J. Serb. Chem. Soc. 76 (9) 1263–1282 (2011) UDC 66.011+54.061/.062:635.71:615.277
JSCS–4202 *Original scientific paper*

Molecular modeling and chemometric study of anticancer derivatives of artemisinin

JARDEL P. BARBOSA[1]*, JOÃO E. V. FERREIRA[1], ANTONIO F. FIGUEIREDO[1], RUTH C. O. ALMEIDA[1], OSMARINA P. P. SILVA[1], JOSÉ R. C. CARVALHO[1], MARIA DA G. G. CRISTINO[1], JOSE CIRÍACO-PINHEIRO[1], JOSÉ L. F. VIEIRA[2] and RAYMONY T. A. SERRA[3]

[1]Laboratório de Química Teórica e Computacional, Faculdade de Química, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, CP 101101, CEP 66075-110 Belém, PA, Amazônia, [2]Faculdade de Farmácia, Instituto de Ciências da Saúde, Universidade Federal do Pará, CP 101101, CEP 66075-110 Belém, PA and [3]Centro de Ciências Biológicas e da Saúde, Universidade Federal do Maranhão, CEP 65085-580 São Luis, MA, Brasil

(Received 27 December 2010)

Abstract: In this work, a molecular modeling and multivariate study involving artemisinin and 28 derivatives with activity against human hepatocellular carcinoma HepG2 is reported. The studied calculations of the compounds were performed at the B3LYP/6-31G** level. MEP maps were used in an attempt to identify key structural features of artemisinin and its derivatives that are necessary for their activities, and to investigate their interaction with the transferrin. The chemometrics methods PCA, HCA, KNN, SIMCA and SDA were employed in order to reduce dimensionality and to investigate which subset of variables could be more effective for classification of the compounds according to their degree of anticancer activity. Chemometric studies revealed that the *ALOGPS_logs*, *Mor29m*, *IC5* and the gap energy descriptors are responsible for the separation into more active and less active compounds. In addition, molecular docking was used to investigate the interaction between ligands and receptor. The results showed that the ligands approached the receptor through the endoperoxide bond.

Keywords: artemisinin; HepG2; MEP maps; chemometrics; molecular docking.

INTRODUCTION

Cancer, malignant neoplasia and malignant tumor are synonymous words for the disease characterized by uncontrolled growth of abnormal cells of an organism. The presence of some characteristics in these cells indicates alteration in genes owing to mutation in DNA.[1]

* Corresponding author. E-mail: jardelquantun@yahoo.com.br
doi: 10.2298/JSC111227111B

According to the WHO, in 2008, the total number of new cases of cancer would reach 12.3 million people, with 7.6 million deaths. The least developed regions are more affected, considering both incidence (56 %) and mortality (63 %).[2] Most deaths are due to lung cancer (≈18.2 %), stomach cancer (≈9.7 %) and liver cancer (≈9.2 %). It is important to state that among the malignant tumors in the liver, the most common and dangerous one is the hepatocellular carcinoma.[2] Thus, it is essential to discover new efficient strategies to treat the disease and to avoid new cases.

Artemisia annua L. (qinghao) is one of the plants that have shown anticancer properties. It contains the active ingredient artemisinin, which is used as an antimalarial, mainly against *falciparum* and *vivax* malaria.[3–6] Lately, the sensitivity to artemisinin and its first generation derivatives artesunate, artemether, arteether and artelinate has been evaluated in some tumoral cells. It was verified that artemisinin and its derivatives exhibit cytotoxicity to mammary cell in nanomolar and micromolar concentrations.[7] Moreover, low concentrations of these compounds show cytotoxicity to leukemia cells, colon cancer, lung cancer, kidney cancer,[8] thyroid cancer,[9] cervical cancer, uterine cancer and ovarian cancer.[10]

Another aspect that has intensified research with artemisinin and its derivatives in anticancer treatment is the lack of resistance that tumoral cells show to these compounds. As a strategy to preserve natural sources of *Artemisia annua* L., endoperoxides similar to artemisinin were synthesized.[11] Some of these second generation compounds showed considerable cytotoxicity to tumoral cells.[12–16]

In this report, a molecular modeling and chemometric study of 29 artemisinins (artemisinin and its derivatives) with different degrees of cytotoxicities against human hepatocellular carcinoma HepG2 is presented.[17] The employed strategy was based on the knowledge that the endoperoxide group presented in artemisinin and its derivatives is responsible for their antimalarial and anticancer activities. Calculations of the studied molecules were performed by the B3LYP/6-31G** method as implemented in the Gaussian 98 program.[18] MEP maps were used in an attempt to identify key structural features of the artemisinin and the derivatives that are necessary for their activities and to investigate the interaction with a molecular receptor (transferrin). PC and HC analyses, KNN and SIMCA methods[19] and SD analysis[20,21] were employed in order to reduce dimensionality and to investigate which subset of variables could be more effective at classifying the compounds according to their degree of anticancer activity. Molecular docking studies were used to investigate the interaction between the ligands and receptor (transferrin). The developed studies could provide valuable insight into the experimental process of syntheses and biological evaluation of new artemisinin derivatives with activity against cancer HepG2.

COMPUTATIONAL

Modeled artemisinin and derivatives

The compounds, the subjects of this study, consisted of artemisinin, amides, esters, alcohols, ketones, derivatives with polar hydroxyl and carboxylic acid groups and five-membered ring derivatives. All compounds have been associated with *in vitro* bioactivity against a human hepatocellular carcinoma cell line, HepG2, and were divided previously into two classes according with their activities: (–) less active (those with $IC_{50} \geq 97$ μM) and (+) more active (those with $IC_{50} < 97$ μM) derivatives. The atom numbering adopted in this study is showed in Fig. 1 (artemisinin **1**).

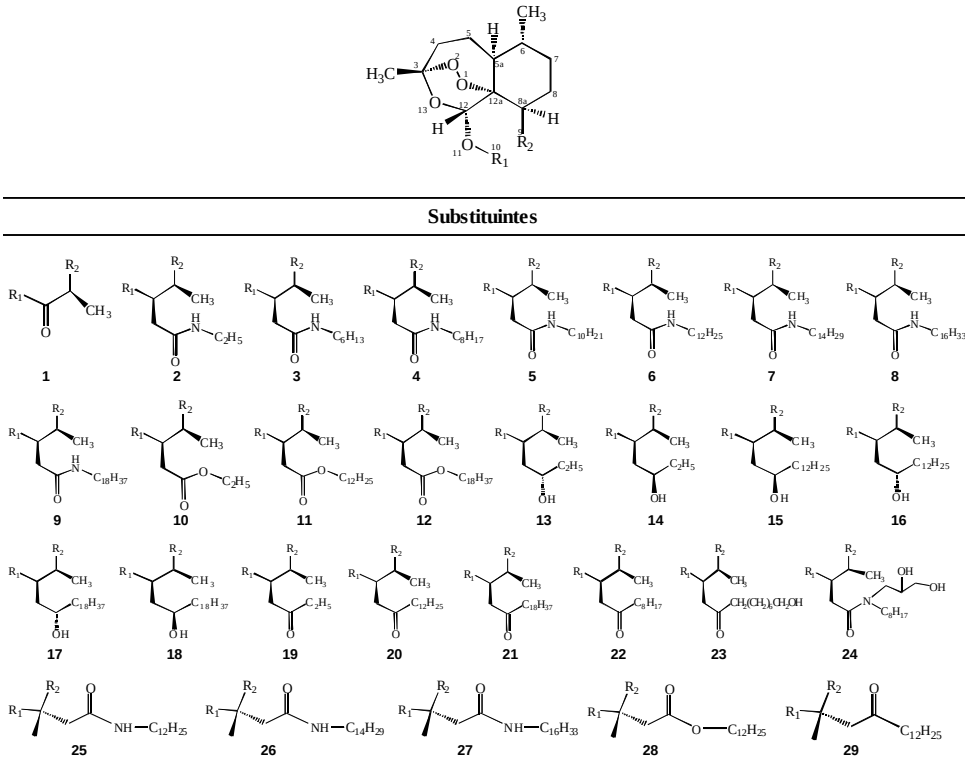

Fig. 1. Artemisinin and derivatives with anticancer HepG2 activity.

Molecular modeling

Quantum chemical approaches implemented in the Gaussian 98 program were used in the modeling of the artemisinin derivatives. Initially, the artemisinin geometry optimization was performed by Hartree-Fock (HF) method[22] and Density Functional Theory (B3LYP)[23,24] with the 3-21G, 6-31G, 6-31G*, 6-31G**, CEP-31G and CEP-31G* basis sets and AM1 and PM3 methods, also available in the Gaussian program.[18] These calculations were performed to find the method that would present the best compromise between computational time and accuracy of the information relative to experimental data.[25] The experimental structure of artemisinin was retrieved from the Cambridge Structural Database CSD[26] with REFCODES: QNGHSU10,[25] crystallographic *R* factor 3.6.

Chemometric methods implemented in the Pirouette program,[27] PCA and HCA, were used to compare the optimized structures by different approaches with the experimental structure of artemisinin QNGHSU to identify the appropriate method and the basis set for further calculations. The analyses were performed on an autoscaled data matrix with dimension 15×18, where each row was related to 14 computed and 1 experimental geometries and each column represented one of 18 geometrical parameters of the 1,2,4-trioxane ring (bond lengths, bond angles and torsion angles). In order to optimize all structures and to perform calculations to obtain the molecular properties, the B3LYP/6-31G** method was also selected.

Molecular descriptors

The descriptors were computed to represent electronic, steric as well as hydrophilic and hydrophobic features, and to allow for quantification of their influence on the biological activity of the studied molecules. The electronic descriptors employed were: total energy, HOMO, HOMO–1, LUMO, LUMO+1 energies, LUMO–HOMO gap energy, Mulliken's electronegativity, and molecular hardness and softness; the steric descriptors were: O1–O2 bond length, C12a–O1–O2–C3 torsion angle, superficial area, molecular volume, Mor29m and IC5; the hydrophilic descriptors were: *HYF*, *ALOGPS_logs* and the hydrophobic descriptor was the log *P* value. The computation of the descriptors was performed employing the Gaussian 98 program, the e-Dragon program[28] and the HyperChem 6.02. program.[29]

Molecular electrostatic potential maps

The MEP is related to the electronic density and it is a very useful descriptor to understand sites for electrophilic attack and nucleophilic reactions as well as for hydrogen bonding interactions.[30–33] The electrostatic potential, $V(r)$, is also well-suited for analyzing processes based on the "recognition" of one molecule by another, as in drug–receptor, and enzyme––substrate interactions, because it is through their potentials that the two species first "see" each other.[34,35] Being a real physical property, $V(r)$ can be determined experimentally by diffraction or by computational methods.[36] To investigate the reactive sites of artemisinin and its derivatives, the MEP was evaluated using the B3LYP/6-31G** method. The MEP at a given point (x,y,z) in the vicinity of a molecule is defined in terms of the interaction energy between the electrical charge generated from the molecule's electrons and nuclei and a positive test charge (a proton) located at r. For the studied compounds, the $V(r)$ values were calculated as described previously using Eq. (1).[37]

$$V(r) = \sum_A \frac{Z_A}{|R_A - r|} - \int \frac{\rho(r')}{|r' - r|} dr' \quad (1)$$

where Z_A is the charge of nucleus A, located at R_A, $\rho(r')$ is the electronic density function of the molecular, and r' is the dummy integration variable.

The MEP was realized by the Molekel program.[38]

Chemometrics

Principal component analysis (PCA). Given the great number of multivariate computed data, an exploratory tool is recommended to uncover unknown trends in the data and to reduce them. The central idea of PCA[19] is to reduce the dimensionality of a data set consisting of large number of interrelated variables, while retaining, as much as possible, the variation present in the data set. This is achieved by transforming them into a new set of variables, the principle components (PCs), which are uncorrelated and ordered so that the first few retain most of the variation present in all of the original variables. The final result of PCA is the selection of a small number of descriptors (molecular properties) that are believed to be best related to the dependent variable, in this case, anticancer activity against HepG2-strains.

Hierarchical cluster analysis (HCA). The statistical analysis required in this study should group compounds of a similar kind into respective categories. HCA[19] is a statistical method developed for this purpose. It is represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of the analysis. The single samples (compounds) are represented by the branches on the bottom of the dendrogram. The similarity among the clusters is given by the length of their branches so that compounds presenting low similarity have long branches whereas compounds of high similarity have short branches.

K-Nearest neighbor (KNN) method. The statistical technique KNN[19] categorizes an unknown object based on its proximity to samples already placed into categories. Specifically, the predicted class of an unknown object depends on the class of its K nearest neighbors, which accounts for the name of the technique. Classification with KNN is related to the compared distance among samples. Multivariate Euclidean distances between every pair of training samples are computed. After the model is built, a test set has its predicted class taking into account the multivariate distance of this sample with respect to the K samples in the training set.

Soft independent modeling of class analogy (SIMCA) method. This method develops principal component models for each training set category. The main goal of SIMCA[19] is the reliable classification of new samples. When a prediction is made in SIMCA, new samples insufficiently close to the PC space of a class are considered non-members. Additionally, the method requires that each training sample be pre-assigned to one of *Q* different categories, where *Q* is typically greater than one. It provides three possible outcome predictions: the sample fits only one pre-defined category, the sample does not fit any of the pre-defined categories and the sample fits into more than one pre-defined category.

Stepwise discriminant analysis (SDA). SDA[20,21] is also a multivariate method that attempts to maximize the probability of correct allocation. This method has two main objectives, which are to separate objects from distinct populations and to allocate new objects into populations previously defined. A stepwise procedure is used in this program, *i.e.*, in each step, the most powerful variable is entered into the discriminant function. The criterion function for selecting the next variable depends on the number of specified groups.

The SDA is a method based on the *F*-test for the significance of the variables. In each step, one variable is selected based on its significance and, after several steps, the more significant variables are extracted from the whole set in question.

Molecular Docking

The geometry of the molecules **1**–**29** was optimized by the B3LYP/6-31G** method while the geometry of the protein receptor was obtained from the Protein Data Bank (PDB) RCSB, identified by the code 1A8E,[39] and optimized by the ROB3LYP/6-31G** method. In order to better describe the biological environment involving the ligand/receptor interaction, only the protein fragment (iron atom and the amino acids bound to it: two tyrosines, one histidine, and one aspartic acid) of the transferrin was considered. The geometry of the complex was optimized through the molecular mechanics method with the force field MM+ implemented in the HyperChem 6.02 program and Polak Ribiere algorithms with a gradient of 0.1 kcal $Å^{-1}$ mol^{-1}*. Flexible docking[40] calculations were performed into a box. When using artemisinin (**1**), the dimensions of the box were *x* =18 Å, *y* =18 Å and *z*=18 Å and the number of molecules of water was 193, while for the other compounds (**2**–**29**) the dimensions were

* 1 kcal = 4.184 kJ

x=20 Å, *y*=20 Å and *z*=29.3 Å and the number of molecules of water was 388. For the simulation involving the complex, the peroxide group (pharmacophore) of artemisinin and its derivatives was positioned toward the iron fragment of the receptor in order to achieve a good interaction.

RESULTS AND DISCUSSION

Method and basis set for the description of the geometries of artemisinin and its derivatives

The theoretical and experimental parameters of the 1,2,13-trioxane ring in artemisinin are given in Table I. They were used with the objective to identify, through PC and HC analyses, which geometry optimization method/basis set give results closest to the experimental data. The advantage in using the PCA and HCA methods in this step of this study was that all structural parameters are considered simultaneously and it takes into account the correlations among them.

The first three *PC*s explain 83.8 % of the original information as follows: *PC1* = 38.73, *PC2* = 28.8 and *PC3* = 16.3 %. The *PC1–PC2* scores plot is shown in Fig. 2a, from which it can be see that the methods are discriminated into two classes according to PC1. The semi-empirical methods (AM1 and PM3) are on the right side; while the other theoretical (HF and B3LYP) and experimental methods are on the left side. Moreover, it can be seen that the B3LYP/6-31G* and B3LYP/6-31G** method are the closest to the experimental method, indicating that either of them could be used in the development of this study.

TABLE I. Theoretical and experimental parameters of the 1,2,4-trioxane ring in artemisinin (**1**)

Geometry	HF/3-21G	HF/6-31G	HF/ /6-31G*	HF/ /6-31G**	HF/ /CEP31G	HF/ /CEP/31G*
O1–O2	1.462	1.447	1.390	1.390	1.439	1.395
O3–C3	1.441	1.435	1.396	1.396	1.447	1.405
C3–O13	1.436	1.435	1.408	1.409	1.449	1.418
O13–C12	1.408	1.403	1.376	1.376	1.413	1.384
C12–C12a	1.529	1.533	1.532	1.532	1.549	1.542
O1–O2	1.462	1.447	1.390	1.390	1.439	1.395
O1–O2–C3	107.089	108.799	109.458	109.460	109.383	109.452
O2–C3–O13	107.274	106.762	107.841	107.818	106.771	108.011
C3–O13–C12	115.684	117.294	115.292	115.309	116.781	114.800
O13–C12–C12a	112.092	112.289	112.268	112.263	112.462	112.473
C12–C12a–O1	111.600	110.964	110.540	110.545	110.588	110.539
C12a–O1–O2	111.296	113.233	112.701	112.700	113.386	112.506
O1–O2–C3–O13	–74.701	–71.865	–73.369	–73.377	–71.992	–73.783
02–C3–O13–C12	32.363	33.414	31.034	31.058	32.955	31.026
C3–O13–C12–C12a	28.188	25.285	27.432	27.402	25.432	27.538
O13–C12–C12a–O1	–50.771	–49.377	–50.157	–50.143	–49.683	–50.321

TABLE I. Continued

Geometry	HF/3-21G	HF/6-31G	HF/ /6-31G*	HF/ /6-31G**	HF/ /CEP31G	HF/ /CEP/31G*
C12–C12a–O1–O2	9.941	12.486	10.918	10.924	12.683	10.721
C12a–O1–O2–C3	50.358	46.723	48.670	48.674	46.638	48.892

Geometry	B3LYP/3-21G	B3LYP/6-31G	B3LYP/6-31G*	B3LYP//6-1G**
O1–O2	1.524	1.525	1.460	1.460
O3–C3	1.456	1.456	1.414	1.414
C3–O13	1.473	1.473	1.441	1.441
O13–C12	1.431	1.426	1.396	1.396
C12–C12a	1.535	1.538	1.539	1.539
O1–O2	1.524	1.525	1.460	1.460
O1–O2–C3	105.587	107.304	108.267	108.286
O2–C3–O13	108.226	107.739	108.490	108.494
C3–O13–C12	113.194	114.993	114.095	114.067
O13–C12–C12a	113.302	113.644	113.261	113.242
C12–C12a–O1	112.413	111.751	111.308	111.285
C12a–O1–O2	109.627	111.405	111.609	111.595
O1–O2–C3–O13	–76.620	–73.463	–73.937	–73.913
02–C3–O13–C12	33.756	34.982	32.863	32.782
C3–O13–C12–C12a	29.076	26.258	27.383	27.507
O13–C12–C12a–O1	–52.217	–51.204	–51.197	–51.344
C12–C12a–O1–O2	9.625	12.763	11.692	11.784
C12a–O1–O2–C3	51.056	46.885	47.879	47.835

Geometry	B3LYP/CEP31G	B3LYP/CEP31G*	AM1	PM3	EXP48
O1–O2	1.515	1.468	1.289	1.544	1.469(2)
O3–C3	1.475	1.427	1.447	1.403	1.416(3)
C3–O13	1.488	1.452	1.427	1.428	1.445(2)
O13–C12	1.440	1.406	1.416	1.403	1.379(2)
C12–C12a	1.559	1.551	2.220	1.555	1.523(2)
O1–O2	1.515	1.468	1.289	1.544	1.469(2)
O1–O2–C3	107.961	108.381	112.530	110.339	108.100(1)
O2–C3–O13	107.602	108.566	103.615	104.815	106.600(2)
C3–O13–C12	114.883	113.886	115.479	116.005	114.200(2)
O13–C12–C12a	113.722	113.457	113.502	115.205	114.500(2)
C12–C12a–O1	111.510	111.322	111.057	113.177	110.700(2)
C12a–O1–O2	111.739	111.573	113.736	112.289	111.200(2)
O1–O2–C3–O13	–73.439	–74.064	–77.795	–73.301	–75.500(2)
02–C3–O13–C12	34.923	33.258	42.001	52.699	36.000(2)
C3–O13–C12–C12a	25.587	26.926	11.490	2.817	25.3000(2)
O13–C12–C12a–O1	–51.007	–51.184	–41.826	–40.526	–51.300(2)
C12–C12a–O1–O2	13.020	11.945	12.049	19.968	12.700(2)
C12a–O1–O2–C3	46.466	47.496	47.069	35.602	47.800(2)

HCA was also used to investigate the most appropriate method for further calculation. Analyzing the dendrogram obtained by HCA with complete linkage

method in Fig. 2b, it can concluded that the theoretical methods are distributed in a similar way as in PCA, *i.e.*, HC analysis confirmed the PC analysis results. Therefore, according to the obtained results, the B3LYP in combination with either of the 6-31G* and 6-31G** basis sets can be used to model the molecular structure of the studied compounds. In this study, the B3LYP/6-31G** method was used.

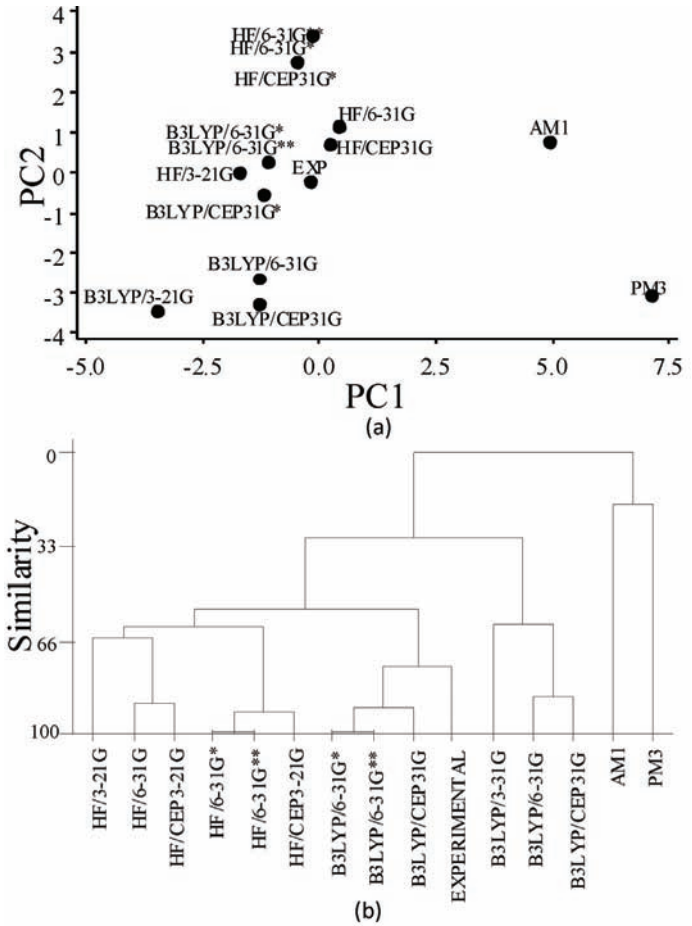

Fig. 2. *PC1–PC2* scores plot (a) and dendrogram (b) of the fourteen theoretical and experimental methods used in the geometry optimization of artemisinin.

Molecular electrostatic potential maps

The MEP maps for artemisinin and the derivatives given in Fig. 1 were similar (Fig. 3). They display contour surfaces close to that of the 1,2,13-trioxane ring, which is characterized by negative electrostatic potentials (red and green colors), on which the lowest value for the charge was about –0.11 a.u. (red color).

Such a characteristic indicates concentration of the electron density due to the lone electron pairs on the oxygen atoms (O_1, O_2 and O_{13}). These molecules also have contour surfaces characterized by positive electrostatic potentials (blue), whereby the highest value was about 0.049 a.u. The distribution of electron density on the molecules around the trioxane ring induces their cytotoxicity against cancer, a belief supported by the fact that the complexation of artemisinin with transferrin involves particularly the interaction between the peroxide bond, the most negatively charged region on the ligand, and the iron(II) ion, the most positive zone on the transferring, the receptor, molecule.[41–43] Hence, the presence of a surface in red near to 1,2,13-trioxane ring suggests artemisinin and derivatives have a reactive site for electrophilic attack and must possess anticancer cytotoxicity; consequently, they are of interest for investigation. Thus, in the case of an electrophilic attack of the iron of transferrin against an electronegative region of the studied compounds, this attack has a great preference to occur through the involvement of the endoperoxide linkage. A pattern of MEP maps is an indication that the artemisinin and derivatives in Fig. 1 are all active against cancer. Thus, by analyzing MEP maps, the selection of inactive compounds is avoided in the proposition step of potential new active derivatives against cancer HepG2.

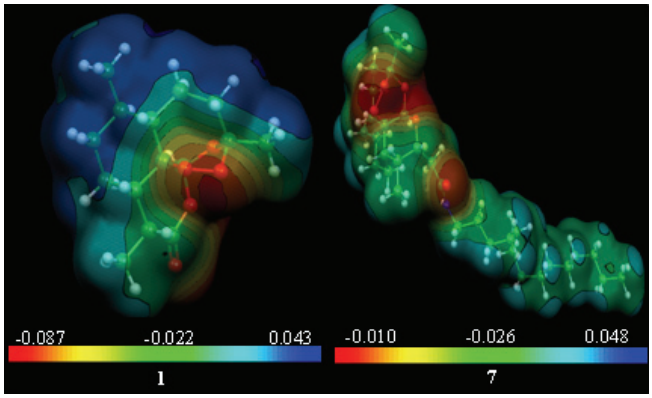

Fig. 3. MEP (a.u.) Maps of artemisinin (**1**) and derivative compound **7** with anticancer HepG2 activity.

PCA Method

The PCA results show the score plot relative to the first and second principal components. In *PC1*, there is a distinct separation of the compounds into two classes (Fig. 4): more and less active. More active compounds are on the left side, while less active are on the right side. The variables responsible for this were *ALOGPS_logs*, *Mor29m*, *IC5* and gap energy. They were chosen from the complete data set (1740 descriptors) and they are assumed to be very important in the investigation the anticancer mechanism involving artemisinins. Other vari-

ables were not selected because either they had a poor linear correlation with activity or they did not give a distinct separation between the more and less active compounds. The values for these properties are listed in Table II. The first three principal components, *PC1*, *PC2* and *PC3* explained 42.97, 28.72 and 14.94 % of the total variance, respectively. According to Table III, the Pearson correlation coefficient between the variables was, in general, low (less than 0.50, in absolute values).

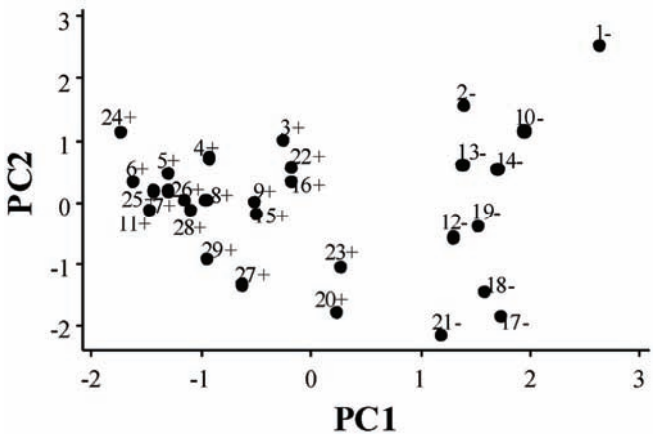

Fig. 4. Plot of the first two principal components score vectors (**PC1**×**PC2**) for the separation of artemisinin and the derivatives into two classes: (+) more active compounds and (–) less active compounds against HepG2 cancer.

TABLE II. Values of the four properties that classify artemisinin and the derivatives and values of experimental IC_{50}

Compound	*ALOGPS_logs*	*Mor29m*	*IC5*	Gap energy, a.u.	IC_{50} / μM	Activity
1	–2.3500	–0.3050	4.8620	0.2616	97	la[a]
2	–3.5200	–0.3070	5.2530	0.2525	>100	la
3	–5.0300	–0.4120	5.5140	0.2522	17.6	ma[b]
4	–5.7600	–0.4430	5.6280	0.2522	9.5	ma
5	–6.3500	–0.4550	5.6840	0.2521	2.8	ma
6	–6.8400	–0.5250	5.6240	0.2524	1.2	ma
7	–7.1600	–0.5140	5.5010	0.2527	0.46	ma
8	–7.3900	–0.5150	5.3640	0.2524	0.79	ma
9	–7.4900	–0.5010	5.2250	0.2525	4.2	ma
10	–3.6400	–0.2360	5.2170	0.2467	>100	la
11	–7.0300	–0.5260	5.5970	0.2462	0.72	ma
12	–7.6800	–0.1790	5.1970	0.2462	>100	la
13	–3.6800	–0.3650	5.2530	0.2367	>100	la
14	–3.6800	–0.3050	5.2530	0.2359	>100	la
15	–6.9700	–0.3940	5.5080	0.2457	2.8	ma
16	–6.9700	–0.2910	5.5080	0.2552	4.4	ma
17	–7.4000	–0.2280	5.1590	0.2217	>100	la

TABLE II. Continued

Compound	*ALOGPS_logs*	*Mor29m*	*IC5*	Gap energy, a.u.	IC_{50} / μM	Activity
18	–7.4000	–0.2280	5.1590	0.2287	>100	la
19	–3.7500	–0.4430	5.1800	0.2194	>100	la
20	–7.1401	–0.3051	5.5711	0.2197	1.4	ma
21	–7.6100	–0.3330	5.1680	0.2177	>100	la
22	–6.0300	–0.3110	5.5720	0.2524	1.8	ma
23	–5.4900	–0.3470	5.6380	0.2199	42.3	ma
24	–4.8200	–0.5180	5.8560	0.2517	3.5	ma
25	–6.7200	–0.5520	5.5430	0.2491	1.3	ma
26	–7.0600	–0.5520	5.4190	0.2492	0.77	ma
27	–7.3500	–0.6010	5.2800	0.2276	0.74	ma
28	–6.8400	–0.5150	5.5160	0.2449	3.7	ma
29	–7.0100	–0.5430	5.4880	0.2323	0.47	ma

[a]The less active compound; [b]The more active compound

TABLE III. Correlation matrix for the descriptors

	ALOGPS_logs	*Mor29m*	*IC5*
Mor29m	0.251	–	–
IC5	–0.269	–0.464	–
Gap energy	0.152	–0.207	0.203

The loading plot relative to the first and second principal components can be seen in Fig. 5. *PC1* is expressed in Eq. (2) as a function of the four selected descriptors. Thus, it is a quantitative variable that provides the overall predictive ability of the different sets of molecular descriptors for all the selected properties. The loadings of *ALOGPS_logs* and *Mor29m* are positive whereas they are negative for *IC5* and gap energy. Incidentally, gap energy is the least important pro-

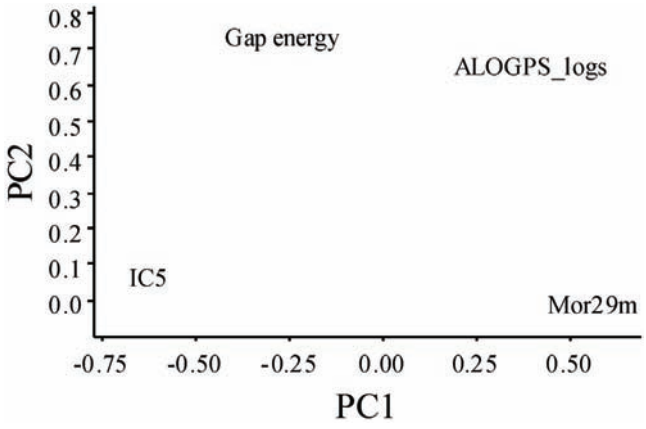

Fig. 5. Plot of the first two principal components loadings vectors (**PC1**×**PC2**) for the four descriptors responsible for the separation of the artemisinin and the derivatives into two classes: (+) more active compounds and (–) less active compounds against HepG2 cancer.

perty contributing to *PC1*, as its coefficient (–0.272) is lower in comparison to those of the others. For a compound to be more active against cancer, it must generally have a combination of these properties, i.e., more negative values for *ALOGPS_logs* and *Mor29m* but more positive values for *Gap energy* and *IC5*.

$$PC1 = 0.393ALOGPS\_logs + 0.619\text{Mor29m} - 0.624\ IC5 - 0.272\text{Gap energy} \quad (2)$$

HCA Method

The HCA method that better classified the compounds into two classes (more and less active compounds) was the complete method. In the complete linkage, the distance between two clusters is the maximum distance between a variable in one cluster and a variable in the other cluster. The descriptors employed to perform HCA were the same as for PCA, *i.e.*, *ALOGPS_logs*, *Mor29m*, *IC5* and gap energy. The dendrogram (Fig. 6) shows HCA graphic as well as the compounds separated into two main classes. The scale of similarity varies from 0 for samples with no similarity to 100 for samples with identical similarity. By analyzing the dendrogram, some conclusions can be made even though the compounds present some structural diversity.

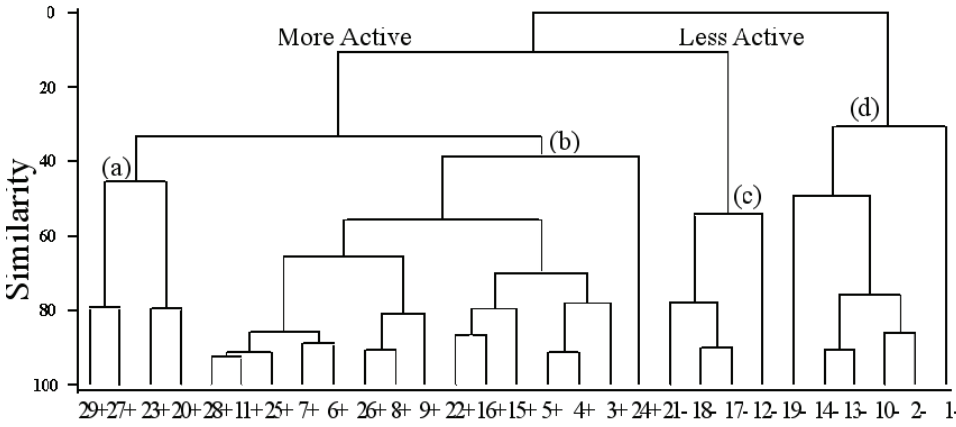

Fig. 6. HCA Dendrogram for artemisinin and the derivatives with anti-HepG2 cancer activity.

The more active analogs are distributed into clusters **a** and **b**. They have both a six-membered ring artemisinin and a five-membered ring artemisinin analogs and substituents with a greater number of carbon atoms. The number of amides in clusters **a** and **b** (more active analogues) is eleven, while in clusters **c** and **d** (less active analogues), it is only one. The values for *IC5* are the highest (*IC5* > 5.20) and for *Mor29m* they are the most negative. The variation in activity is the greatest in cluster **a** (IC_{50} = 0.47 μM, for **29**, and IC_{50} = 42.3 μM, for **23**). This cluster also presents the lowest values for gap energy among the most active compounds.

The less active class was divided into two main clusters (**c** and **d**). All samples, excluding **1**, present a six-membered ring artemisinin. These two clusters have a clear difference involving values for *ALOGPS_logs* (cluster **c**, –7.40 to –7.68; cluster **d**, –2.35 to –3.75) and *Mor29m* (cluster **c**, –0.179 to 0.333, which are the lowest of all compounds; cluster **d**, –0.236 to –0.443). Since artemisinin (**1)** is found to bear different substitution from the others structures, it was classified alone. It shows the lowest negative value for *ALOGPS_logs* and the lowest positive value for *IC5* (–2.35). Cluster **c** contains ester **12**, alcohols **17** and **18** (isomers), and ketone **21**. These compounds bear a long carbon chain of $C_{18}H_{37}$. Samples in cluster **d** are amide **2,** ester **10**, alcohols **13** and **14** (isomers), and ketone **19**. They bear a short carbon chain, C_2H_5 (the exception is **1**).

Interestingly, alcohols with a $C_{12}H_{25}$ carbon chain developed cytotoxicities much higher than alcohols with a short C_2H_5 carbon chain or with a long $C_{18}H_{37}$ carbon chain. Moreover, ketones with C_8H_{17} and $C_{12}H_{25}$ carbon chains were much more active than ketones with a short C_2H_5 carbon chain or a long $C_{18}H_{37}$ carbon chain. Moreover, as the length of the carbon chain of the six-membered ring amides increased, the activity increased from **2** (C_2H_5, IC_{50} = 100 μM) through **3** (C_6H_{13}, IC_{50} = 17.6 μM), **4** (C_8H_{17}, IC_{50} = 9.5 μM), **5** ($C_{10}H_{21}$, IC_{50} = 2.8 μM), **6** ($C_{12}H_{25}$, IC_{50} = 1.2 μM) to **7** ($C_{14}H_{29}$, IC_{50} = 0.46 μM). Further increases in the carbon chain length resulted in slight drops in the cytotoxicity as can be noted for **8** ($C_{16}H_{33}$, IC_{50} = 0.79 μM) and **9** ($C_{18}H_{37}$, IC_{50} = 4.2 μM). For the five-membered ring amides, as the length of carbon chains increased, the activity increased slightly from **25** ($C_{12}H_{25}$, IC_{50} = 1.3 μM), **26** ($C_{14}H_{29}$, IC_{50} = = 0.77 μM) to **27** ($C_{16}H_{33}$, IC_{50} = 0.74 μM).

KNN Method

The results obtained with the KNN method using one to eight (K = 1 to K = = 8) nearest neighbors are given in Table IV and the same was used for validation of the initial set (Fig. 1). For K = 1, 2, 4, 6 and 8, the percentage correct information was 100 % but 8NN was used because the higher the number of nearest neighbors, the better the reliability of the method in question.

TABLE IV. Classification obtained with the KNN method

Category	Number of compounds	Compounds incorrectly classified							
		1NN	2NN	3NN	4NN	5NN	6NN	7NN	8NN
Class: more active	19	0	0	1	0	1	0	0	0
Class: less active	10	0	0	0	0	0	0	1	0
Total	29	0	0	1	0	1	0	1	0
% Correct information	–	100	100	96.36	100	96.36	100	96.36	100

SIMCA Method

The results obtained with the SIMCA method are given in Table V. The method was also used for validation of the initial set (Fig. 1). For the 19 more active compounds, one was incorrectly classified. The model was built with three PCs for the more active class and two PCs for the less active class.

TABLE V. Classification obtained by using SIMCA method

Category	Number of compounds	Correct classification
Class: more active	19	18
Class: less active	10	10
Total	29	29
% Correct information	–	96.5

SD Analysis

From the two-discrimination function obtained with the SDA study, it can see that the variables *ALOGPS_logs*, *MOR29m*, *IC5* and *Gap energy* have a large contribution in the classification methodology. According to the results using PCA, HCA, KNN, SIMCA and SDA, it can also be seen that the descriptors are key properties for explaining the anticancer HepG2 activity of the derivatives compounds artemisinin (Fig. 1).

The discrimination functions for the more active and less active groups are given, respectively, by Eqs. (3a) and (3b).

Group more active:

$$-1.28-1.18ALOGPS\_logs - 1.45Mor29m + 2.27IC5 + 0.772\text{Gap energy} \quad (3a)$$

Group less active:

$$-4.64 + 2.23ALOGPS\_logs + 2.75Mor29m - 4.32IC5 - 1.47\text{Gap energy} \quad (3b)$$

Through the discrimination functions and the value of each variable for the compounds and using all compounds of the training set, the classification is obtained (Table VI). The classification error rate was 0 %, resulting in a satisfactory separation of the more and less active compounds. The allocation rule derived from the SDA results, when the activity against cancer of new artemisinin derivative is investigated, is: a) initially, for the new derivatives, calculate the value of the more important variables obtained in the construction of the SDA model (descriptors); b) substitute these values in the two discrimination functions performed in this work; c) check which discrimination function (group more active compounds or group less active compounds) presents the higher value. The new derivative is more active if it is related to the discrimination function of the more active group and *vice versa*.

To determine if the model obtained is reliable, a cross-validation test which uses the leave-one-out technique was employed. In this procedure, one com-

pound is omitted from the data set and the classification functions are built based on the remaining compounds.

TABLE VI. Classification matrix obtained using SDA

Classification group or class	Number of compounds	True group	
		More active	Less active
Group (Class): more active	19	19	0
Group (Class): less active	10	0	10
Total	29	–	–
% Correct information	–	100	100

Afterwards, the omitted compound is classified according to the classification functions generated. In the next step, the omitted compound is included and a new compound is removed, and the procedure continues until the last compound is removed. The obtained results with the cross-validation methodology are summarized in Table VII.

TABLE VII. Classification matrix obtained by using SDA with cross validation

Classification group or class	Number of compounds	True group	
		More active	Less active
Group (Class): more active	19	19	0
Group (Class): less active	10	0	10
Total	29	–	–
% Correct information	–	100	100

In the application of the chemometric step in this study, it was verified that the most important properties to describe the anticancer activity are: *ALOGPS_logs*, *Mor29m*, *IC5* and gap energy. Thus, some considerations on the most important variables can be drawn for activity from molecules. The *ALOGPS_logs* indicates that drug solubility is one of the important factors, which affect the movement of a drug from the site of administration into the blood. Knowledge of drug solubility is important. It is well-known that insufficient solubility of a drug can lead to poor absorption.[44] From Table II, it can be seen that aqueous solubility is lower for the more active compounds than for the less inactive ones. This is an indication that the lipophilic carbon chain plays an important role in determining the cytotoxicities of the more active compounds.

The *Mor29m* is the 3D-MORSE (Molecule representation of structures based on electron diffraction) code of signal 29, weighted with atomic masses. It is calculated by summing atom weights viewed by a different angular scattering function. The 3D-MORSE code allows the representation of the three dimensional structure of a molecule by a fixed number of values.[45] This fact indicates the importance of atomic mass, a steric property, and gives the basic idea that the larger the molecule is, the higher is the activity, because the activity also in-

creases with decreasing value of *Mor29m*. In fact, the less active compounds (Table II), **12**, **17** and **18**, have *Mor29m* values of –0.1790 and –0.2280, respectively, and the more active compounds, **7**, **8**, **11**, **26**, **27** and **29**, present values of –0.5140, –0.5150, –0.5260, –0.5520, –0.6010 and –0.5430, respectively. It seems clear that larger compounds present a bigger steric effect, augmenting the anticancer activity as consequence of the augmented lipophilicity.

The information content index is calculated based on the pair-wise equivalence atoms in a hydrogen-filled molecule.[46] In Table III, in general, compounds with higher values for *IC5* are more active. These compounds are the same that have a lipophilic carbon chain.

The gap energy is the energy separation between the LUMO and HOMO energies.[47] This property gives information associated to the electronic structure of a molecule and is the measure of the stability of a molecule. A smaller gap energy shows that a molecule is more reactive. In TABLE III, as can be seen, when the gap energy values of the more active and less active compounds are compared, it is not possible to verify a clear tendency for their distinction. This can be associated to the smaller contribution of this property in *PC1* (Eq. (2)). It is possible that the main contribution to a higher anticancer activity against HepG2 developed by a compound is due to its lipophilicity.

Molecular docking

Docking calculations for the compounds from Fig. 1 show that the polar region of the ligands close to the peroxide linkage is directed toward the iron ion of the receptor. The results achieved for the interaction of artemisinin (**1**) and compound **7** with the receptor are presented in Fig. 9. For artemisinin, the Fe–O1

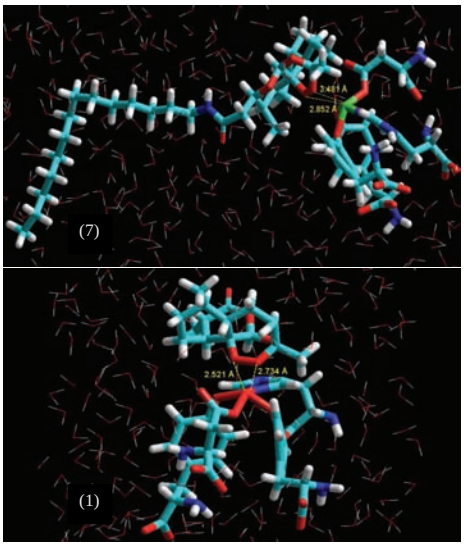

Fig. 9. Docking formed for artemisinin (**1**) and the molecule receptor and derivative **7** and the molecule receptor (transferrin).

and Fe–O2 distances are 2.52 and 2.74 Å, respectively. However, for **7**, they are 2.85 e 3.48Å, respectively. The explanation for a closer approximation of the artemisinin to the iron ion (Fe^{2+}) is probably due to the greater number of bulky substituents in the artemisinin derivative, which makes it difficult to approach the transferrin. Another point to emphasize is that the anticancer activity of the derivatives is probably also related to lipophilicity owing to the great number of carbon atoms present in the molecule, besides the endoperoxide group.[48] The histidine unity in transferrin is usually coordinated to the iron ion through its sp^2 the nitrogen atom. This allows for the iron ion to acquire a hexacoordinated octahedral arrangement after binding to the artemisinin and the derivatives.[49]

CONCLUSIONS

In this work, the use PCA and HCA in the selection step of the method and the basis set for the molecular modeling and development of quantum chemistry calculations revealed that the B3LYP/6-31G* and B3LYP/6-31G** theory were the most adequate. The use of MEP maps to identify key structural features of the artemisinin and its derivatives necessary for their activities and to investigate the interaction with the molecular receptor (transferrin) showed that the presence of a red surface near to the 1,2,13-trioxane ring suggested that these compounds have a reactive site for electrophilic attack and they must possess anticancer cytotoxicity and which in the case of the an electrophilic attack of the ion of tranferrin against an electronegative region of the studied compounds, this attack has a great preference to occur through the involvement of the endoperoxide linkage. Principal component analysis (PCA), hierarchical cluster analysis (HCA), the K-nearest neighbor method (KNN), soft independent modeling of class analogy method (SIMCA) and the stepwise discriminant analysis showed that the studied artemisinin and derivatives can be classified into two classes or group: more active and less active according to their degree of anticancer HepG2 activity. The properties *ALOGPS_logs*, *Mor29m*, *IC5* and gap energy are responsible for the separation into the more active and less active studied molecules and it is interesting to notice that these properties represent three distinct classes of interactions between the molecules and the transferring receptor: electronic (gap energy), steric (*Mor29m* and *IC5*) and hydrophilic (*ALOGPS_logs*). The molecular docking study showed that the ligands approached the receptor (transferrin) through the endoperoxide bond. The developed studies with MEP maps, PCA, HCA, KNN, SIMCA and SDA and molecular docking can provide valuable insight into the experimental process of syntheses and biological evaluation of new artemisinin derivatives with activity against cancer HepG2.

Acknowledgements. We acknowledge the financial support of the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico. We also thank the Instituto de Química-Araraquara for the use of the GaussView software and the Swiss Center for Scien-

tific Computing for the use of the Molekel software and the Laboratório de Química Teórica e Computacional, Universidade Federal do Pará.

И З В О Д

МОЛЕКУЛСКО МОДЕЛОВАЊЕ И ХЕМОМЕТРИЈСКЕ СТУДИЈЕ АНТИКАНЦЕР ДЕРИВАТА АРТЕМИЗИНИНА

JARDEL P. BARBOSA[1], JOÃO E. V. FERREIRA[1], ANTONIO F. FIGUEIREDO[1], RUTH C. O. ALMEIDA[1], OSMARINA P. P. SILVA[1], JOSÉ R. C. CARVALHO[1], MARIA DA G. G. CRISTINO[1], JOSE CIRÍACO-PINHEIRO[1], JOSÉ L. F. VIEIRA[2] и RAYMONY T. A. SERRA[3]

[1]Laboratório de Química Teórica e Computacional, Faculdade de Química, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, CP 101101, CEP 66075-110 Belém, PA, Amazônia, [2]Faculdade de Farmácia, Instituto de Ciências da Saúde, Universidade Federal do Pará, CP 101101, CEP 66075-110 Belém, PA u [3]Centro de Ciências Biológicas e da Saúde, Universidade Federal do Maranhão, CEP 65085-580 São Luis, MA, Brasil

Изложени су резултати молекулског моделовања и мултиваријантне студије која обухвата артемизинин и 28 његових деривата, који делују против хуманог хепатоцелуларног карцинома HepG2. Израчунавања су вршена на нивоу теорије B3LYP/6-31G**. Коришћене су MEP мапе да би се идентификовали структурни детаљи у артемизинину и његовим дериватима, неопходни за њихову активност. Примењене су хемометријске методе PCA, HCA, KNN, SIMCA и SDA да би се установило који подскуп варијабли највише одговара за предвиђање антиканцерогене активиности. Хемометријска разматрања су показала да су *LOGPS_logs*, *Mor29m*, *IC5* и gap energy дескриптори одговорни за разликовање између активних и мање активних једињења. Такође је примењено молекулско доковање да би се истражила интеракција између лиганда и рецептора. Резултати показују да се лиганд везује за рецептор преко ендопероксидне везе.

(Примљено 27. децембра 2010)

REFERENCES

1. S. A. Rosenberg, *Nature* **411** (2001) 380
2. *Word Health Organization – International Agency for Research on Cancer*, Geneva, Switzerland, http://globocan.iarc.fr (accessed August 2010)
3. R. Price, M. van Vugt, F. Nosten, C. Luxemburger, A. Brockman, L. Phaipun, T. Chongsuphajaisiddhi, N. White, *Am. J. Trop. Med. Hyg.* **59** (1998) 883
4. D. M. Opsenica, B. A. Šolaja, *J. Serb. Chem. Soc.* **74** (2009) 1155
5. A. K. Bhattacharjee, K. A. Carvalho, D. Opsenica, B. A. Solaja, *J. Serb. Chem. Soc.* **70** (2005) 329
6. J. E. V. Ferreira, A. F. Figueiredo, J. P. Barbosa, M. G. G. Crispino, W. J. C. Macedo, O. P. P. Silva, B. V. Malheiros, R. T. A. Serra, J. C. Pinheiro. *J. Serb. Chem. Soc.* **75** (2010) 1533
7. A. C. Beekman, H. J. Woerdenbag, W. van Uden, N. Pras, A. W. Konings, H. V. Wikstrom, *J. Pharm. Pharmacol.* **49** (1997) 1254
8. T. Efferth, H. Duntan, A. Sauerbrey, H. Miyachi, C. R. Chitambar, *Int. J. Oncol.* **18** (2001) 767
9. B. Rinner, V. Siegl, P. Purstner, T. Efferth, B. Brem, H. Greger, R. Pfragner, *Anticancer* **24** (2004) 495
10. H. H. Chen, H. J. Zhou, X. Fang, *Pharmacol.* **48** (2003) 231

11. W. Hofheinz, H. Burgin, E. Gocke, C. Jaquet, R. Masciadri, G. Schmid, H. Stohler, H. Urwyler, *Trop. Med. Parasitol*. **45** (1994) 261
12. G. H. Posner, J. Northrop, I. H. Paik, K. Borstnik, P. Dolan, T. W. Kensler, S. Xie, T. A. Shapiro, *Bioorg. Med. Chem.* **10** (2002) 227
13. G. H. Posner, A. J. McRiner, I. H. Paik, S. Sur, K. Borstnik, S. Xie, T. A. Shapiro, A. Alagbala, B. Foster, *J. Med. Chem.* **47** (2004) 1299
14. A. M. Galal, S. A. Ross, M. A. El-Sohly, F. S. El-Feraly, M. S. Ahmed, A. T. McPhail, *J. Nat. Prod*. **65** (2002) 184
15. Y. Li, J. M. Wu, F. Shan, G. S. Wu, J. Ding, D. Xiao, J. X. Han, G. Atassi, S. Leonce, D. H. Caignard, P. Renard, *Bioorg. Med. Chem.* **11** (2003) 1391
16. J. P. Jeyadevan, P. G. Bray, J. Chadwick, A. E. Mercer, A. Bayme, S. A. Ward, B. K. Park, D. P. Williams, R. Cosstick, J. Davies, A. P. Higson, E. Irving, G. H. Posner, P. M. O'Neill, *J. Med. Chem*. **47** (2004) 1290
17. Y. Liu, V. K.-W. Wong, B. C.-B. Ho, M.-K. Woang, C.-M. Che, *Org. Lett.* **7** (2005) 1561
18. *Gaussian 98*, revision A.7, Gaussian, Inc., Pittsburgh, PA, 1998
19. K. R. Beebe, R. J. Pell, M. B. Seashottz, *Chemometrics: A Practical Guide*, Wiley, New York, 1998, p. 81
20. R. A. Johnson, D. W. Wichem, *Applied Multivariate Statistical Analysis*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1992
21. K. V. Mardia, J. T. Kent, J. M. Bibbly, *Multivariate Analysis*, Academic Press, New York, 1979
22. C. C. Roothaan, *Rev. Mod. Phys.* **23** (1951) 69
23. A. D. Becke, *J. Chem. Phys*. **98** (1993) 5648
24. C. Lee, W. Yang, R. G. Parr, *Phys. Rev., B* **37** (1988) 785
25. J. N. Lisgarten, B. S. Potter, C. Bantuzeko, R. A. Palmer, *J. Chem. Cryst*. **28** (1998) 539
26. F. H. Allen, *Acta Cryst., B* **58** (2002) 380
27. *Pirouette 3.01*, Informetrix, Inc., Woodinville, WA, 2001
28. *Virtual Computational Laboratory*, VCCLAB 2005, http://www.vcclab.org (accessed February 2010)
29. *ChemPlus, Modular Extensions to HyperChem Release 6.02*, Molecular Modeling for Windows, Hyper, Inc., Gainesville, FL, 2000
30. H. Tanak, *J. Mol. Struct. THEOCHEM* **5** (2010) 950
31. E. Scrocco, J. Tomasi, *Adv. Quantum Chem.* **11** (1979) 115
32. F. J. Luque, J. M. Lopz, M. Orozco, *Theor. Chem. Acc.* **103** (2000) 343
33. N. Okulik, A. H. Jubert, *Int. Electron J. Mol. Des.* **4** (2005) 17
34. P. Politzer, P. R. Laurence, K. Jayasuriya, J. McKinney, *Environ. Health Perspect.* **61** (1985) 191
35. E. Scrocco, J. Tomasi, *Top. Curr. Chem.* **42 (**1973) 45
36. P. Politzer, D. G. Truhlar, *Chemical Applications of Atomic and Molecular Electrostatic Potentials*, Plenum, New York, 1981, p. 1
37. P. Politzer, J. S. Murray, *Theor. Chem. Acc.* **108** (2002) 134
38. *Molekel 4.3*, Swiss Center for Scientific Computing, Manno, Switzerland, 2000
39. R. T. A. MacGillivray, S. A. Moore, J. Chen, B. F. Anderson, H. Baker, Y. Luo, M. Bewley, C. A. Smith, M. E. P. Murphy, Y. Wang, A. B. Mason, R. C. Woodworth, G. D. Brayer, E. N. Baker, *Biochemistry* **37** (1998) 1719
40. D. E. Koshland, G. Nemethy, D. Filmer, *Biochemistry* **5** (1996) 365

41. H. C. Lai, T. Sasaki, N. P. Singh, (University of Washington), US2004/0067875A1 (2004)
42. T. Sasaki, H. C.-Y. Lai, N. P. Singh, S. J. Oh, (University of Washington) US2007/ /0231300A1 (2007)
43. P. M. O'Neill, V. E. Barton, S. A. Ward, *Molec.* **15** (2010) 1705
44. V. Tetko, V. I. Tanchuk, T. N. Kasheva, A. E. Villa, *J. Chem. Inf. Comput. Sci*. **41** (2001) 1488
45. B. F. Rasulev, N. D. Abdullaev, V. N. Syrov, J. Leszczynski, *QSAR Comb. Chem.* **24** (2005) 1056
46. V. R. Magnuson, D. K. Harriss, S. C. Basak, *Studies in Physical and Theoretical Chemistry*, Elsevier, Amsterdam, 1983, p. 178
47. M. Karelson, V. S. Labanov, A. R. Katritzky, *Chem. Rev*. **96** (1999) 1027
48. M. S. Costa, R. Kiralj, M. M. C. Ferreira, *Quím. Nova* **30** (2007) 25
49. R. K. Haynes, S. C. Vonwiller, *Tetrahedron Lett*. **37** (1996) 253.