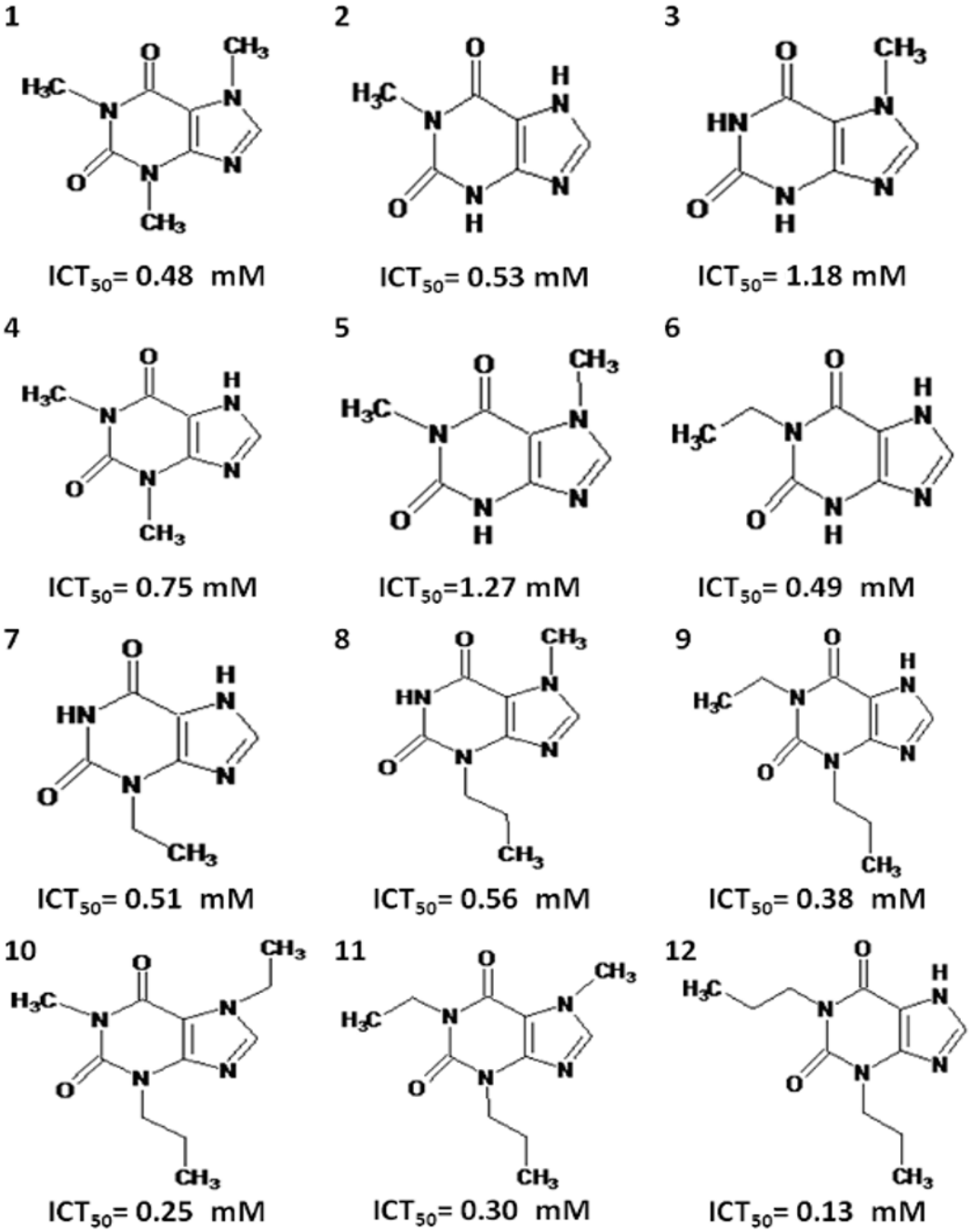

Fig. 2. Continuation

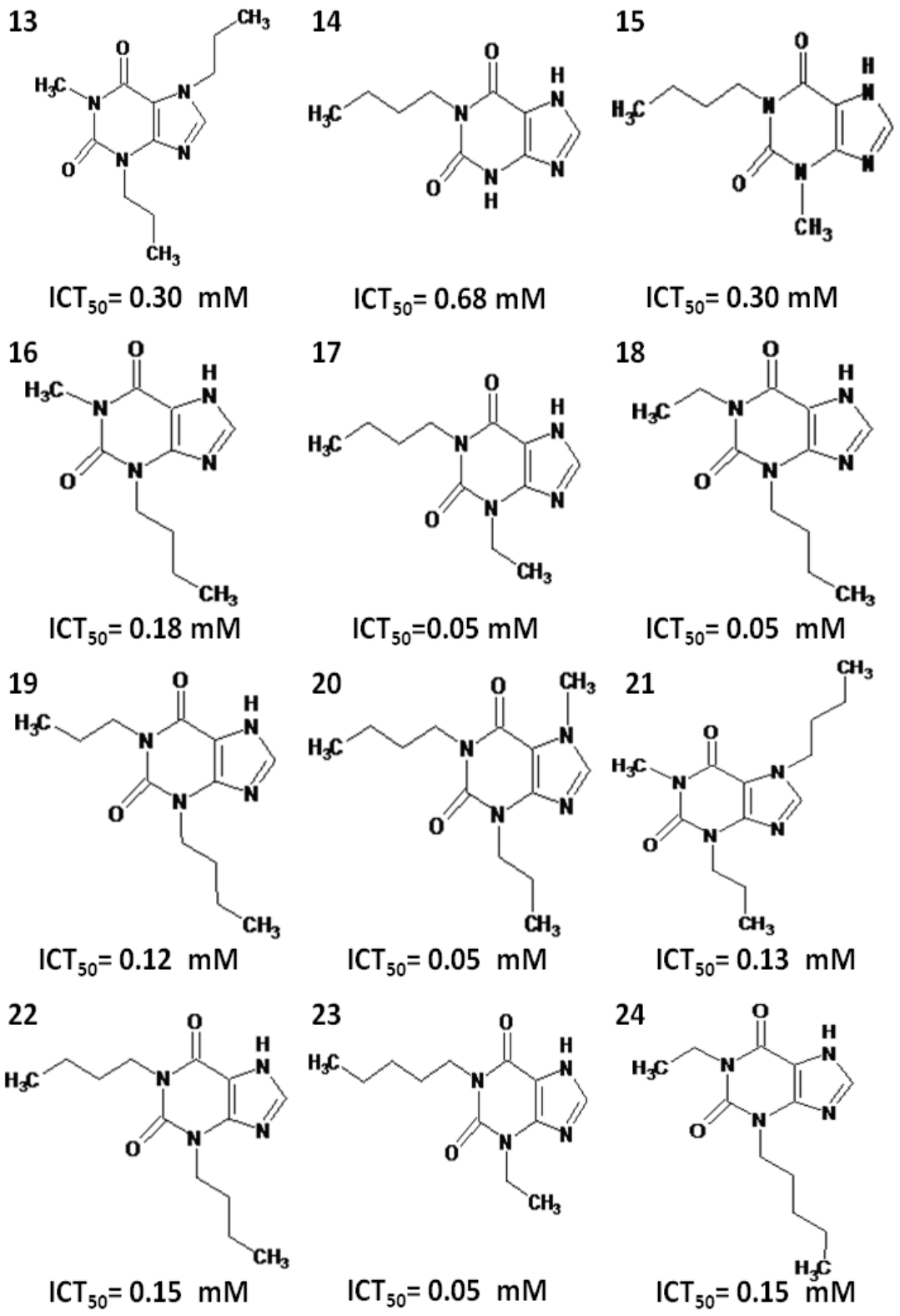

Fig. 2. Continuation

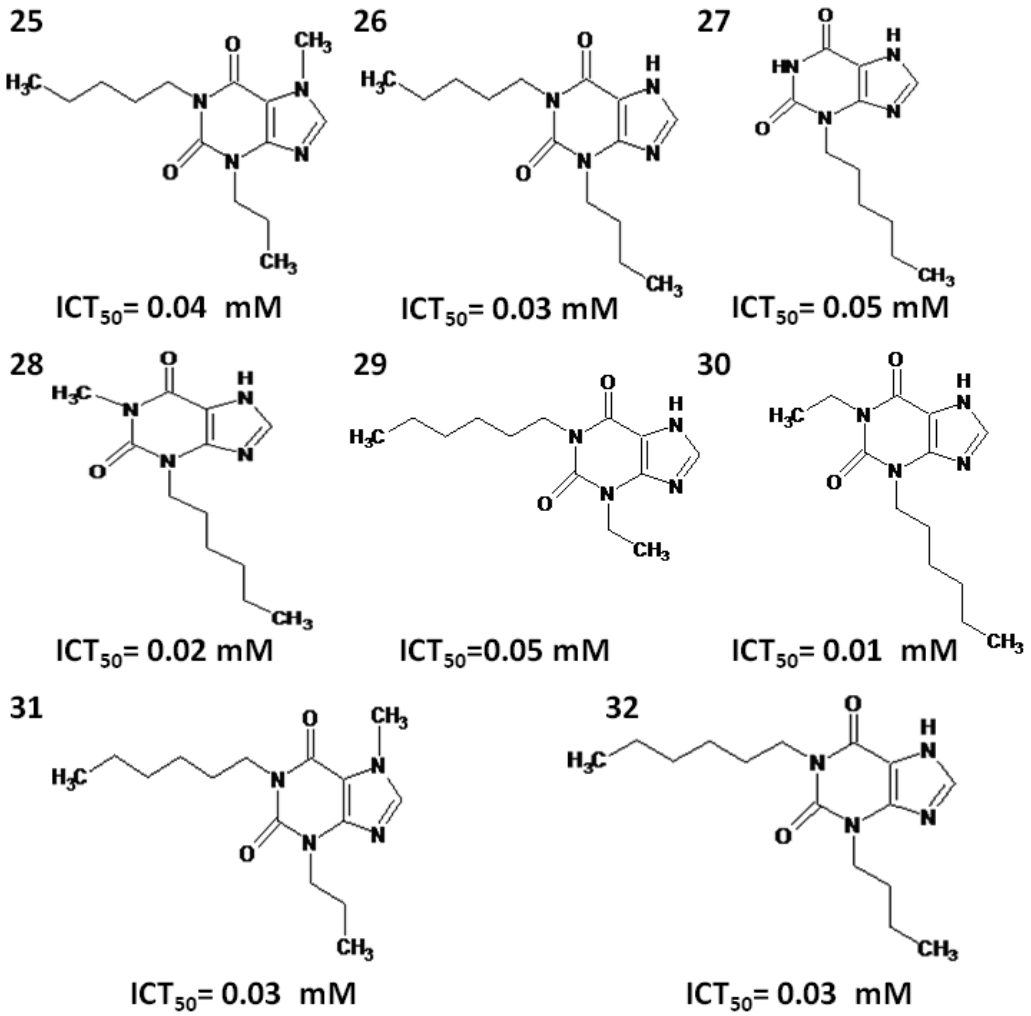

Fig. 2. Structure and biological activity of caffeine and its derivatives with epithelial anticancer activity

2.1 Compounds Studied with Epithelial Anticancer Activity

Initially, caffeine and 31 derivatives with epithelial anticancer activity were selected from the literature; see Fig. 2 [31]. The compounds were based on chemical structures, which differ in the length of the side chain alkyls (radicals). This size difference of the radicals, radical 1 (R1), radical 3 (R3) and radical 7 (R7), showed that the antitumor activity of these compounds suggested that inhibition of cell transformation to 1,3,7-trialkylxanthines depends on the number of carbon atoms for the alkyl group R1, R3 or R7. However, this study of Rogozin depicted that some xanthine analogues exert anticancer activity in epidermal cell cultures in animal models. Therefore, it is based on prevention of epidermal growth factor (EGF). All the compounds studied have been associated with *in vivo* activity against inducing malignant transformation in epidermal cells of rats JB6 susceptible to the development (P+) C141 (JB6 P+), based on the values of 50% inhibition of cell transformation (ICT_{50}) [31].

In this paper, the compounds were classified based on the anticancer responses adopted: compounds with ICT_{50}<0.48 mM, ranging from 0.0100 to 0.3800 mM, were assumed to be more potent compounds (9-13 and 15-32), and those with $ICT_{50}\geq$0.48 mM, ranging from 0.4800 to

1.2700 mM, were considered to be less potent compounds (1-8 and 14).

2.2 Molecular Modeling of Caffeine and Its Derivatives and Calculations of Descriptors or Molecular Properties

Compounds were modeled with GaussView 3.0 program [32], following the described strategy: initially the structure of caffeine was built and optimized with the DFT method and B3LYP/6-31G** base set implemented in Gaussian 03 program [33]. After obtaining of more stable geometry of caffeine, other compounds were built, and subsequently, obtained their most stable structures with lower energy in the same method and base set of caffeine.

The molecular descriptors are important for description of molecular structure and to finding appropriate predictive models [34]. The calculations of the molecular descriptors were performed employing the following software: Gaussian 03 program, Molekel [35] and HyperChem 6.02 [36].

In our study we calculated the chemical descriptors: total energy (TE), lowest unoccupied molecular orbital energy (LUMO), a level above the energy of the lowest unoccupied molecular orbital (LUMO+1), energy of the highest occupied molecular orbital (HOMO), a level below the energy of the highest occupied molecular orbital (HOMO-1), difference in energy between HOMO and LUMO (GAP=HOMO-LUMO), Mulliken electronegativity (χ), molecular softness ($1/\eta$), molecular hardness (η), torsion angle, dihedral angle, dipole moment (μ) and partial atomic charge. The atomic charges used in this study were obtained with the key word POP=CHELPG using the electrostatic potential [37], and atomic charges offer the general advantage of being physically more satisfactory than Mulliken charges [38]. In addition, the binding orders were calculated, distance radical 1 (dR1), distance radical 3 (dR3), distance radical 7 (dR7), hydrophobic descriptors such as the partition coefficient (logP), partition coefficient for radical 1 (logPR1), partition coefficient for radical 3 (logPR3), partition coefficient for radical 7 (logPR7) of all compounds studied. The molar refractivity, polarizability molar (MP), molar volume (MV), molar area, molar mass and hydration energy (HE) descriptors were calculated with the HyperChem 6.02 program.

2.3 Variable Selection and Construction of PLS and PCR Models

After determining all descriptors, a data matrix was built to start the multivariate analysis step. Multivariate analysis step was necessary to make the standardization or autoscale of the data matrix $X=(n, m)$, formed of thirty-two (32) lines (anticancer compounds) and seventy-six (76) columns (descriptors or molecular properties for each molecule), where n is the number of compounds and m is the number of variables.

The aim of standardized data matrix is to give equal weight to each variable in mathematical terms [39]. Variable selections were based on the analysis of the Pearson correlation matrix between variables and the epithelial anticancer activity (ICT_{50}). The descriptors or molecular properties with small or no correlation were discarded. After this analysis, two complementary methods were employed (PCA and HCA) to select the properties that contribute for classification of the compounds into two groups (more potent compounds and less potent compounds) [40-42].

PCA was used to reduce the dimensionality and find descriptors that could be useful in characterizing the behavior of the compounds with epithelial anticancer activity, and analyze the natural grouping of data and discrepant samples. During the execution of PCA, several attempts were made to get a good classification. The score plot gives information about the compounds (similarity and differences), and the loading plot gives information about the variables [29,30,42-46]. The descriptors selected by PCA were used to carry out HCA and for build of PLS and PCR models.

The aim of HCA is to show the compounds distributed in similarity, and the results should confirm of the PCA. The compounds become an agglomeration type, because each compound was first defined as its own cluster, and then others were grouped together to form new clusters until all the compounds were part of a single cluster [29,30,42-46].

The QSAR models were constructed by the PCR and PLS methods based on the autoscaled data and the leave-one-out cross validation procedure [29,30,42-46]. The statistical parameters used to evaluate the quality of the models were: Prediction Residual Error Sum of Squares (PRESS), Equation (1), Standard Error of

Validation (SEV), Equation (2), total variance explained, R^2 (correlation between the estimated values predicted by the model built with the full data set and actual values of y), Q^2 (the cross-validated correlation coefficient) and S_{PRESS} (standard deviation of cross-validation) given by Equations (3)–(5), respectively [29,30,42-46].

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (1)$$

$$SEV = \sqrt{\frac{PRESS}{n}} \quad (2)$$

$$R^2 = 1 - \left[\frac{PRESS_{cal}}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \right] \quad (3)$$

$$Q^2 = 1 - \left[\frac{PRESS_{val}}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \right] \quad (4)$$

$$S_{PRESS} = \frac{\sqrt{PRESS}}{n - k - 1} \quad (5)$$

In Equations (1) and (2), *n* is the number of compounds used for the validation model, y_i is the experimental data of the molecular properties for the sample and $\hat{y}_i$ is the value predicted by a calibration or validation model. In Equations (3) and (4), $PRESS_{cal}$ is the Calibration Prediction Error Sum of Squares and $PRESS_{val}$ is the Validation Prediction Error Sum of Squares. Both $PRESS_{cal}$ and $PRESS_{val}$ are evaluated from Equation (1) by changing $\hat{y}_i$ for a calibration or validation model. The values of explained variance (R^2 ajust, i.e., adjusted R^2), standard deviation (s) and F (Fisher test) were determined. The aim of the PLS and PCR methods were the constructions of a mathematical model that can be used to predict epithelial anticancer activity of fourteen new caffeine compounds (test set) with unknown anticancer activity. PCA, HCA, PLS and PCR were performed using the Pirouette 3.01 [47] and Statistica 6.2 [48] programs.

3. RESULTS AND DISCUSSION

3.1 Principal Component Analysis (PCA) Results

PCA results showed that the important molecular properties were the following: area (A^2), distance radical 1 (dR1), distance radical 3 (dR3), radical partition coefficient 1 (logPR1) and radical partition coefficient 3 (logPR3). They were selected from the data set of 76 molecular properties, and other variables were not selected because had a poor linear correlation with epithelial anticancer activity.

The values of the important molecular properties of each selected compound identified via PCA as well as the values of ICT_{50} (50% inhibition of cell transformation), are shown in Table 1. Table 1 shows the Pearson correlation matrix between the molecular properties and ICT_{50}, and the correlation between pairs of molecular properties is less than 0.9833, while the correlation between the molecular properties and ICT_{50} is more than -0.7885. The molecular properties selected by PCA represent the characteristics to separate between the more and less active with epithelial anticancer activity of the compounds.

The results of the PCA model are presented in Table 2. The model was built with three main components (3 PCs). The first principal component (PC1) describes 92.1316% of the total information, the second principal component (PC2) describes 58.8150%, and the third (PC3) 2.5021%. PC1 contains 59.4397% of the original data, and the combination of the firts two components (PC1+PC2) contains 97.3850%, and all three (PC1+PC2+PC3) explain 98.9992% of the total information, losing only 1.008 of the original information. The dR3 (0.4261), logPR3 (0.4396) and A^2 (0.5642) descriptors contribute to PC1, while in PC2, the main contributors are the dR1(0.5190) and logPR1 (0.5283) descriptors.

The main components can be written as a linear combination of the selected descriptors. Mathematical expressions for PC1 (6) and PC2 (7) are shown below:

$$PC1 = 0.3963dR1 + 0.4261dR3 + 0.3869logPR1 + 0.4396logPR3 + 0.5642Â2 \quad (6)$$

$$PC2 = 0.5190dR1 - 0.4866dR3 + 0.5283logPR1 - 0.4632logPR3 + 0.0015Â2 \quad (7)$$

Table 1. Molecular properties selected by principal component analysis, experimental ICT_{50} (mM) values and the Pearson correlation matrix

Compounds	dR1	dR3	logPR1	logPR3	A^2	ICT_{50} (mM)
1-	2.1060	2.1040	1.0900	1.0900	359.0799	0.4800
2-	2.1090	0.0000	1.0900	0.0000	237.9100	0.5300
3-	1.0140	1.0110	0.0000	0.0000	312.1600	1.1800
4-	2.1080	2.1060	1.0900	1.0900	337.6099	0.7500
5-	2.1080	1.0110	1.0900	0.0000	337.4100	1.2700
6-	3.4390	0.0000	1.3000	0.0000	324.3999	0.4900
7-	1.0140	3.4340	0.0000	1.3000	333.1000	0.5100
8-	1.0140	4.6730	0.0000	1.6900	386.5899	0.5600
9+	3.4380	4.6730	1.3000	1.6900	411.2399	0.3800
10+	2.1080	4.6730	1.0900	1.6900	436.8900	0.2500
11+	3.4380	4.6740	1.3000	1.6900	435.5799	0.3000
12+	4.6770	4.6740	1.6900	1.6900	449.4800	0.1300
13+	2.1080	4.6740	1.0900	1.6900	468.4800	0.3000
14-	5.9860	0.0000	2.0900	0.0000	388.3399	0.6800
15+	5.9850	2.1050	2.0900	1.0900	422.6600	0.3000
16+	2.1080	5.9820	1.0900	2.0900	422.6400	0.1800
17+	5.9850	3.4350	2.0900	1.3000	445.6099	0.0500
18+	3.4380	5.9820	1.3000	2.0900	446.0599	0.0500
19+	4.6760	5.9810	1.6900	2.0900	477.6000	0.1200
20+	5.9860	4.6740	2.0900	1.6900	497.5499	0.0500
21+	2.1080	4.6730	1.0900	1.6900	501.8099	0.1300
22+	5.9840	5.9810	2.0900	2.0900	508.4899	0.1500
23+	7.2450	3.4340	2.4900	1.3000	476.7999	0.0500
24+	3.4380	7.2420	1.3000	2.4900	476.6600	0.1500
25+	7.2460	4.6740	2.4900	1.6900	533.3400	0.0400
26+	7.2450	5.9820	2.4900	2.0900	541.2100	0.0300
27+	1.0140	8.5410	0.0000	2.8800	459.5100	0.0500
28+	2.1090	8.5420	1.0900	2.8800	484.2000	0.0200
29+	8.5460	3.4350	2.8800	1.3000	510.7399	0.0500
30+	3.4380	8.5410	1.3000	2.8800	511.2000	0.0100
31+	8.5460	4.6740	2.8800	1.6900	564.5700	0.0300
32+	8.5440	5.9830	2.8800	2.0900	573.6500	0.0300
dR1		0.0283	0.9635	0.0660	0.6515	-0.4976
dR3			0.0056	0.9833	0.6988	-0.6986
logPR1				0.0523	0,6302	-0.4984
logPR3					0.7114	-0.7574
A^2						-0.7885

Fig. 3 shows the scores for the 32 compounds studied. According to the graph, PC1 separates the compounds between more and less potent. The most potent compounds are shown in the right part (9-13 and 15-32), while the less potent compounds are in the left side of the graph (1-8 and 14).

Fig. 4 shows the loadings for the five (5) descriptors that are important for the classification of compounds. The most potent compounds have high contributions of the A^2 and logPR3 descriptors, while less potent compounds has a high contribution of the dR1, dR3 and logPR1 descriptors. Thus, the dR1, dR3 and logPR1 descriptors are responsible for the location of less potent compounds at the left side of the graph. The A^2 descriptor represents the more potent compounds in the right part of the graph. Fig. 3 also shows that the higher the contribution of the A^2 and logPR3 descriptors in the first principal component, i.e., the higher value for a certain compound, the higher score value will be, indicating that the compound is more potent than others.

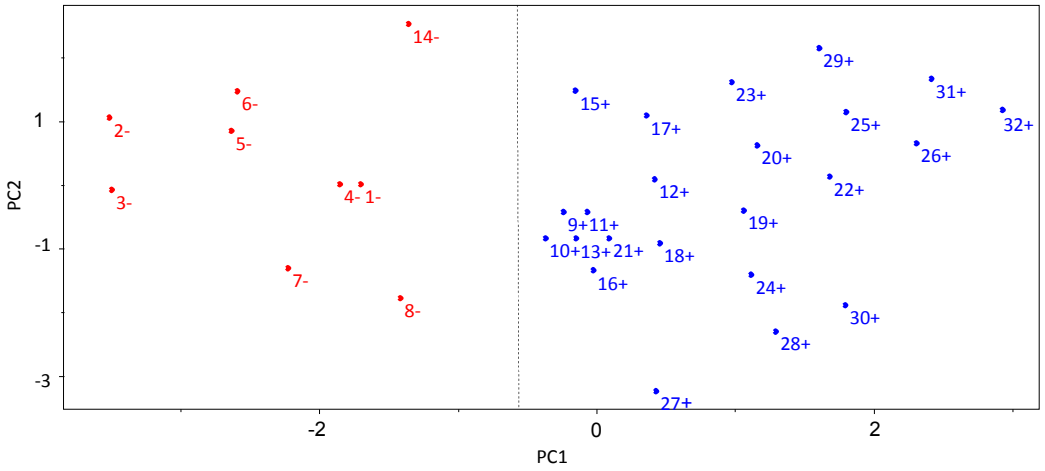

Fig. 3. Plot of PC1–PC2 scores for caffeine and derivatives with epithelial anticancer activity. Positive values indicate more potent analogs (in blue) and negative values indicate less potent analogs (in red)

3.2 Hierarchical Cluster Analysis (HCA) Results

The statistical analysis used in this study should group similar compounds into categories. The categories are represented by a two-dimensional diagram known as dendrogram which illustrates the fusions or division made in each successive analysis phase. Samples (compounds) are represented by the branches in the lower part of the dendrogram. The similarity between the groups is given by the length of its branches, thus compounds with low similarity have long branches, whereas compounds of high similarity have short branches. HCA method classified the compounds in two classes (more active and less active), and was based on the Euclidean distance and the complete method [49,50]. The descriptors used to carry out HCA were the same as for the PCA method, that is, area (A^2), distance radical 1 (dR1), distance radical 3 (dR3), radical partition coefficient 1 (logPR1) and radical partition coefficient 3 (logPR3).

The dendrogram in Fig. 5 shows the HCA result as well as the compounds separated into two classes. The scale of similarity varies from 0 for samples with no similarity and 1 for samples with identical similarity. When analyzing the dendrogram, some conclusions can be made, although the compounds show some structural diversity.

HCA showed similar results obtained by PCA. Compounds were grouped according to their biological activities. The most potent compounds are 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31 and 32. The less potent compounds are 1, 2, 3, 4, 5, 7, 8, e 14. Compound 30 has the lowest value of ICT_{50}=0.0100, among the compounds classified as most potent. Whereas, the compound 5 has the highest value of ICT_{50}= 1.2700, while the variation between the activities of the compounds 5 and 30 is ±1.2600 between them.

3.3 Partial Least Squares (PLS) and Principal Component Regression (PCR) Results

The statistical quality [49] of the PLS and PCR models was calibrated by parameters: squared correlation coefficient (R^2), explained variance (R^2 ajust), standard deviation (s), variance ratio (F), cross-validated correlation coefficient (Q^2), standard error of validation (SEV), predicted residual error sum of squares (PRESS) and standard deviation of cross-validation (S_{PRESS}) [50–52]. The best regression models were selected based on high values of R^2, R^2 ajust, Q^2 and F (a statistic of assessing the overall significance) and low values of s, SEV, PRESS and S_{press}.

The calculated properties and the experimental values for the compounds were used to build the PLS and PCR models (see Table 3). The models built were based on three latent variables and 32 compounds. The equations obtained for PLS (Equation (8)) and PCR (Equation (9)) models are the following:

$$ICT_{50} = -0.1685dR1 - 0.2366dR3 - 0.1688logPR1 - 0.2564logPR3 - 0.2670Â2 \quad (8)$$

n = 32, R^2 = 0.9260, R^2_{ajust} = 0.9157, s = 0.1587, $F_{(5,27)}$ = 65.0702, Q^2 = 0.8831, SEV = 0.1867, PRESS=0.7056, S_{PRESS} = 0.0191.

$$ICT_{50} = -0.1955dR1 - 0.2102dR3 - 0.1908logPR1 - 0.21681logPR3 - 0.2783Â2 \quad (9)$$

n = 32, R^2 = 0.9210, R^2_{ajust} = 0.7159, s = 0.1640, $F_{(5,27)}$ = 15.8441, Q^2 = 0.8951, SEV=0.1750, PRESS=0.7529, S_{PRESS} = 0.0218.

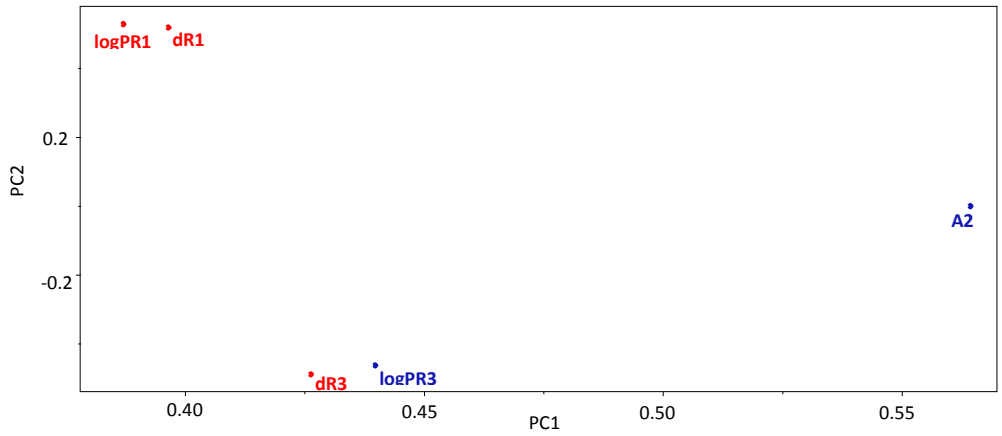

Fig. 4. Plot of the PC1–PC2 loadings with the five descriptors selected to build the PLS and PCR models of caffeine and derivatives with biological activity against epithelial cancer

Table 2. Principal component analysis of the SAR model and contribution of selected descriptors based on step multivariate analysis

Parameters	Main component		
	PC1	PC2	PC3
Variance (%)	92.1316	58.8150	2.5021
Cumulative variance (%)	59.4397	97.3850	98.9992
Molecular descriptors	Contribution		
		PC1	PC2
dR1		0.3963	0.5190
dR3		0.4261	-0.4866
logPR1		0.3869	0.5283
logPR3		0.4396	-0.4632
A^2		0.5642	0.0015

The results obtained with the PLS and PCR models were close, with variation between PLS and PCR of R^2 = ±0.005, R^2 ajust = ±0.1998, s = ±0.0053, $F_{(5,27)}$ = ±49.2261, Q^2 = ±0.012, SEV = ±0.0117, PRESS = ±0.0473 and S_{PRESS} = ±0.0087. The quality of the models was demonstrated by comparing the measured and the predicted activities. The validation errors obtained by the leave-one-out cross-validation method are shown in Table 3. For the PLS model, only four compounds (2, 3, 5, 17 and 32) had high validation errors, and the PCR model yielded five compounds (2, 3, 5, 27 and 32) with high residual values.

The measured versus predicted values using our PLS and PCR models are presented in Figs. 6a,b, respectively. The PLS and PCR plots identify compounds with higher activity (blue) and compounds with lower activity (red). The validation parameters support the fact that the models are efficient and hence satisfactory given the complexity of the anticancer mechanisms and the small number of descriptors (five) selected to build the QSAR model.

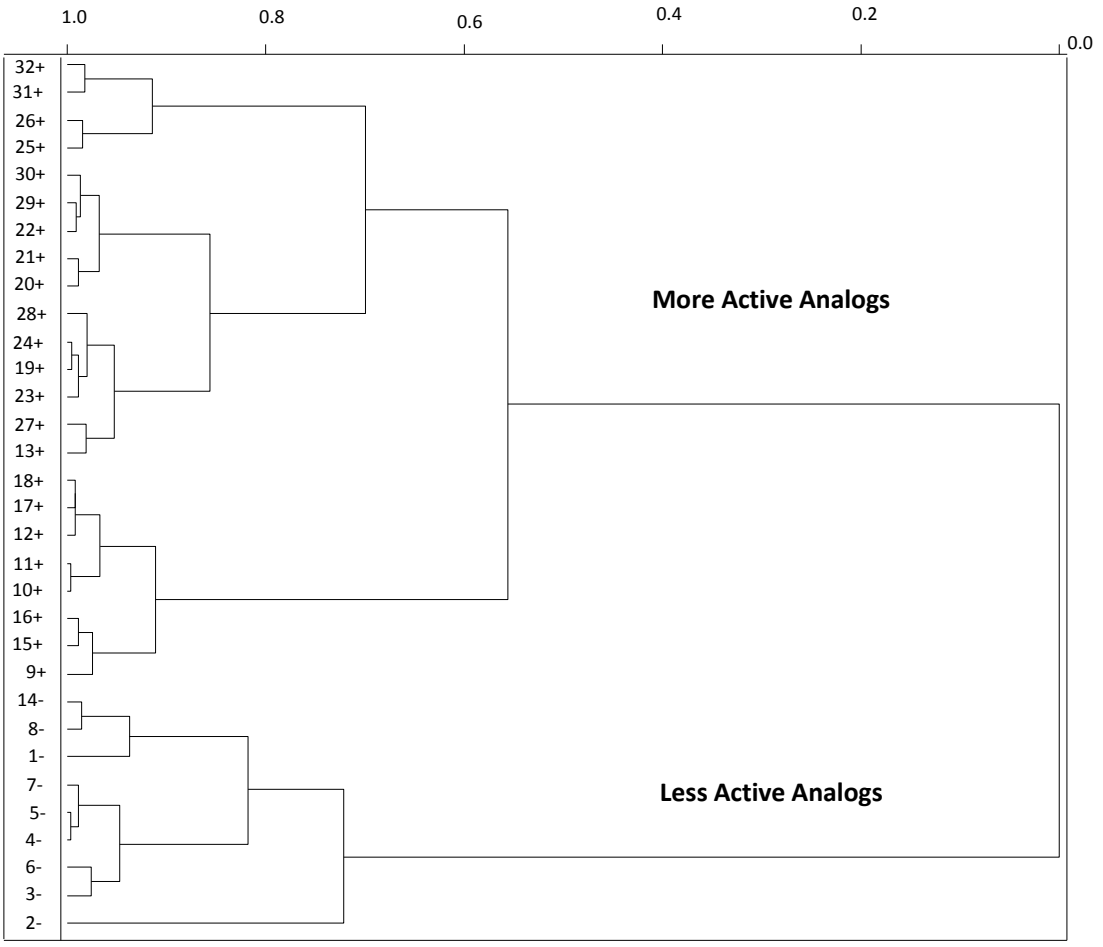

Fig. 5. HCA dendrogram for caffeine and derivatives with epithelial anticancer activity. Positive values indicate more potent analogs and negative values indicate less active compounds

3.4 Application of PLS and PCR Models to the Compounds of the Test Set

The compounds of the test set were molded from the most stable structure of caffeine using GaussView 3.0 program. The optimization of the geometry of each compound was performed with DFT method and the basis set of separated valence 6-31G** by Gaussian 03 program. After obtain the most stable geometry of each compound was determined only selected descriptors in PCA and used in the building of the QSAR (PLS and PCR) models, namely area (A^2), distance radical 1 (dR1), distance radical 3 (dR3), radical partition coefficient 1 (logPR1) and radical partition coefficient 3 (logPR3), shown in Table 4.

The QSAR models (PLS and PCR) were used to predict the unknown epithelial anticancer activity of fourteen new caffeine derivatives shown in Fig. 7, compounds 33-46. Table 5 shows the results of the ICT_{50} by PCR and PLS models. According to Table 5 the PLS model showed that eight compounds of the test set (37 and 40-46) are predicted to be more active, they had values of ICT_{50}<0.48 mM. However, the PCR model only seven compounds of all test sets (33-36, 38, 39 and 42) were predicted as less active, which showed values of $ICT_{50}\geq$0.48 mM, a total of 7 compounds proposed as more active of fourteen suggested compounds. However, compounds 37, 40, 41, 43, 44, 45 and 46 were the ones that had values of ICT_{50}<0.48 mM in both models (PLS and PCR), suggesting that these new compounds in the two models are more potent than caffeine may be synthesized and tested for epithelial anticancer activity.

Table 3. Predicted PLS and PCR results and validation errors for ICT_{50} (experimental)

Compounds	Predicted		Validation error		Experimental
	PLS	PCR	PLS	PCR	ICT_{50} (mM)
1-	0.5700	0.5962	-0.0900	-0.1162	0.4800
2-	0.9583	0.8770	0.4283	-0.347	0.5300
3-	0.7892	0.8167	0.3908	0.3633	1.1800
4-	0.5756	0.6193	0.1744	0.1307	0.7500
5-	0.6628	0.7259	0.6072	0.5441	1.2700
6-	0.7760	0.6858	-0.286	-0.1958	0.4900
7-	0.6375	0.7065	-0.1275	-0.1965	0.5100
8-	0.4830	0.4930	0.0770	0.0670	0.5600
9+	0.3180	0.3602	0.0620	0.0198	0.3800
10+	0.3404	0.3238	-0.0904	-0.0738	0.2500
11+	0.2935	0.3338	0.0665	-0.0338	0.3000
12+	0.2281	0.1966	-0.0981	-0.0666	0.1300
13+	0.3027	0.3596	-0.0027	-0.0596	0.3000
14-	0.5399	0.5213	0.1401	0.1587	0.6800
15+	0.3431	0.3125	-0.0431	-0.0431	0.3000
16+	0.2738	0.2596	-0.0938	-0.0796	0.1800
17+	0.2617	0.2343	-0.2117	-0.1843	0.0500
18+	0.2061	0.1810	-0.1561	-0.1310	0.0500
19+	0.1111	0.1457	0.0008	-0.0257	0.1200
20+	0.1203	0.0752	-0.0703	-0.0252	0.0500
21+	0.2730	0.3236	-0.1430	-0.1936	0.1300
22+	0.0140	-0.0189	0.1360	0.1689	0.1500
23+	0.1711	0.1238	-0.1211	-0.0738	0.0500
24+	0.0798	0.0648	0.0702	0.0852	0.1500
25+	0.0222	-0.0021	0.0178	0.0421	0.0400
26+	-0.0768	-0.1208	0.1068	0.1508	0.0300
27+	0.1925	0.3046	-0.1425	-0.2546	0.0500
28+	0.0415	0.0265	-0.0215	0.0065	0.0200
29+	0.0727	0.0100	-0.0227	0.0400	0.0500
30+	-0.0403	-0.0562	0.0503	0.0662	0.0100
31+	-0.0774	-0.1130	0.1074	0.1430	0.0300
32+	-0.1851	-0.2226	0.2151	0.2526	0.0300

Table 4. Molecular properties selected by analysis of main components of test set with anticancer activity unknown

Compounds	dR1	dR3	logPR1	logPR3	A^2
33	1.0150	0.0000	0.0000	0.0000	279.2799
34	1.0140	2.1050	0.0000	1.0900	312.9500
35	1.0140	2.1050	0.0000	1.0900	338.0199
36	3.4370	2.0750	1.3000	1.0900	357.7099
37	3.4370	3.4340	1.3000	1.3000	381.5599
38	4.6770	0.0000	1.6900	0.0000	352.5299
39	1.0140	4.6730	0.0000	1.6900	365.8500
40	4.6770	2.1050	1.6900	1.0900	385.9500
41	2.1080	4.6730	1.0900	1.6900	408.0700
42	1.0140	4.6730	0.0000	1.6900	416.3999
43	1.0140	4.6730	0.0000	1.6900	443.4800
44	1.0140	5.9820	0.0000	2.0900	394.4899
45	7.2450	0.0000	2.4900	0.0000	415.9299
46	8.5440	0.0000	2.8800	0.0000	450.9700

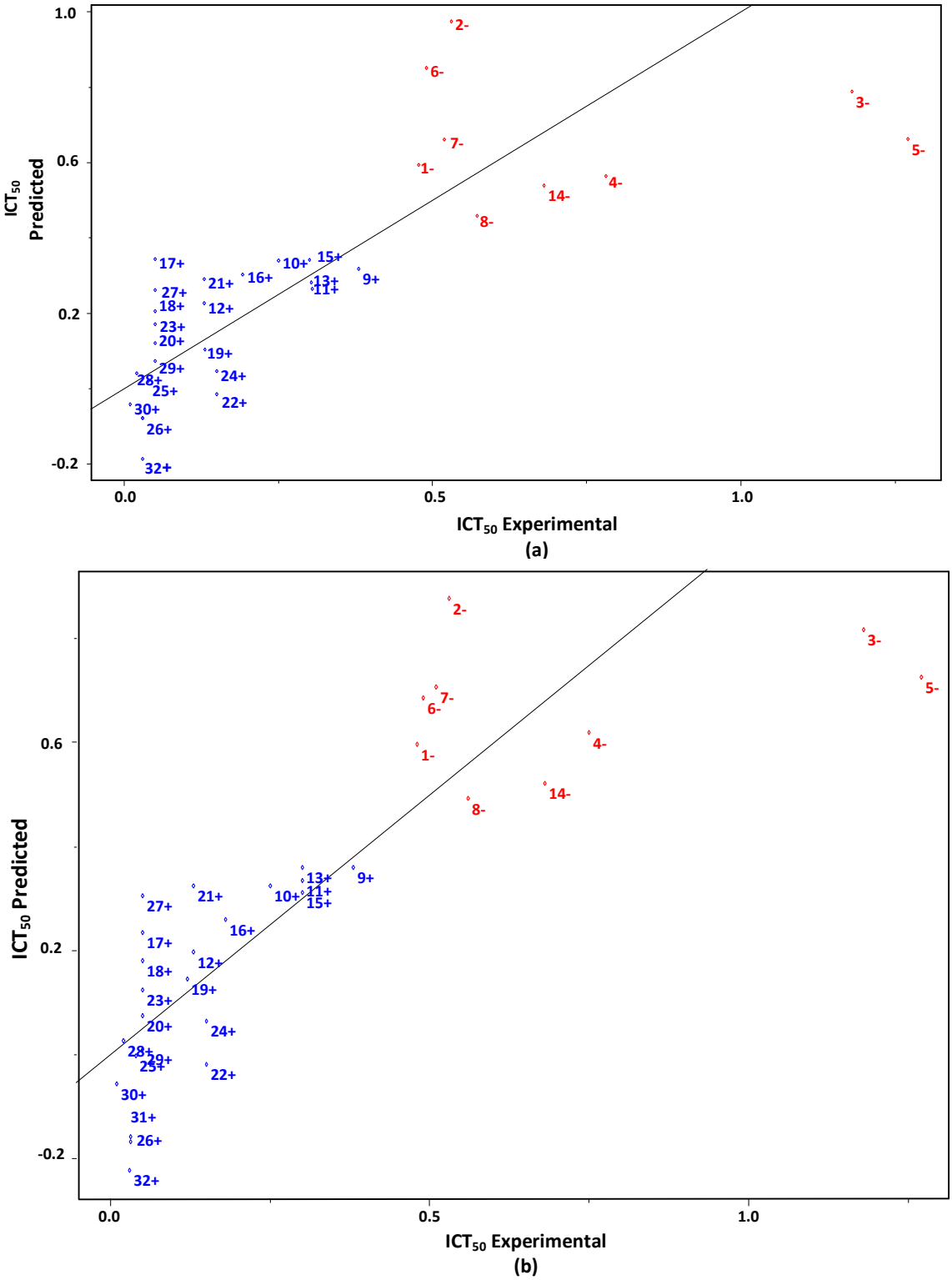

Fig. 6. Plot of experimental versus predicted values for ICT_{50} modeled by (a) PLS and (b) PCR

Table 5. Anticancer activity predicted (ICT_{50}) by PCR and PLS models for the test set compounds and residues of prediction between models

Test set compounds	Predicted (ICT_{50})		Residues of prediction
	PLS	PCR	(PLS-PCR)
33	0.9261	0.9179	0.0082
34	0.7112	0.7261	-0.0149
35	0.6840	0.6978	-0.0138
36	0.5229	0.5156	0.0073
37	0.4318	0.4317	0.0001
38	0.6521	0.6127	0.0394
39	0.5101	0.5412	-0.0311
40	0.4375	0.4212	0.0163
41	0.3674	0.3832	-0.0158
42	0.4552	0.4840	-0.0288
43	0.4258	0.4534	-0.0276
44	0.3962	0.4370	-0.0408
45	0.4725	0.4140	0.0585
46	0.3794	0.3112	0.0682

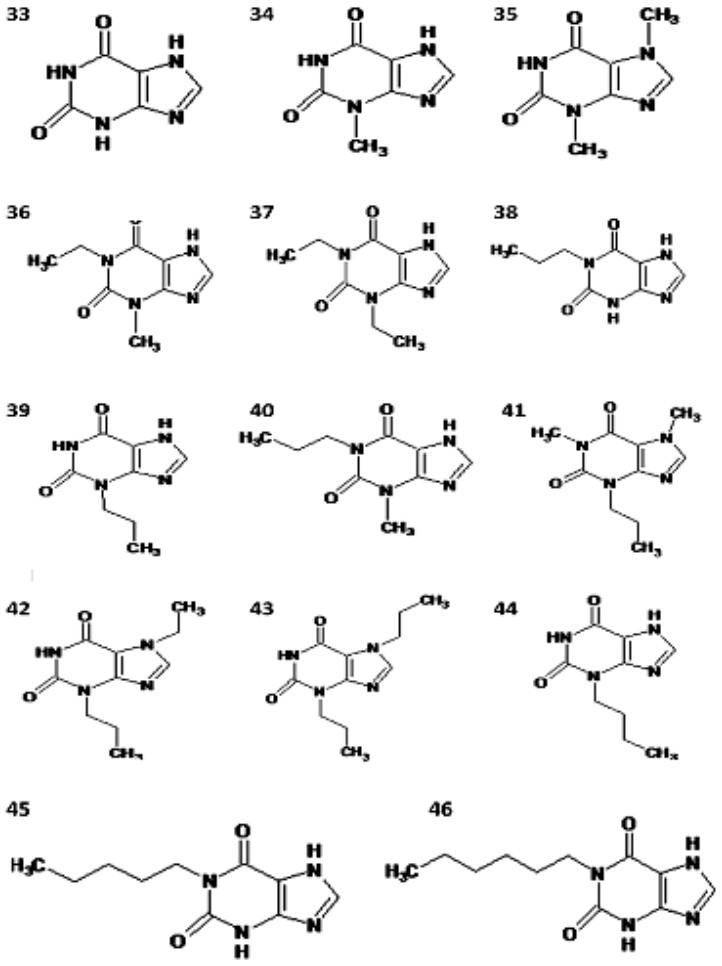

Fig. 7. Compounds of the test set caffeine derivatives with unknown anticancer activity against epithelial cancer

4. CONCLUSION

The DFT method and B3LYP/6-31G** basis set revealed themselves to be adequate to optimize the structures of caffeine and derivatives for subsequent study. PCA and HCA methods classified the compounds into groups according to their degree of epithelial anticancer activity. The descriptors dR1, dR3, logPR1, logPR3 and Â2 were responsible for distinguishing compounds with higher and lower activity. PLS and PCR models obtained showed good predictive ability. The test set showed for seven new caffeine compounds satisfactory results for epithelial anticancer activity. This strategy is fundamental for use in experimental syntheses and biological evaluation, and to understand the structural requirements for designing new ligands as anticancer agents. Consequently, further studies need be done to evaluate the different proposals as well as their actions, toxicity, and potential use for treatment of epithelial cancer.

CONSENT

It is not applicable.

ETHICAL APPROVAL

It is not applicable.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support provided by the Brazilian Agency National Council of Scientific and Technological Development (CNPq-Brazil). The authors would like to thank the Scientific Initiation Program (IC/CNPq/UNIFAP), and the Laboratory of Modeling and Computational Chemistry of Federal University of Amapá for computational support.

COMPETING INTERESTS

The authors confirm that this article content has no conflicts of interest and part of the methodology of this article has been previously published in *Molecules* 2014, 19 (8), 10670-10697; doi:10.3390/molecules190810670.

REFERENCES

1. Montanari CA, Bolzani VS. Planejamento racional de fármacos baseado em produtos naturais. Quim. Nova. 2001; 24(1):105-111.
2. Alves AB. Determinação simultânea de teobromina, teofilina e cafeína em chás por cromatografia líquida de alta eficiência. Revista Brasileira de Ciências Farmacêuticas. 2002;38(2):237-243.
3. Altimari LR, Moraes AC, Tirapegui J, Moreau RLM. Cafeína e performance em exercícios anaeróbios. Revista Brasileira de Ciências Farmacêuticas. São Paulo. 2006;42(1):18-27.
4. Carvalho JE. Fitoterápicos: Alimento ou Medicamento? Ciência de Alimentos: avanços e perspectivas, Faculdade de Engenharia de Alimentos da Unicamp: Campinas. 2001;2:196-202.
5. Chaturvedi D. Sesquiterpene lactones: Structural diversity and their biological activities. Opp. Chal. and Sco. of Nat. Prod. in Med. Chem. 2011;2:313-334.
6. Rosenberg SA. Progress in human tumour immunology and immunotherapy. Nature. 2001;411(6835):380-384.
7. Cairns J. The origin of human cancers. Nature. 1981;289:353-357.
8. Brentani RR, Chammas R, Coelho FRG. Mecanismos de invasão e metástases. Bases da Oncologia. 1998;91-98.
9. Ferlay J, Soerjomataram I, Ervik, M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: iarc cancer base. International Journal of cancer. 2010;1(11).
10. World Health Organization, Skin cancer. 2009;51.
11. Prolla PA. Síndromes de suscetibilidade hereditaria ao câncer de pele. Genética Molecular do Câncer. 2002;213-222.
12. Friedman R, Sherman Jr CD. Câncer de pele. Manual de Oncologia Clínica. 1999; 245-253.
13. Popim RC, Corrente JE, Marino JAG, Souza CA. Câncer de pele: uso de medidas preventivas e perfil demográfico de um grupo de risco na cidade de Botucatu. Ciência & Saúde Coletiva. 2008; 13(4):1331-1336.
14. Almeida Jr C. Farmacêuticos em oncologia: uma nova realidade. São Paulo: Atheneu; 2004.
15. Bergfeld WF. The aging skin. Int. J. Fertil womens Med. 1997;42(2):57-66.
16. Armmings SR, Tripp MK, Hermann NB. Approaches to the prevention and control

of skin cancer. Cancer Metastasis Rev. 1997;16(3-4):309-322.
17. Carper J. Alimentos: o melhor remédio para a boa saúde. 4a ed. Campus; 1995.
18. Longo EN, Navarro ET. Manual dietoteráripo. 2º ed. Artmed; 2001.
19. Mujagiae H, Chabner BA, Mujagiae Z. Mechanisms of action and potential therapeutic uses of Thalidomide. Croat Med J. 2002;43:274-85.
20. Diggle GE. Thalidomide: 40 years on. Int J Clin Pract. 2001;55:627-31.
21. Gasparini G, Morabito A, Magnani E, Gattuso D, Capaccetti B, Alberti AM. Thalidomide: an old sedative-hypnotic with anticancer activity? Curr Opin Investig Drugs. 2001;2:1302-8.
22. Lenz W. Malformations caused by drugs in pregnancy. Am J Dis Child. 1966;112:99-106.
23. Lenz W. A short history of thalidomide embryopathy. Teratology. 1988;38:203-15.
24. Hales BF. Thalidomide on the comeback trail. Nat Med. 1999;5:489-90.
25. Aitken RA, Parker D, Taylor RJ, Gopal J, Kilenyi RN. Asymmetric synthesis, Blackie Academic & Professional; 1992.
26. Rajkumar SV. Thalidomide in the treatment of multiple myeloma. Expert Rev Anticancer Ther. 2001;1:20-8.
27. Gramatica P. Principles of QSAR models validation: Internal and external. QSAR Comb. Sci. 2000;26:694-701.
28. Geladi P, Kowalski BR. Partial least-squares regression: A tutorial. Anal. Chim. Acta. 1986;185:1-17.
29. Geladi P. Notes on the history and nature of partial least squares (PLS) modeling. J. Chemom. 1988;2:231-246.
30. Ferreira MMC. Multivariate QSAR. J. Braz. Chem. Soc. 2002;13:742-753.
31. Rogozin EA, Lee KW, Kang NJ, Yu H, Nomura M, Miyamoto KI, Conney AH, Bode AM, Dong Z. Inhibitory effects of caffeine analogues on neoplastic transformation: structure-activity relationship. Carcinogenesis. 2008;29(6): 1228-1234.
32. GaussView 3.07; Gaussian Inc.: Pittsburgh, PA, USA; 1997.
33. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery Jr JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda, R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD Strain MC; Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA. Gaussian 03, Revision C.02, Gaussian, Inc. Wallingford, CT; 2004.
34. Estrada E, Molina E. Novel local (fragment-based) topological molecular descriptors for QSPR/QSAR and molecular design. J. Mol. Graph. Model. 2001;20: 54–64.
35. Molekel 4.3; Swiss Center for Scientific Computing: Manno, Switzerland; 2000.
36. ChemPlus, Modular Extensions to HyperChem, Release 6.02; Molecular Modeling for Windows,Hyper, Inc.: Gainesville, FL, USA; 2000.
37. Breneman CM, Winberg KB. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. J. Comput. Chem. 1990;11:361–373.
38. Singh UC, Kollman PA. An approach to computing electrostatic charges for molecules. J. Comput. Chem. 1984;5: 129–145.
39. Santos CBR, Vieira JB, Lobato CC, Hage-Melim LIS, Souto, RNP, Lima CS, Costa EVM. Brasil DSB, Macêdo WJC, Carvalho JCT. A SAR and QSAR Study of new artemisinin compounds with antimalarial activity. Molec. 2014;19:367–399.
40. Santos CBR, Vieira JB, Formigosa AS, Costa EVM, Pinheiro MT, Silva JO, Macêdo WJC, Carvalho JCT. Validation of computational methods applied in molecular modeling of artemisinin with antimalarial activity. J. Comput. Theor. Nanosci. 2014;11:553–556.
41. Vieira JB, Braga FS, Lobato CC, Santos CF, Costa JS, Bittencourt JAHM, Brasil DSB, Hage-Melim LIS, Macêdo WJC, Carvalho JCT, Santos CBR. A QSAR,

pharmacokinetic and toxicological study of new artemisinin compounds with anticancer activity. Molec. 2014;19:10670-10697.
42. Santos CBR, Lobato CC, Sousa MAC, Macêdo WJC, Carvalho JCT. Molecular modeling: Origin, fundamental concepts and applications using structure-activity relationship and quantitative structure-activity relationship. Rev. Theor. Sci. 2014; 2:91–115.
43. Santos CBR, Lobato CC, Vieira JB, Brasil DSB, Brito AU, Macêdo WJC, Carvalho JCT, Pinheiro JC. Evaluation of quantum chemical methods and basis sets applied in the molecular modeling of artemisinin. Comp. Mol. Biosc. 2013;3:66–79.
44. Pimentel M, Neto BB. Calibração: Uma Revisão para Químicos Analíticos. Quim. Nova. 1996;19:268–277.
45. Custódio R, Andrade JC. Augusto FO. Ajuste de funções matemáticas a dados experimentais. Quim. Nova. 1997;20:219–225,
46. Gaudio AC, Zandonade E. Proposição, validação e análise dos modelos que correlacionam estrutura química e atividade biológica. Quim. Nova. 2001;24: 658–671.
47. Pirouette Software, Version 3.01, Infometrix Inc.: Seattle, WA, USA; 2001
48. STATISTICA (data analysis software system); version 6.2, StatSoft, Inc.; 2004. Available: www.statsoft.com
49. Snedecor GW, Cochran WG. Statistical Methods; Oxford and IBH: New Delhi, India; 1967.
50. Chatterjee S, Hadi AS, Price B. Regression analysis by examples, 3rd ed.; Wiley VCH: New York, NY, USA; 2000.
51. Diudea MV. QSPR/QSAR studies for molecular descriptors; Nova Science: Huntingdon, NY, USA; 2000.
52. Bikash D, Shovanlal G, Subrata B, Soma S, Tarun J. QSAR study on some pyridoacridine ascididemin analogues as anti-tumor agents. Bioorg. Med. Chem. 2003;11:5493–5499.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=1177&id=14&aid=9537