ELSEVIER

Journal of Molecular Structure (Theochem) 572 (2001) 35–44

THEO CHEM

www.elsevier.com/locate/theochem

Antimalarial activity of dihydroartemisinin derivatives against *P. falciparum* resistant to mefloquine: a quantum chemical and multivariate study

J.C. Pinheiro[a,*], M.M.C. Ferreira[a], O.A.S. Romero[b]

[a]*Departamento de Físico-Química, Instituto de Química, Universidade Estadual de Campinas, 13081-970. Campinas, São Paulo, Brazil*
[b]*Departamento de Química, Centro de Ciências Exatas e Naturais, Universidade Federal do Pará, CP 11101, 66075-110 Belém, Para, Brazil*

Received 20 December 2000; accepted 23 April 2001

Abstract

Dihydroartemisinin derivatives with antimalarial activity against *P. falciparum* resistant to mefloquine are proposed with the aid of quantum chemistry and multivariate analysis methods (PCA, KNN, and SIMCA). The principal component analysis (PCA) and hierarchical cluster analysis (HCA) showed that the descriptors: molecular softness (MS), total surface area (TSA), Randic's index, path-1 molecular connectivity-average (CHI1A), bond information index (BIC), shape index based on paths of length 2 of Kier (2K), and directional (related to molecular size, dimension: axis 1 and weight: van der Waals' volume) and non-directional (related to linear contribution to the total molecular size and weight: van der Waals' volume) WHIM-3D indices (L1v and Tv), respectively, are responsible for the classification between the higher and lower antimalarial activity of the derivatives. The compounds predicted as of high activity by the three methods are 22 and 28 in Fig. 2. © 2001 Published by Elsevier Science B.V.

Keywords: Antimalarial activity of dihidroartemisinin derivatives; A quantum chemical multivariate study; PCA; HCA; KNN; SIMCA

1. Introduction

Artemisinin or qinghaosu (Fig. 1, compound 1) is a sesquiterpene containing the 1,2,4-trioxane ring system which have been used in China for the treatment of *P. falciparum* malaria [1], a disease responsible through approximately two million of deaths per year [2].

Computational and quantitative structure–activity relationship (QSAR) studies have been developed to unravel the drug's mechanism of action and give guidelines for synthesizing new derivatives with improved efficiency and stability. Thomson et al. studied theoretically the structure of Artemisinin and related molecules using both semiempirical and ab initio quantum chemistry methods, and investigated the structure activity relationship of these molecules as antimalarials using molecular electrostatic potentials maps [3]. Bernardinelli et al. have done a systematic study of the structure of Artemisinin-like molecules using semiempirical and ab initio methods. Molecular electrostatic potential maps have been evaluated and used in an attempt to identify the key features of the molecules that are necessary for their activity [4]. Avery et al. hava built a comparative

* Corresponding author. Permanent address: Departamento de Química, Laboratório de Química Teórica e Computacional, Centro de Ciências Exatas e Naturais, Universidade Federal do Pará, CP 11101, 66075-110 Belém, Pará, Brazil.

0166-1280/01/$ - see front matter © 2001 Published by Elsevier Science B.V.
PII: S0166-1280(01)00522-X

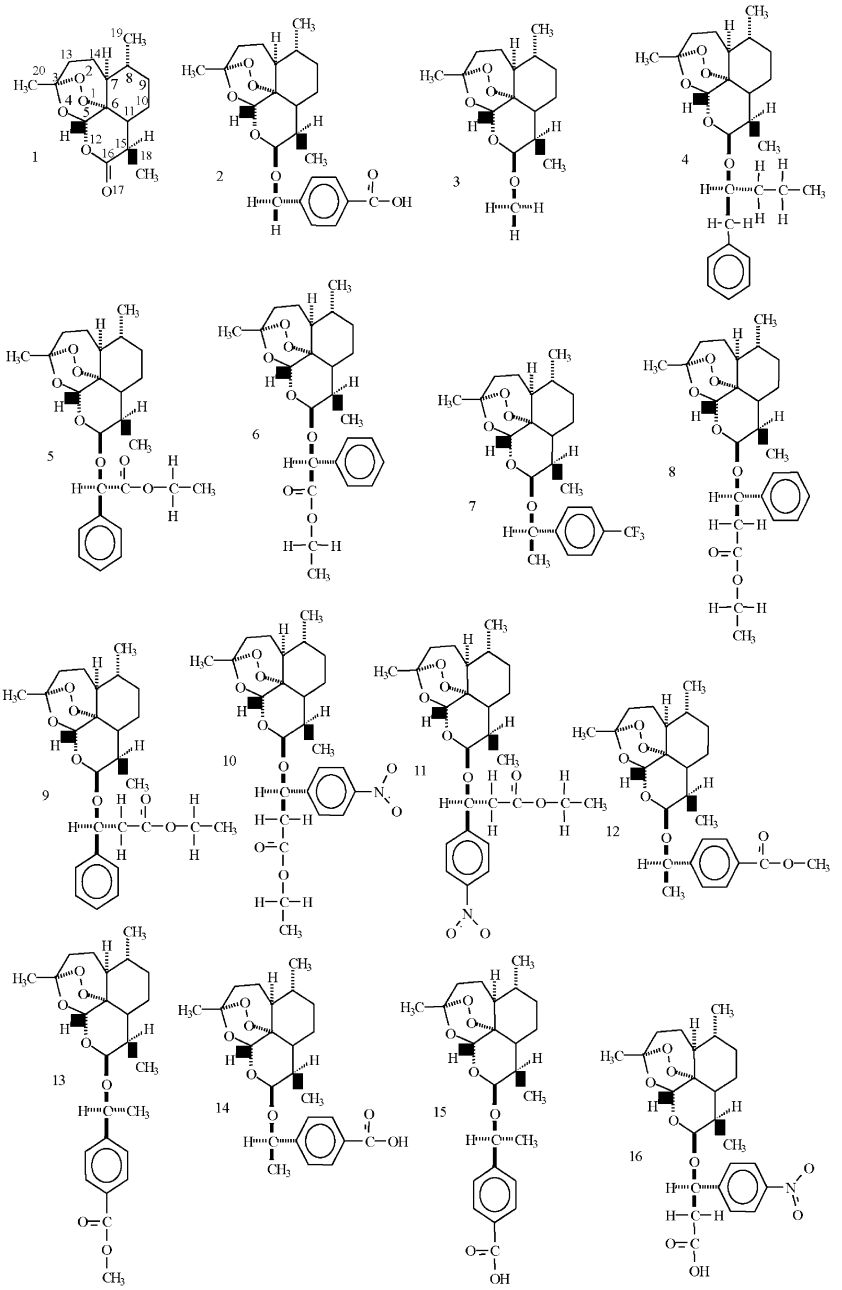

Fig. 1. Dihydroartemisinin derivatives with antimalarial activity (Training Set).

molecular field analysis model for C-9 analogs of Artemisinin and 10-deoxoartemisinin [5,6]. Also, Suter et al. have correlated the three-dimensional molecular electrostatic potentials, obtained through quantum chemistry and projected on two-dimensional surfaces, with the biological activity of Artemisinin derivatives by using neural networks [7]. More recently, Nguyen–Cong et al. described an experiment with the multivariate adaptive regression splines method along with multiple linear regression,

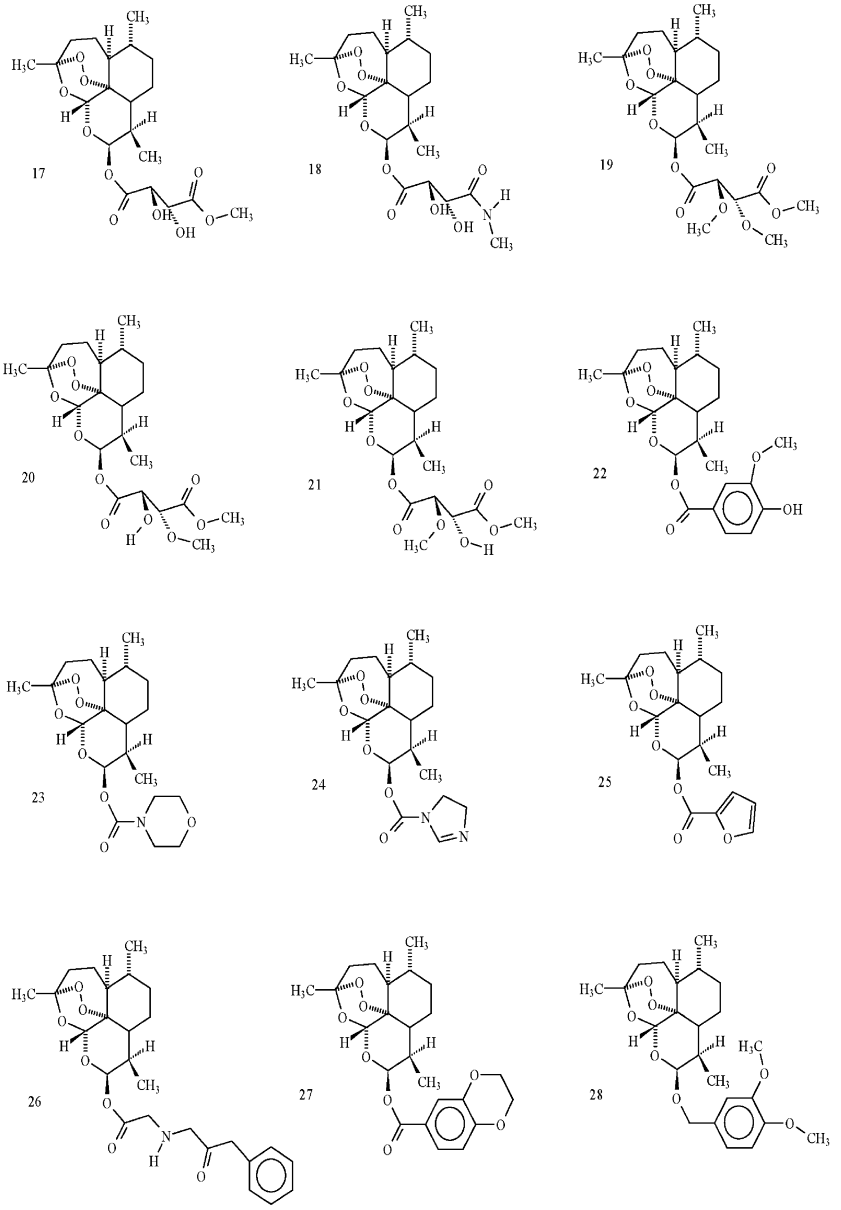

Fig. 2. Dihidroartemisinin derivatives with antimalarial activity unknown (Test Set).

alternation conditional expectations, and projection pursuit regression, on a series of diastereomeric dihydroartemisinin α-alkylbenzylic ethers using Mulliken's atomic net charges, obtained in the PM3 semiempirical molecular orbital approach, as descriptors [8].

In this article, we present a quantum chemical and multivariate study of 16 Artemisinin derivative compounds reported in the literature as showing different degrees of antimalarial activity against *P. falciparum* resistant to mefloquine [9]. In a first step, we built the geometry of the compounds with the strategy described in the Section 2.2 followed by the calculation of the molecular descriptors (See Section

Table 1
Calculated and experimental values of the 1,2,4-trioxane ring parameters in Artemisinin (distances in Å and angles in degrees)

Parameters[a]	ZINDO[b]	AM1[c]	PM3[d]	6-31G[e]	3-21G[f]	Experimental[g]
dO1O2	1.2376	1.2882	1.5439	1.4467	1.4619	1.475
dO2C3	1.4000	1.4471	1.4026	1.4351	1.4405	1.417
dC3O4	1.3960	1.4267	1.4283	1.4347	1.4359	1.448
dO4C5	1.3929	1.4160	1.4033	1.4026	1.4074	1.388
dC5C6	1.5137	1.5370	1.5551	1.5325	1.5294	1.528
dC6O1	1.4161	1.4683	1.4255	1.4687	1.4772	1.450
AO1O2C3	114.31	112.53	110.34	108.80	107.10	107.6
AO2C3O4	105.37	103.60	104.81	106.76	107.28	107.2
AC3O4C5	115.843	115.48	116.01	117.30	115.67	113.5
AO4C5C6	113.27	113.51	115.20	112.28	112.08	114.5
AC5C6O1	107.29	111.07	113.18	110.91	111.57	111.1
AC6O1O2	118.38	113.74	112.29	113.24	111.29	111.5
DO1O2C3O4	−70.403	−77.80	−73.31	−71.84	−74.67	−75.5
DO2C3O4C5	36.37	42.07	52.70	33.39	32.30	36.3
DC3O4C5C6	17.42	11.40	2.811	25.32	28.29	24.8
DO4C5C6O1	−46.61	−41.77	−40.51	−49.41	−50.86	−50.8
DC5C6O1O2	18.11	12.05	19.94	12.51	9,989	12.3
DC6O1O2C3	40.13	47.05	35.63	46.70	50.33	47.7

[a] The atoms are numbered according to compound 1 in Fig. 1.
[b] Method from Ref. [16].
[c] Method from Ref. [17].
[d] Method from Ref. [18].
[e] Basis sets from Ref. [14].
[f] Basis sets from Ref. [15].
[f] Values from Ref. [20]

2.3). And subsequently, the multivariate methods, of principal component analysis (PCA) and hierarchical clustering (HCA) [10] were used to analyze the data and obtain a relationship between the calculated descriptors and the antimalarial activity, classifying the compounds into two categories, high activity (HA) and low activity (LA). The *K*-nearest neighbor (KNN) [11] and soft independent modeling of class analogy (SIMCA) [12], two well established pattern recognition and classification modeling methods, were used for model building and prediction of new derivatives of the compounds studied.

2. Methods

2.1. The compounds studied

The molecular structure of compounds used in present study are shown in Figs. 1 (Training Set) and 2 (Test Set). The atomic numbering, which we have adopted to study the compounds, is shown in compound 1 (Fig. 1). The dihydroartemisinin derivatives in Fig. 1 were tested in vitro against the human malaria, *P. falciparum* resistant to mefloquine. The $IC_{50's}$ for these compounds correspond to average values of at least three experiments reported by Lin and Miller [9]. In our study, we labeled the compounds synthesized and tested by Lin and Miller (compounds 4–16) along with Artemisinin, artelinic acid, and artemether (compounds 1–3, respectively) in two classes: HA compounds (those with $IC_{50} < 1.00$ ng/ml) and low antimalarial Activity compounds (those with $IC_{50} > 1.00$ ng/ml).

2.2. Molecular modeling

In the molecular modeling step, the point of departure was the construction of the structure of compound 1 (Artemisinin) with aid of the GAUSSVIEW program [13]. Complete geometry optimization was performed with the 3-21G basis set [14]. In order to check the

reliability of the geometry obtained, we also optimized the geometry of compound 1 with 6-31G basis set [15] and using the ZINDO [16], AM1 [17] and PM3 [18] methods. The semiempirical calculations were performed with aid of ZININDO 3.5 and MOPAC 6.0 (using EF and PRECISE keywords [19] programs. Table 1 shows the calculated and experimental values [20] of the 1,2,4-trioxane ring parameters in Artemisinin for the different methods used. In Artemisinin all the methods describe well the torsion angles of the twist boat conformation adopted. It is interesting to note that ZINDO, AM1, and PM3 methods, in general give larger errors in the torsion angles than do the ab initio. The agreement between the ab initio 6-31G and 3-21G results is very good, especially considering that the dO1O2 bond length is closer to experimental value than in any of the semiempirical results.

To the compounds 2–16 (Fig. 1) and 17–28 (Fig. 2) the geometries were built with the optimized geometry of the Artemisinin using also GAUSSVIEW program. The ab initio calculations were carried out, using GUASSIAN 94 program and the DIRECT-SCF method [21].

2.3. Molecular descriptors

In order to perform the calculations of descriptors the conformation which is most stable for a given compound was used, and the following molecular descriptors were calculated: total energy (E_T), highest occupied molecular orbital energy (HOMO), lowest unoccupied molecular orbital energy (LUMO), Mulliken's electronegativity (χ), molecular hardness (MH), molecular softness (MS), dipole moment (μ), atomic charge q_i, $i = 1, 2,\ldots,20$, sum of absolute values of the atomic charges Q_i, $i = 1, 2,\ldots,6$, sum of absolute values of the charges of all atoms in the molecules (Q_T), total surface area (TSA), molecular volume (Vol), molecular refractivity (MR), octanol–water partition coefficient (log P), torsion angle formed by 1, 2, 3, and 4 atoms (DO1O2C3C4), and distance of the atoms 1 to 2 (dO1O2). We also included topological [22] and three-dimensional molecular descriptors [23] with the purpose of representing different sources of chemical information in terms of size, shape of a molecule, symmetry and atom distribution in the molecule. The analysis was started with 172 molecular descriptors selected so that they represent electronic, hydrophobic, and steric features of the antimalarial compounds. These features are supposed to be important for investigating structure–activity relationship (SAR) of dihydroartemisinin derivatives against *P. falciparum* resistant to mefloquine.

The quantum chemical descriptors were calculated using the 3-21G standard basis set available in the GAUSSIAN 94 program [21]. The descriptors TSA, Vol, MR, and log P were calculated with a program for SAR and quantitative structure–activity relationship (QSAR) methods [24], and finally the topological and three-dimensional descriptors were obtained with aid of the 3d-weighted holistic invariant molecular program [25]. In order to obtain the atomic charges, the keyword CHELPG was chosen. This strategy was used to make possible the derivation of the atomic charges from the electrostatic potential. The electrostatic potential is obtained through the calculation of a set of punctual atomic charges so that it represents the possible best quantum molecular potential for a set of points defined around the molecule [26].

2.4. Multivariate analysis

2.4.1. Principal components analysis

PCA [10,27] is widely used to simplify large data sets in a way that patterns and relationships can be readily recognized and understood. The underlying purpose of the technique is the dimension reduction.

The method generates a new set of variables called principal components, PCs, as linear combination of all the initial variables so that the first new variable, PC1, describes the largest variance in the data set, the second new variable, PC2, must be chosen orthogonal (uncorrelated) to the first one and in the direction to describes as much variance left as possible and so on.

The initial data matrix, represented by **X**, is decomposed into two matrices, **T** and **P** where

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \tag{1}$$

where **T**, known as 'scores' matrix, represents the position of the samples in the new coordinate system. The second matrix, **P**, is the 'loadings' matrix and describes how the new axis, i.e. the PCs, are built from the original variables. The samples are mapped through scores and the variables by the loadings in the

Table 2
Training set compounds' more important molecular parameters

Compounds		MS (a.u.) $-$ [1]	TSA (Å)^2	CHI1A	BIC	2K	L1v	Tv	$IC_{50's}$ (ng/mL)[a]
1[b]	Artemisinin	8.792	442.440	0.411	0.773	4.038	4.64	9.098	1.0150
2[b]	Artenilic acid	7.346	662.049	0.421	0.77	7.368	14.657	19.097	2.0818
3[c]	Artemether	12.566	478.94	0.417	0.818	4.651	5.005	9.936	0.3008
4[c]		10.66	675.94	0.427	0.787	8.429	12.31	18.639	0.6968
5[c]		9.976	656.09	0.427	0.801	8.462	8.936	16.087	0.0938
6[c]		9.864	658.9	0.427	0.801	8.462	10.98	17.094	0.2265
7[b]		7.888	646.020	0.418	0.767	7.678	14.004	18.664	1.1475
8[c]		10.257	696.070	0.429	0.786	9.029	11.375	18.194	0.2134
9[b]		10.341	691.539	0.429	0.786	9.029	10.051	17.214	0.1437
10[c]		6.065	730.69	0.429	0.786	9.842	12.726	19.617	0.2297
11[c]		6.076	727.62	0.429	0.786	9.842	10.278	18.238	0.0487
12[c]		8.05	676.03	0.424	0.809	8.156	17.355	21.928	0.3368
13[c]		8.104	677.549	0.424	0.809	8.156	13.361	18.505	0.3353
14[b]		7.797	640.82	0.421	0.789	7.616	12.31	18.505	2.5350
15[b]		7.853	643.4	0.421	0.789	7.616	10.066	16.406	1.2308
16[b]		6.008	672.288	0.424	0.774	8.716	13.049	18.412	1.3470

[a] In vitro antimalarial activity of artemisinin derivatives against *P. falciparum* resistant to mefloquine.
[b] Low antimalarial activity.
[c] High antimalarial activity.

new low dimensional vector space defined by the principal components.

2.4.2. Hierarchical cluster analysis

HCA [10] has become, together with principal components, another important tool in multivariate data analysis. Its primary purpose is to display the data in such a way as to emphasize its natural clusters and patterns in a two-dimensional space. The results are presented in the form of dendograms. In HCA, the distances between samples or variables are calculated and compared through the similarity index which ranges from zero, i.e. no similarity and large distance among samples, to one, for identical samples.

2.4.3. The K-nearest neighbors method

The KNN method [10,11] classifies the objects based on distance comparison among them. The multivariate Euclidean distances between every pair of samples with known class membership is calculated. The closest K samples are used to build the model. The optimal K is determined by crossvalidation applied to the training set samples. The classification of a test samples is determined based on the

Table 3
Correlation matrix of the seven descriptors responsible for the classification into higher and lower antimalarial activity

Variables	Variables						
Variables	MS	TSA	CHI1A	BIC	2K	L1v	Tv
MS	1.0000	−0.4235	−0.0805	0.4851	−0.4075	−0.2346	−0.5025
TSA	−0.4235	1.0000	0.8799	−0.0497	0.9873	0.4352	0.8723
CHI1A	−0.0805	0.8799	1.0000	0.1753	0.9173	0.1762	0.6277
BIC	0.4851	−0.0497	0.1753	1.0000	−0.0730	−0.0806	−0.1061
2K	−0.4075	0.9873	0.9173	−0.0730	1.0000	0.3388	0.8086
L1v	−0.2346	0.4352	0.1762	−0.0806	0.3388	1.0000	0.7244
Tv	−0.5025	0.8723	0.6277	−0.1061	0.8086	0.7244	1.0000

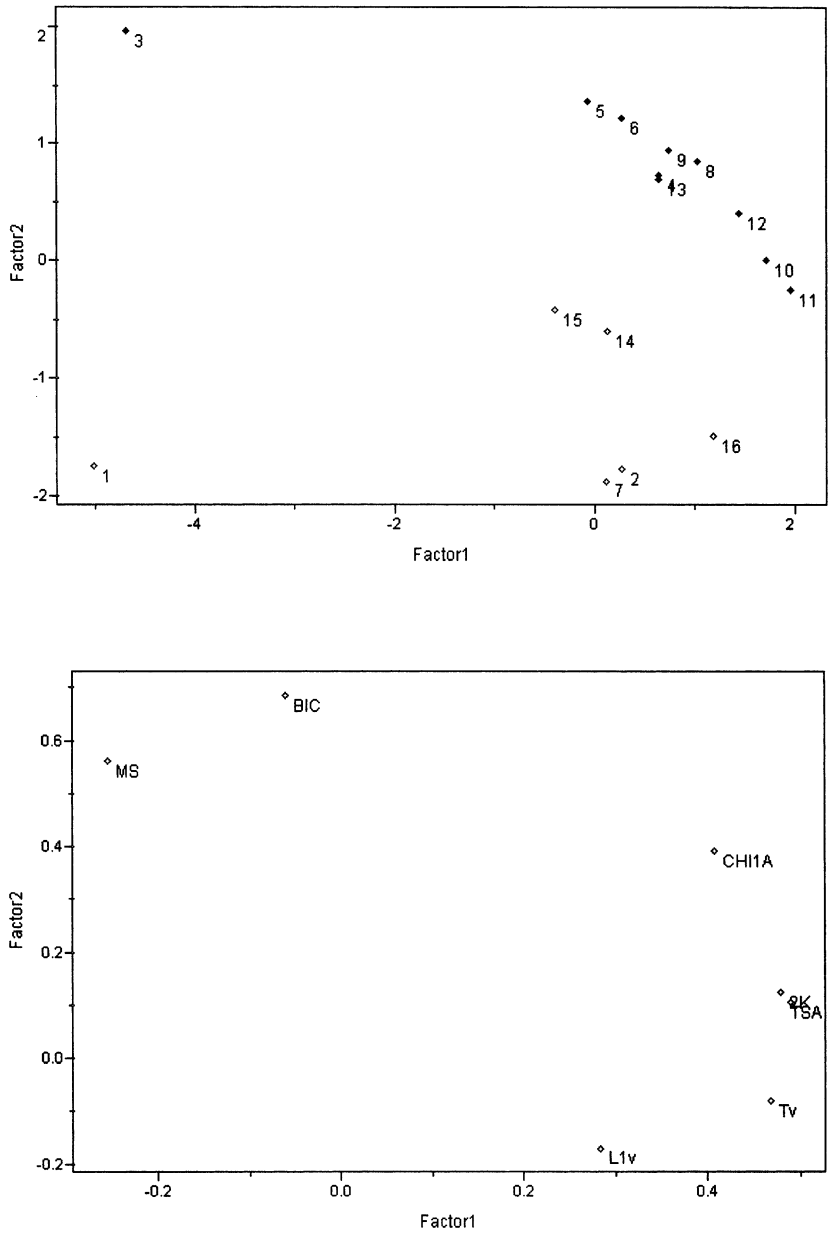

Fig. 3. (a) Plot of the first two PC score vectors (PC1 and PC2) for the dihydroartemisinin derivatives with antimalarial activity (Training Set). The PC analysis leads to a separation in two groups: High activity (HA) and Low activity (LA). (b) Plot of the first two PC loading vectors (PC1 and PC2) for the seven descriptors responsible for the separation of the dihydroartemisinin derivatives (Training Set).

multivariate distance of this sample with respect to the K samples in the training set. In this method no assumption is made about the size and shape of the training set classes.

2.4.4. Soft independent modeling of class analogy method

The SIMCA method [10,12] builds principal components models for each class in the training

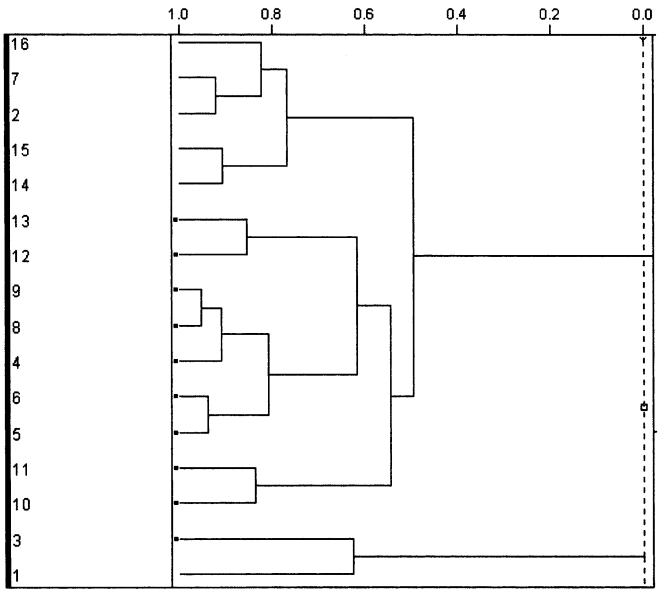

Fig. 4. HCA for dihydroartemisinin derivatives (Training Set).

set. In this method, a multidimensional (determined by the number of PCs necessary to describe the class) box is built for each class, which means that the shape and position of the samples in the classes are taken into account. The classification of a test sample is achieved by determining which space the sample occupies. It can be a member of one, more than one or none of the classes (boxes).

The multivariate analysis were performed with PIROUETTE program [28].

3. Results and discussion

In order to give each variable equal weight in the analysis before applying the PCA, HCA, KNN, and SIMCA methods, each one of the variables was auto-scaled. In a first step, HCA and PCA analysis were carried out for these 16 compounds shown in Fig. 1 (Training Set). After several attempts to obtain a good discrimination of the compounds in question into high activity and low activity classes, the best separation was obtained with a small set of variables. They are: MS, TSA, Randic's index, path 1 — molecular connectivity (CHI1A), bond information index (BIC), shape index based on paths of length 2 of Kier (2K) [22], and the directional (related to molecular size, dimension: axis 1 and weight: van der Waals' volume) and non-directional (related to linear contribution to the total molecular size and weight: van der Waals' volume) 3D-WHIM indices (L1v and Tv), respectively [23]. The data set can be see in Table 2 together with the experimental activities. Table 3 shows the correlation matrix of the seven

Table 4
The results of the multivariate methods for the compounds in Fig. 2 (Test Set) (HA = high activity and LA = low activity)

Compounds	PCA	KNN	SIMCA
17	LA	LA	LA
18	LA	HA	LA
19	LA	HA	LA
20	HA	HA	LA
21	HA	HA	LA
22	HA	HA	HA
23	LA	HA	0.0
24	LA	LA	LA
25	HA	LA	LA
26	LA	LA	LA
27	LA	LA	LA
28	HA	HA	HA

descriptors responsible for the classification into higher and lower antimalarial activity.

The results of the calculation the first three PCs explained 92.097% of the total variance in the data as follows: PC1 = 57.257, PC2 = 21.568, PC3 = 13.272%. The plots of the scores and loadings for the first two PCs are shown in Fig. 3 From Fig. 3, we can see that PC2 discriminates between HA-compounds: 3, 4, 5, 6, 9, 10, 11, 12, and 13 and LA-compounds: 1, 2, 7, 14, 15, and 16. Still in the Fig. 3 the compounds 1 and 3 due the their structural features appear distant of the other compounds of high activity and low activity, respectively.

The results of the HCA analysis are similar to the PCA analysis and are displayed in the dendogram showed in Fig. 4 We can see that the two classes (high activity and low activity) are the same obtained by PCA (Fig. 3). The compounds 1 and 3 due to their structural features (as mentioned before) form one cluster and according to PCA analysis are classified into low activity (compound 1) and high activity (compound 3) classes.

Starting from the separation of the 16 compounds (Training Set), we proposed the new 12 compounds (Test Set) for analysis, for which the antimalarial activity against *P. falciparum* resistant to mefloquine are still unknown. With the application of the PCA method, we obtained a classification into two classes: High Activity (compounds: 20, 21, 22, 25, and 28) and Low Activity (compounds: 17, 18, 19, 23, 24, 26, and 27).

In the construction of the KNN model were used six nearest neighbors in the training set. With the application of the KNN model for the test set also two classes were predicted: HA (compounds: 18, 19, 20, 21, 22, 23, and 28) and LA (compounds: 17, 24, 25, 26, and 27).

The SIMCA model used two PCs for the low activity class and three PCs for the high activity class. The SIMCA model applied for the test set also predicted two classes: HA (compounds: 22 and 28) and LA (compounds: 17, 18, 19, 20, 21, 24, 25, 26, and 27).

Two compounds were predicted as of high activity by the three methods (compounds 22 and 28) and the results of this prediction are summarized in Table 4.

4. Conclusions

The methods of multivariate analysis (PCA and HCA) shows that the 16 Dihidroartemisinin derivatives studied here can be classified into two classes according to their degree of antimalarial activity against *P. falciparum* resistant to mefloquine. The variables (descriptors) MS, TSA, CHI1A, BIC, 2K, L1v, and Tv are those responsible for the discrimination between the derivatives with higher and lower antimalarial activity. The construction of KNN and SIMCA models was useful in the prediction of activity of new derivatives against *P. falciparum* resistant to drug cited. Among the 12 compounds tested, two have been predicted as of high activity (compounds 22 and 28, Fig. 2).

Acknowledgements

We gratefully acknowledge the financial support from the Brazilian agencies CNPq and FAPESP. We employed computing facilities at the IQ-USP-São Carlos, CISC-São Carlos, IQ-Araraquara, and at the CENAPAD-UNICAMP.

References

[1] Y.-J. Rong, Y.-L. Wu, J. Chem. Soc. Perkins Trans. I (1993) 2149–2150.
[2] P.M. O'Neill, D.J. Willock, S.R. Hawley, P.G. Bray, R.C. Storr, S.A. Ward, B.K. Park, J. Med. Chem. 40 (1997) 437–488.
[3] C. Thomson, M. Cory, M. Zerner, Int. J. Quant. Chem.: Quant. Biol. Symp. 18 (1991) 231–245.
[4] G. Bernardinelli, C.W. Jefford, D. Maric, C. Thomson, J. Weber, Int. J. Quant. Chem.: Quant. Biol. Symp. 21 (1994) 117–131.
[5] M.A. Avery, F. Gao, W.K.M. Chong, S. Mehrotra, W.K. Millhous, J. Med. Chem. 36 (1993) 4264–4275.
[6] M.A. Avery, S. Mehrotra, T.L. Johnson, J.D. Bonk, J.A. Vroman, R. Miller, J. Med. Chem. 39 (1996) 4144–4155.
[7] H.U. Suter, D.M. Maric, J. Weber, C. Thomson, Chimia 49 (1995) 125–127.
[8] V. Nguyen-Cong, G.V. Dang, B.M. Rode, Eur. J. Med. Chem. 31 (1996) 797–803.
[9] A.J. Lin, R.E. Miller, J. Med. Chem. 38 (1995) 764–770.
[10] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, Wiley, New York, 1998.

[11] B.R. Kowalski, C.F. Bender, J. Am. Chem. Soc. 94 (1972) 5632–5639.
[12] S. Wold, Pattern Recognition 8 (1976) 127–139.
[13] Gaussian, Inc., gaussview 1.0., Pittsburgh PA, 1997.
[14] J.S. Binkley, J.A. Pople, W.J. Hehre, J. Am. Chem. Soc. 102 (1980) 939–946.
[15] W.J. Hehre, R. Ditchfield, J.A. Pople, J. Chem. Phys. 56 (1972) 2257–2261.
[16] J.D. Baker, M.C. Zerner, Chem. Phys. Lett. 175 (1990) 192–196.
[17] M.J.S. Dewar, J. Am. Chem. Soc. 107 (1985) 3902–3909.
[18] J.J.P. Stewart, J. Comput. Chem. 10 (1989) 209–264.
[19] J.J.P. Stewart, MOPAC — a semiempirical molecular orbital program, J. Comput.-Aid. Mol. Des. 4 (1990) 1–105.
[20] I. Leban, L. Golic, Acta Pharm Jugosl. 38 (1998) 71–77.
[21] M.J. Frisch, G. Trucks, M. Head-Gordon, P.M.W. Gill, B.G. Johnson, M.A. Robb, J.R. Cheesxman, T.A. Keith, G.A. Peterson, J.A. Montgomery, K. Raghavachari, M.A. Al-Lahan, V.G. Zakrzewski, J.V. Ortiz, J.B. Foresmam, J. Cioslowski, B.B. Stefanov, A. Norayakkara, M. Challocombe, C.Y. Peng, P.Y. Ayala, W. Chen, M.W. Womg, J.L. Andes, E.S. Replogle, R. Gomperts, R.L. Martin, D.J. Fox, J.S. Binkley, D.J. Defrees, J. Baker, J.P. Stewart, M. Head-Gordon, C. Gonzalez, J.A. Pople, GAUSSIAN 94, Gaussian, Inc., Pittsburgh PA, 1995.
[22] H. van de Waterbeemd, Chemometric Methods in Molecular Design, VCH, New York, 1996.
[23] R. Todeschini, P. Gramatica, Quant. Struct.-Act. Relat. 16 (1997) 113–119.
[24] ChemPlus, Extension for HyperChem. Molecular Modeling for Windows HyperClub Inc., 1993.
[25] R. Todeschini, P. Gramatica, WHIM-3D 3.3 Milano, 1997.
[26] C.M. Breneman, K.B. Wiberg, J. Comput. Chem. 11 (1990) 361–373.
[27] R.A. Johnson, D.W. Wichern, Applied Multivariate Stastistical Analysis, Prentice-Hall, New Jersey, 1982.
[28] Infometrix, Inc., PIROUETTE 2.02., Woodinville, 1996.