Article

J. Braz. Chem. Soc., Vol. 23, No. 12, 2183-2190, 2012.
Printed in Brazil - ©2012 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Structure-Activity Relationship Study of Rutaecarpine Analogous Active Against Central Nervous System Cancer

Gabriel R. Martins, Hamilton B. Napolitano, Lilian T. F. M. Camargo and Ademir J. Camargo*

Unidade de Ciências Exatas e Tecnológicas, Universidade Estadual de Goiás, Campus Henrique Santillo, BR 153, km 98, 75001-970 Anápolis-GO, Brazil

Com o objetivo de relacionar os descritores geométricos e eletrônicos dos derivados análogos da rutaercarpina com a atividade biológica contra o câncer do sistema nervoso central (CNS), cálculos de química quântica molecular em nível B3LYP/6-31(d) e análise estatística foram realizados para os 21 derivados análogos da rutaecarpina. Dos 86 descritores moleculares calculados, 5 foram selecionados na construção do modelo de análises de componentes principais (PCA). A componente PC1, a qual responde por 46,11% da variância total dos dados, foi capaz sozinha de discriminar completamente os compostos em duas classes: ativos e inativos. Todos os descritores moleculares selecionados pelo modelo PCA são parâmetros eletrônicos. A análise de agrupamento hierárquico (HCA) foi também aplicada aos descritores selecionados pelo modelo PCA. Baseado nos 5 descritores selecionados é possível sugerir novos derivados ativos da rutaercarpina para serem sintetizados. Além disso, um modelo de mínimos quadrados parciais para análise discriminante (PLS-DA) supervisionado foi construído e aplicado com sucesso na discriminação dos análogos à rutaercarpina, o qual foi validado usando um conjunto independente de compostos.

In order to relate the geometric and electronic descriptors of the rutaecarpine analogous to their biological activity against cancer of the central nervous system (CNS), molecular quantum chemical calculations at B3LYP/6-31(d) level and statistical analysis were carried out for 21 rutaecarpine analogous. Out of the 86 molecular descriptors calculated, 5 were selected to build the principal component analysis (PCA) model. The PC1 component that accounts for 46.11% of the total variance of the data was alone able to discriminate completely the compounds into two classes: active and inactive. All molecular descriptors selected by PCA model are electronic parameters. The hierarchical cluster analysis (HCA) was also applied on the selected descriptors. Based on the 5 electronic descriptors selected, it is possible to suggest new compounds to be synthesized with activity against CNS cancer. In addition to that, a supervised partial least squares discriminant analysis (PLS-DA) model was built and successfully applied to discriminate rutaecarpine analogues, being validated through an independent test set and considered robust to overfitting.

Keywords: central nervous system cancer, quinazoline-beta-carboline-5-one, B3LYP, rutaecarpine analogous

Introduction

Brain tumors are rare, but their incidence and mortality rates have been increasing over the past decades in several countries, especially among older people.[1-3] Although very little about brain tumors is known, it is believed that genetic factors, hormonal and environmental factors are related to the evolution of this disease.[4,5] The central nervous system (CNS) comprises the brain and spinal cord. Patients who experience personality changes, apathy, early dementia, constant headache or even depression can have a brain tumor.[6,7] Data from the American Cancer Society (ACS) show that about 27% of the childhood cancers are represented by the CNS tumors, especially in developed countries.[8] On the other hand, in most African countries, this type of cancer represents less than 5%.[7] The ACS predicts that in 2012, 22,910 (12,630 men and 10,280 women) CNS tumors will be diagnosed in the United States. These numbers would be probably much higher in case benign tumors were also included in the

*e-mail: ajc@ueg.br

equation. Estimates from ACS foresees that in 2012 at least 13,700 people (7,720 men and 5,980 women) will die of CNS tumors.[8]

Studies show that the interaction between the drug and its receptor site in biological system involves many intermolecular interactions such as electrostatic, hydrophobic, polar and steric factors.[9] It is well know that the drug recognition by the bioreceptor of the biomacromolecule are dependent on the drug structure, including the spatial arrangement of their functional groups, which are complementary to the binding site located in the bioreceptor and also to the electronic parameters. Structure-activity relationship (SAR) indicates the molecular structure modifications that increase the drug effectiveness. In general, reports show that these modifications are made throughout small changes in the leading compound structure, followed by trials in laboratory to quantify the variations in the biological activity due to changes in the molecular structure.

With this purpose, in 2004, Baruah *et al.*[10] isolated from a fruit of a Chinese medicinal plant *Evodia rutaecarpa* the alkaloid rutaecarpine (quinazoline-beta-carboline-5-one), which has been shown to have important medicinal properties. From the leading compound rutaecarpine, it was synthesized 21 rutaecarpine analogous (Figure 1), which were tested *in vitro* and *in vivo* against eight types of human cancer, including the evaluation against CNS cancer using the growth inhibition (GI_{50}) index, i.e., the concentration in μmol needed to reduce the growth of treated cancer cells (U251 cells line) to half of the untreated cancer cells. According to the rutaecarpine GI_{50} index, the compounds shown in Figure 1 can be divided into two classes: active compounds (1, 2, 5, 6, 7, 8, 9, 10 and 11) with GI_{50} less than or equal to 6 μmol, and inactive (3, 4, 12, 13, 14, 15, 16, 17, 18, 19, 20 and 21) with GI_{50} greater than 6 μmol. Furthermore, the natural product has proved to be a great

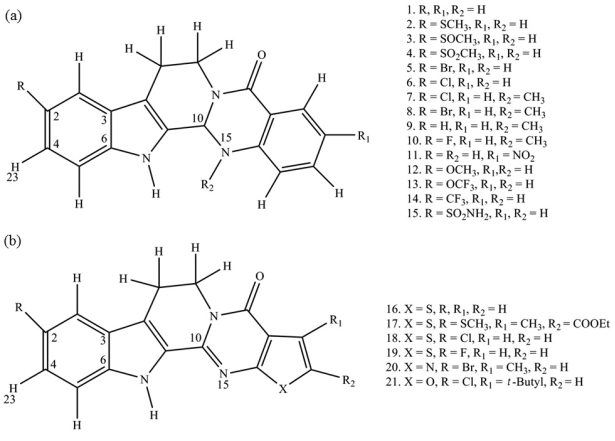

Figure 1. Chemical structures of the rutaecarpine analogous.

source of drugs and inspiration for drug discoveries.[11,12] and the majority of FDA approved drugs are inspired or derived from natural products.

Many research groups working in this area have increasingly been using computational molecular modeling in order to shorten the development and optimization process of a new chemical compound.[13] In this sense, the major aim of QSAR/SAR (quantitative structure-activity relationship) is to establish a relationship between molecular descriptors and biological activity. The molecular descriptors employed in the QSAR/SAR analyses can be theoretical descriptors, derived from quantum chemistry calculations, empirical or derived from readily available experimental characteristics of the structures.[14] Descriptors derived from quantum chemistry calculation are more appropriate to describe the electronic effect than those derived from empirical method.[15]

The aim of this work was to investigate the relationship between the calculated molecular descriptors (geometrics and electronics) for 21 rutaecarpine analogous synthesized and tested against CNS cancer using quantum chemical methods and principal component analysis (PCA) to make the statistical analysis. In addition, a partial least squares discriminant analysis (PLS-DA) model was developed for classifying molecules as active or inactive.

Methodology

The molecular conformation can affect many quantum chemical descriptors[16] and, in addition, the spatial arrangement of the molecular functional groups must be complementary to the binding site of the receptor. As some rutaecarpine analogous have several conformations in the substituent groups, it was necessary to carry out a systematic conformational search on those molecules. This was carried out using HyperChem 7.15 program[17] with the Austin Model 1 (AM1)[18] semi-empirical method. For each analogous derivative, the conformation with lower energy, that most closely resembles the most stable conformation of the most active derivative, was further selected for the full geometric optimization using the Gaussian 03 program[19] with the exchange-correlation hybrid functional B3LYP[20] and the 6-31G(d) basis set.

From the optimized structures, the following molecular descriptors were obtained using the Gaussian 03 software at the B3LYP/6-31G(d) level: the frontier molecular orbital energies E_{HOMO} (the highest occupied molecular orbital energy) and E_{LUMO} (the lowest unoccupied molecular orbital energy), bond angles (A), dihedral angles (D) and the electric dipole moment (μ) calculated as $\mu = |\boldsymbol{\mu}|$, where $\boldsymbol{\mu}$ is given by

$$\mu = \int \rho(\mathbf{r})\, \mathbf{r} d\mathbf{r} \quad (1)$$

and $\rho(\mathbf{r})$ stands for electrical charge density, Mulliken electronegativity (χ) calculated as

$$\chi = \frac{1}{2}\left(-E_{HOMO} - E_{LUMO}\right) \quad (2)$$

energy gap (Δ) obtained as

$$\Delta = E_{LUMO} - E_{HOMO} \quad (3)$$

where E_{HOMO} and E_{LUMO} are the same as above, and hardness (η) defined as

$$\eta = \frac{1}{2}\left(-E_{HOMO} + E_{LUMO}\right) \quad (4)$$

The bond order indexes were calculated using the NBO[21,22] program, which is part of the Gaussian 03 package. The partial atomic charges used in this work were the charges derived from molecular electrostatic potential using the CHELPG scheme by Breneman and Wiberg.[23] As the initial interaction between the ligand and the active site has an electrostatic nature, the partial atomic charges, derived from molecular electrostatic potential with CHELPG scheme, are more suitable for a SAR study and were used in this work. The partition coefficient (log *P*), molecular volume (V) and the average polarizability (α) were calculated using the HyperChem 7.15 program[17] with the AM1 method.[18]

The extraction of relevant information from the chemical experiment involves the analysis of a large number of variables. Often, a small number of these variables contains the most significant chemical information, while the most of them add little or nothing to the interpretation of the results from the chemical point of view.[24] Fisher weights[25] is a statistical technique used to provide a measure of the discriminating power of a descriptor in order to classify compounds in active and inactive classes. Fisher weights for those categories, $w_i(A,I)$, are defined as the ratios of the square of the interclass means to the sum of the intraclass variances, i.e.,

$$w_i(A,I) = \frac{\left[\bar{x}_i(A) - \bar{x}_i(I)\right]^2}{s_i^2(A) + s_i^2(I)} \quad (5)$$

where $\bar{x}_i(A)$ is the mean value of the descriptor *i* for the class *A* (active compounds), $\bar{x}_i(I)$ is the mean value of the descriptor *i* for class *I* (inactive compounds), and $s_i^2(A)$ and $s_i^2(I)$ are the variances for the classes *A* and *I*, respectively. The best descriptors to discriminate the two classes are those with large values of Fisher weights. Therefore, the descriptors with higher values were selected for PCA analysis,[26,27] an unsupervised classification method that reduces the dimensionality of a data set, explaining the variance-covariance structure. This is achieved through linear transformations of the original data set of variables into a smaller number of uncorrelated significant principal components (PCs). Geometrically, this transformation represents the rotation of the original coordinate system, and the direction of the maximum residual variance is given by the first PC axis. The second PC, orthogonal to the first one, has the second maximum variance and so on. Usually, only the first few PCs account for the greatest amount of the total data variance and can be utilized to represent the whole data set in a simpler manner.[28]

Using the selected descriptors by the Fisher weights and PCA analysis, the compound can be grouped based on its similarity using the hierarchical cluster analysis (HCA). The similarity S_{ij} between compound *i* and *j* can be computed using the equation 6:

$$S_{ij} = 1.0 - \frac{d_{ij}}{d_{max}} \quad (6)$$

where d_{ij} is the distance between the compounds *i* and *j*, and d_{max} is the maximum distance observed between all compounds. Thus, the two most distant points in the distance matrix have similarity zero and identical points have similarity 1.0. The hierarchical classification starts by assuming that each point is a group. After that, each point is linked to the next most similar to it. Then, the average point for each point pair is calculated and its link is made to the next most similar average point. This procedure is repeated until to form one single group. The result of this procedure is a diagram called dendrogram. There are several procedures to group the compounds hierarchically such as single linkage, complete linkage, centroid linkage, incremental linkage, etc.[29] In this work, the Euclidian distance as described in equation 7,[29] was used:

$$d_{ij} = \left[\sum_{k=1}^{nd}\left(x_{ik} - x_{jk}\right)^2\right]^{1/2}, \quad (7)$$

where d_{ij} is the distance between the compounds *i* and *j*, *nd* stands for the number of descriptors, and x_{ik} and x_{jk} stand for descriptor x_k for compound *i* and *j*, respectively.

HCA is also an exploratory technique generally used to validate the PCA analysis. HCA results are shown as a dendrogram, which allow us to visualize the clusters and relationship between the compounds. In a dendrogram produced by HCA analysis, the vertical lines represent the compounds and the horizontal lines represent the similarity

between them. The final result of a dendrogram analysis allows us to see how the samples are grouped together and also observe a similarity relationship between the groups. Small distances indicate that samples have some similarity. In principle, it is expected that the points representing the active compounds are grouped in a limited area in the score plot from PCA model, while the points representing the inactive compounds are plotted in different regions of the score plot.

PLS-DA[30] is a multivariate inverse least squares discrimination method used to classify samples and has found importance in some recent applications in QSAR.[31,32] For each class, a model is set up according to correlation $\hat{\boldsymbol{c}} = \boldsymbol{T} \cdot \boldsymbol{q}$, where $\boldsymbol{T}$ is a matrix with the PLS scores obtained from the original data and $\boldsymbol{q}$ is a vector, the length equaling the number of significant latent variables (LVs), and $\hat{\boldsymbol{c}}$ is a class membership function; this is obtained by PLS regression from an original $\boldsymbol{c}$ vector, whose elements are called dummy variables, i.e., they have values of 1 if an object is a member of a class (active) and 0 otherwise (inactive), and an $\boldsymbol{X}$ matrix consisting of the original preprocessed data. The closer each predicted element is to 1, the more likely an object is to be a member of a particular class. In this work, the commonly used threshold value of 0.5 was adopted. All the normal procedures of training and test sets, and cross-validation, can be used with PLS-DA. The predictive ability of the model was also quantified in terms of the Q^2 which is defined as

$$Q^2 = 100 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

where y_i and $\hat{y}_i$ are the observed and predicted values for sample i, respectively. $\bar{y}$ is the observed mean value. The major difference between Q^2 and the normal squared correlation coefficient (R^2) is that the former may also assume negative values, indicating that the model has worse predictive ability than using the mean value as predicted value for each compound. Q^2 should be > 0.5 for the model to be considered to have reasonable practical predictive performance.[33] In this work, the PLS-DA analysis was carried out with PLS_Toolbox 2.1 from Eigenvector Research, Inc.[34]

Results and Discussion

Fisher weights were estimated for 86 geometric and electronic descriptors calculated using quantum chemical methods for 21 rutaecarpine analogous derivatives. 12 out of 86 descriptors were selected by Fisher weights as being relevant to the discrimination of the active and inactive classes. The selected descriptors by Fisher weights are C_{20}, C_{21}, C_{23}, C_{25}, A_2, D_1, $B_{2,4}$, $B_{3,6}$, $B_{6,9}$, $B_{7,11}$, $B_{10,15}$ and $B_{13,14}$, where C, A, D, and B stand for partial charges, angle, dihedral angle, and bond orders, respectively, and the subscripts stand for atomic numbering as shown in Figure 1. The Fisher weight showed to be very useful in this study, allowing reduction of the dimensionality of the data set, which makes easier the subsequent PCA analysis. Before applying the PCA analysis on data set selected by Fisher weights, all selected data were autoscaled to unit variance, i.e., each variable is mean centered and then divided by its standard deviation. This data preprocessing is necessary to remove any inadvertent weighting that arises from arbitrary units. This procedure ensures that all descriptors have the same importance in the statistical analysis. This is important in structure-activity relationship studies since all variables have the same importance and should be compared on the same scale. Working on the 12 descriptors selected by Fisher weights and trying different combinations of them, several different PCA model were tested. A good PCA model should use the fewest descriptors as possible and provide a good discrimination between active and inactive compounds with the lowest numbers of PCs.

The best PCA model was obtained using five descriptors: partial charge on atom 4 (C_4), partial charge on atom 23 (C_{23}), bond orders between atoms 2 and 4 ($B_{2,4}$), atoms 3 and 6 ($B_{3,6}$), and atoms 10 and 15 ($B_{10,15}$). As can be seen in Figure 2, PC1 and PC2 are able to discriminate all the 21 compounds into two classes: active and inactive. PC1 and PC2 account for 46.11 and 35.00% of the total variance, respectively, totalizing 81.11%. Figure 2 displays the score plots and Figure 3 displays the loading plots for this PCA model.

It is worth noting that score plots are related to the samples (compounds) and the loading plots are related to the molecular descriptors. Thus, these two plots should be analyzed together. As can be seen in Figure 2, PC1 alone is responsible for the perfect discrimination of the compounds into two groups. Table 1 shows the positive scores to active compounds with values greater than 0.3, while the inactive compounds have score values lower than 0. Therefore, in this work, the PC1 component can be considered similar to a discriminant function as in a PCA discriminant analysis (PCA-DA).[35]

The contribution of each descriptor for the PCA model is shown in Figure 3 and Table 2. These descriptors are linearly combined to produce the PC1 scores, as shown in equation 8:

$$\mathrm{PC1} = 0.192C_4 - 0.341C_{23} + 0.569B_{2,4} + 0.527B_{3,6} - 0.495B_{10,15} \quad (8)$$

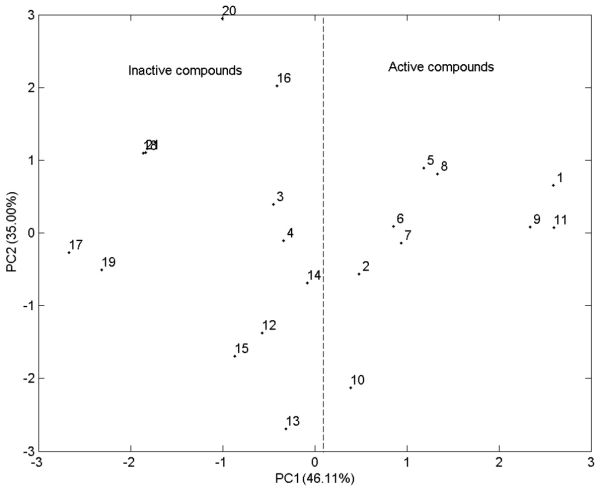

Figure 2. Scores plot for the 21 rutaecarpine analogous using the MATLAB™ software (The Math Works, Natick, USA). The active compounds are placed on the right of the dashed vertical line and the inactive compounds are on the left.

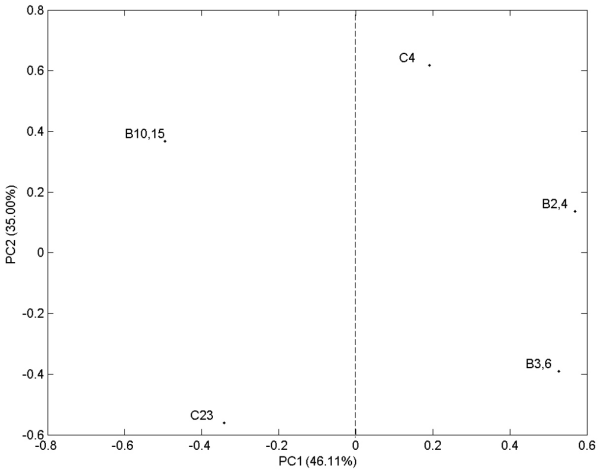

Figure 3. Loading plot of the molecular descriptors selected by PCA model.

The theoretical values obtained from the quantum chemical calculation for the five selected descriptors used to build the PCA model are also included in Table 1. It can be seen that all selected descriptors by PCA model are electronic descriptors. This observation suggests that the electronic properties are really relevant for the rutaecarpine analogous mechanism of action against CNS.

All selected descriptors in the PCA model have high loading values in the PC1 component as can be seen in Table 2. This means that all selected descriptors are important to explain the rutaecarpine derivative activity. As mentioned before, the charges derived from electrostatic potential stand for long range interaction. In this sense, it can be postulated that the atomic partial charges C_4 and C_{23} are important to explain the interaction of rutaecarpine derivatives with the active site. As the scores of the active

Table 1. Scores for PC1 and molecular descriptor values not autoscaled selected by PCA model. The listed descriptor values were obtained at B3LYP/6-31G(d) level of theory. See text for the meanings of the used symbols

Compounds	PC1 score	Selected descriptors by PCA model					Activity
		C_4	C_{23}	$B_{2,4}$	$B_{3,6}$	$B_{10,15}$	
1	2.593	–0.0475	0.0845	1.3693	1.2497	0.9878	active
2	0.480	–0.1464	0.1064	1.3336	1.2449	0.9850	active
3	–0.446	–0.0470	0.1200	1.3301	1.2352	0.9892	inactive
4	–0.341	–0.0815	0.1269	1.3339	1.2371	0.9902	inactive
5	1.184	0.0600	0.1118	1.3453	1.2443	0.9858	active
6	0.856	–0.0669	0.1048	1.3393	1.2448	0.9856	active
7	0.938	–0.0833	0.1067	1.3407	1.2455	0.9660	active
8	1.330	0.0562	0.1113	1.3467	1.2450	0.9662	active
9	2.342	–0.0946	0.0948	1.3697	1.2490	0.9646	active
10	0.388	–0.2198	0.1387	1.3383	1.2491	0.9655	active
11	2.597	–0.0712	0.0939	1.3680	1.2522	0.9750	active
12	–0.572	–0.2266	0.1287	1.3303	1.2398	0.9876	inactive
13	–0.314	–0.1978	0.1800	1.3442	1.2458	0.9895	inactive
14	–0.082	–0.1151	0.1198	1.3231	1.2440	0.9895	inactive
15	–0.865	–0.1565	0.1539	1.3226	1.2420	0.9899	inactive
16	–0.411	–0.0618	0.0912	1.3539	1.2341	1.4808	inactive
17	–2.666	–0.1172	0.1390	1.2985	1.2385	1.4680	inactive
18	–1.862	–0.0948	0.1143	1.3260	1.2328	1.4852	inactive
19	–2.310	–0.2062	0.139	1.3232	1.2358	1.4835	inactive
20	–1.001	0.0954	0.0997	1.3359	1.2319	1.4955	inactive
21	–1.837	–0.0944	0.1137	1.3260	1.2327	1.4739	inactive
Fisher weight	–	0.17	0.78	0.67	0.83	0.56	–

Table 2. PC1 and PC2 loadings for the molecular descriptors selected by PCA model that are able to discriminate the rutaecarpine analogous in active and inactive classes.

Molecular descriptors	Loadings	
	PC1	PC2
C_4	0.192	0.617
C_{23}	–0.341	–0.560
$B_{2,4}$	0.569	0.137
$B_{3,6}$	0.527	–0.391
$B_{10,15}$	–0.495	0.367

compounds are greater than 0.3 and the loadings of the C_4 and C_{23} are positive and negative, respectively, for a new compound to be active it should have high partial charge on atom 4 and low partial charge on atom 23. This means that the electrostatic potential surface on atom 4 should be as positive as possible and the electrostatic potential on atom 23 should be as negative as possible. As can be seen in equation 8, the loadings for the bond order between atoms 2 and 4 ($B_{2,4}$) and the bond order between atoms 3 and 6 ($B_{3,6}$) are positives, while for bond order between atoms 10 and 15 ($B_{10,15}$) are negative. The loading values for these descriptors listed in Table 2 are all positive. This means that for a rutaecarpine analogous to be classified as active compound it should have high $B_{2,4}$ and $B_{3,6}$ values and, in turn, low $B_{10,15}$ value. Bond orders are quantum descriptors related directly to the electron density between two atoms and have direct relation with bond length and the chemical reactivity. Increasing the bond order leads to shortening the bond length and increasing the reactivity of the respective bond. As a result of equation 8 and Table 1, it is possible to infer interactions between some atoms from receptor site with the bond between atoms 2 and 4 and the bond formed between atoms 3 and 6. On the other hand, the bond order between atoms 10 and 15 should be as small as possible, which suggests this bond should not interact with the receptor site.

In summary, the calculation results suggest that increasing the atomic partial charge C_4 and the bond orders between atoms 2 and 4 and atoms 3 and 6 and, in addition, decreasing the charges C_{23} and the bond order between atoms 10 and 15, the probability of the rutaecarpine analogues to become active increases. These features are important in designing new rutaecarpine analogues with anticancer activity. The lack of information about the receptor site makes difficult to describe the interaction of the rutaecarpine analogous with the active receptor site. However, the calculation results allow us to raise some hypothesis about the interaction, pointing out for the prime importance of the benzene ring region and the region between atoms 10 and 15, suggesting that the pharmacophore group is located in this molecular region.

HCA allows us to visualize the clustering formation between the rutaecarpine analogues and the similarity between them. The Figure 4 shows the dendrogram from HCA analysis using the five selected descriptors by PCA model. All selected descriptors were autoscaled to unit variance. It can be observed in Figure 4 that, at the level of similarity 0.3, there are formations of 3 well defined clusters. The active compounds form one cluster on the right of the dendrogram. Just two compounds were classified in wrong clusters. The compound 14 is active, but was classified as inactive and the compound 10 is active but was classified as inactive by HCA analysis. The formation of these well defined clusters shows that the descriptors selected by PCA model were able to accurately assess the similarities that exist among the active and inactive compounds. The similarity among the active compound is about 0.6, showing that these compounds are very similar to each other considering the selected molecular descriptors.

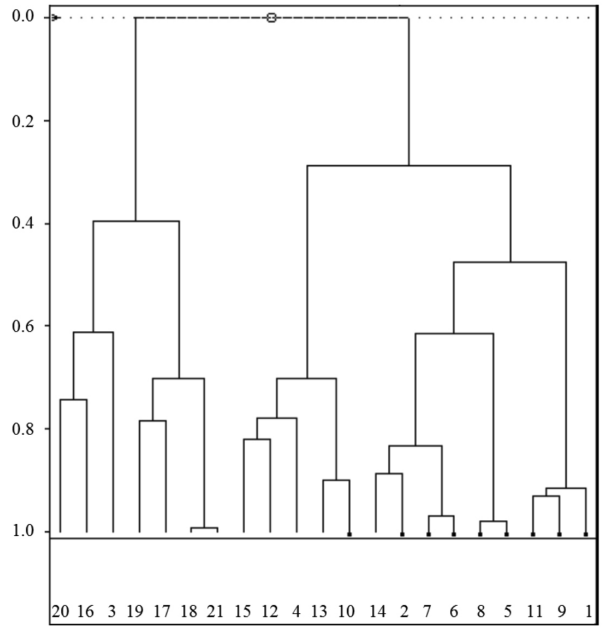

Figure 4. Dendrogram obtained from HCA analysis. The active compounds are on the right cluster, except for the compound 10 that is active and is located on the middle cluster. Also the compound 14 is inactive and is located in the active cluster.

Figure 4 can also be used to make prediction of activity for a new rutaecarpine analogous. For this purpose, it should be calculated the selected descriptors by PCA model (C_4, $C_{23,}$ $B_{2,4}$, $B_{3,6}$ and $B_{10,15}$), autoscaled them and calculated the similarity between the new compound and the compounds of the active group. Thus, if the similarity between the new compound proposed on the basis of the results are greater

than 0.6, it is possible to consider the new rutaecarpine analogous as an active one.

In the development of the PLS-DA model, the 21 rutaecarpine analogues were divided into a training set (15 compounds), used to build the model, and a test set (6 compounds). The compounds used in the test set (1, 4, 9, 17, 21, 6) were randomly selected and have good active/inactive ratio. Vector **y** was built with values 1 for active and 0 for inactive compounds. Predicted values greater than 0.5 were rounded to 1 and predicted values below 0.5 were rounded to 0. In the development of the PLS-DA model, the data were previously autoscaled. The best model was obtained with 3 LVs (minor error of cross validation), accounting for 84.6 and 80.0% of the total variance of **X** and **Y** blocks, respectively. The predictive ability of the model is presented through a confusion matrix (Table 3), a visualization tool typically used in supervised learning, in which each column represents the instances in a predicted class, while each row represents the occurrence in an actual class. As can be seen, all the compounds of the independent test set were correctly discriminated. The model was also used to classify the compounds of the training set with only a false positive. Compound incorrectly predicted was 14. Globally, PLS-DA model correctly classified 100 and 91.7% of the active and inactive compounds, respectively. Estimated regression coefficients without autoscaling the data for PLS-DA model are given by equation 9.

$$y = 0.778C_4 - 9.473C_{23} - 4.241B_{2,4} + 64.551B_{3,6} - 0.119B_{10,15} - 72.723 \quad (Q^2 = 0.80) \qquad (9)$$

Q^2 = 0.80 shows that the equation 9 is relevant to discriminate the compounds into active and inactive classes.

Table 3. Confusion matrix for active/inactive classification obtained with PLS-DA.

		Predicted			
		Training set		Test set	
	Class	Active	Inactive	Active	Inactive
Actual	active	6	0	3	0
	inactive	1	8	0	3

Conclusions

The PCA model result shows that five molecular descriptors are able to completely discriminate the rutaecarpine analogous tested against CNS cancer into active and inactive classes. All selected descriptors are electronic, calculated at B3LYP/6-31G(d) level: the partial charge on atom 4 (C_4), partial charge on atom 23 (C_{23}), bond orders between atoms 2 and 4 ($B_{2,4}$), atoms 3 and 6 (B3,6), atoms 10 and 15 ($B_{10,15}$). PC1 alone is responsible for the compound discriminations and accounts for 46.11% of the total variance of the data. HCA applied on the autoscaled descriptor selected by the PCA model was also able to discriminate the compounds into active and inactive clusters at similarity 0.3. The HCA similarity among the active compounds is greater than 0.5. In addition, a supervised PLS-DA model was build and successfully used to classify rutaecarpine analogues as active or inactive, being validated through an independent test set and considered robust to overfitting. The selected electronic descriptors enable to hypothesize regarding the rutaecarpine analogous mechanism of action and, in addition, can guide us in designing new rutaecarpine analogous with activity against CNS cancer.

Acknowledgments

The authors are grateful to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for their financial support.

References

1. Davis, D. L.; Hoel, D.; Percy, C.; Ahlbom, A.; Schwartz, J.; *Ann. N.Y. Acad. Sci*. **1990**, *609*, 191.
2. Greig, N. H.; Ries, L. G.; Yancik, R.; Rapoport, S. I.; *J. Natl. Cancer Inst*. **1990**, *82*, 1621.
3. Modan, B.; Wagener, D. K.; Feldman, J. J.; Rosenberg, H. M.; Feinleib, M.; *Am. J. Epidemiol*. **1992**, *135*, 1349.
4. Bohnen, N. I.; Kurland, L. T.; *J. Neurol. Sci*. **1995**, *132*, 110.
5. Gold, E. B. In *Reviews in Cancer Epidemiology*; Lillienfeld, A. M., ed.; Elservier: Amsterdam, North-Holland, 1982.
6. Braga, P. E.; Latorre, M. R. D. O.; Curado, M. P.; *Cadernos de Saúde Pública* **2002**, *18*, 33.
7. Parkin D. M.; Kramarova E.; Draper G. J.; Masuyer E.; Michaelis J.; Neglia J.; Qureshi S.; Stiller C. A.; *IARC Sci. Publ*. **1998**, *2*, 102.
8. Siegel, R.; Naishadham, D.; Jemal, A.; *CA-Cancer J. Clin.* **2012**, *62*, 10.
9. Barreiro, E. J.; Fraga, C. A. M.; *Química Medicinal, As Bases Moleculares da Ação dos Fármacos*, 2ª ed; Artmed: São Paulo, Brasil, 2008.
10. Baruah, B.; Dasu, K.; Vaitilingam, B.; Mamnoor, P.; Venkata, P. P.; Rajagopal, S.; Yeleswarapu, K. R.; *Bioorg. Med. Chem*. **2004**, *12*, 1991.
11. Newman, D. J.; Cragg, G. M.; *J. Nat. Prod*. **2007**, *70*, 461.

12. Koehn, F. E.; Carter, G. T.; *Nat. Rev. Drug Discovery* **2005**, *4*, 206.
13. Dias, R. L. A.; Côrrea, A. G.; *Quim. Nova* **2001**, *24*, 236.
14. Karelson, M.; Lobanov, V. S.; *Chem. Rev*. **1996**, *96*, 1027.
15. Cartier, A.; Rivail, J. L.; *Chemom. Intell. Lab. Syst*. **1987**, *1*, 335.
16. Karelson, M.; Labanov, V. C.; *Chem. Rev*. **1996**, *96*, 1027.
17. *HyperChem(TM) Professional 7.51*; Hypercube, Inc.: Gainesville, Florida, USA.
18. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P.; *J. Am. Chem. Soc.* **1985**, *107*, 3902.
19. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; 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, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A.; *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, USA, 2004.
20. Lee, C.; Yang, W.; Parr, R. G.; *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.
21. Foster, J. P.; Weinhold, F.; *J. Am. Chem. Soc.* **1980**, *102*, 7211.
22. Reed, A. E.; Weinhold, F.; *J. Chem. Phys*. **1983**, *78*, 4066.
23. Breneman, C. M.; Wiberg, K. W.; *J. Comput. Chem.* **1990**, *11*, 361.
24. Neto, J. M. M.; Moita, G. C.; *Quim. Nova* **1998**, *21*, 467.
25. Sharaf, M. A.; Illman, D. L.; Kolwalski, B. R.; *Chemometrics*, Wiley, New York, USA, 1986.
26. Arantes F. F. P.; Barbosa, L. C. A.; Maltha, C. R. A.; Demuner, A. J.; Fidêncio, P. H.; Carneiro, J. W. M.; *J. Chemom*. **2011**, *25*, 401.
27. Camargo A. J.; Honório, K. M.; Mercadante, R.; Molfetta, F. A.; Alves, C. N.; da Silva, A. B. F.; *J. Braz. Chem. Soc*. **2003**, *14*, 809.
28. Dehmer, M.; Varmuza, K.; Bonchev, D.; Emmert-Streib, F.; *Statistical Modeling of Molecular Descriptors in QSAR/QSPR*, 1st ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Germany, 2012.
29. Otto, M.; *Chemometrics: Statistical and Computer Application in Analytical Chemistry*; Wiley-VCH: Weinheim, Germany, 1999.
30. Brereton, R. G.; *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*; Wiley: Chichester, UK, 2003.
31. Evers, A.; Hessler, G.; Matter, H.; Klabunde, T.; *J. Med. Chem.* **2005**, *48*, 5448.
32. Carlsson, C.; Harju, M.; Bahrami, F.; Cantillana, T.; Tysklind, M.; Brandt, I.; *Arch. Toxicol.* **2004**, *78*, 706.
33. Schreiber, S. L.; Kapoor, T. M.; Wess, G.; *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*; Wiley-VCH: Weinheim, Germany, 2007.
34. Wise, B. M.; Gallagher, N.; *PLS_Toolbox 2.1 for Use with MATLAB™*; Eigenvector Research, Inc.: Manson, WA, USA, 1998.
35. Defernez, M.; Kemsley, E. K.; *TrAC, Trends Anal. Chem.* **1997**, *16*, 216.

Submitted: June 15, 2012
Published online: December 21, 2012