# A Study on the Antipicornavirus Activity of Flavonoid Compounds (Flavones) by Using Quantum Chemical and Chemometric Methods

Jaime Souza, Jr., Fábio A. Molfetta, Káthia M. Honório, Regina H. A. Santos, and Albérico B. F. da Silva*

Departamento de Química e Física Molecular, Instituto de Química de São Carlos, Universidade de São Paulo, CP 780, 13560-970, São Carlos, SP, Brasil

Received June 16, 2003

The AM1 semiempirical method is employed to calculate a set of molecular properties (variables) of 45 flavone compounds with antipicornavirus activity, and 9 new flavone molecules are used for an activity prediction study. Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), Stepwise Discriminant Analysis (SDA), and K-Nearest Neighbor (KNN) are employed in order to reduce dimensionality and investigate which subset of variables should be more effective for classifying the flavone compounds according to their degree of antipicornavirus activity. The PCA, HCA, SDA, and KNN methods showed that the variables MR (molar refractivity), $B_9$ (bond order between $C_9$ and $C_{10}$ atoms), and $B_{25}$ (bond order between $C_{11}$ and $R_7$ atoms) are important properties for the separation between active and inactive flavone compounds, and this fact reveals that electronic and steric effects are relevant when one is trying to understand the interaction between flavone compounds with antipicornavirus activity and the biological receptor. In the activity prediction study, using the PCA, HCA, SDA, and KNN methodologies, three of the 9 new flavone compounds studied were classified as potentially active against picornaviruses.

## INTRODUCTION

Picornaviruses comprise a family of small, nonenveloped viruses containing a positive-sense RNA genome encased in a protein capsid.[1] The picornavirus family is subdivided into five genera: apthoviruses, cardioviruses, hepatoviruses, rhinoviruses, and enteroviruses. The hepatoviruses, rhinoviruses, and enteroviruses are responsible for a wide array of human illnesses. The enteroviruses, which include the polioviruses, echoviruses, and coxsackieviruses, are associated with poliomyelitis, myocarditis, aseptic meningitis, and encephalitis.[2] Over 100 human rhinovirus serotypes have been identified, and they are considered to be the most frequent cause of the common cold.[3]

Several flavonoids have been shown to inhibit the replication of picornaviruses. Two classes of compounds can be distinguished according to their antiviral spectrum and mechanism of action: in the first class are the chalcones and flavans and in the second class are the flavones. The first compound class (chalcones and flavans) inhibits selectively serotypes of rhinoviruses. Compounds such as 4′-ethoxy-2′-hydroxy-4,6′-dimethoxy-chalcone and 4′,6-dichloroflavan interact directly with specific sites on the viral capsid proteins, thereby liberating the viral RNA.

A second class of compounds consists of flavones and various substituted derivatives (active against a wide range of picornaviruses except Mengo and vesicular stomatitis virus) which were isolated from several plants. They interfere in an early step in the viral RNA synthesis. Although the molecular mechanism is not completely understood yet, they probably inhibit the formation of minus-strand RNA of poliovirus by interacting with one protein involved in the binding of the virus replication complex to vesicular membranes at which poliovirus replication takes place.[4]

The attractive mechanism of action, the pronounced and broad-spectrum antiviral activity, and the lack of resistance-induction by these flavones prompted us to explore this class of flavonoids and establish a link between chemical structure and biological activity.

Structure−activity relationship (SAR) studies with biological molecules can aid in the developing of more effective compounds (by using molecular properties that can be responsible for the biological activity) with the goal of understanding the interaction mechanism between drugs and their biological receptors.[5−8]

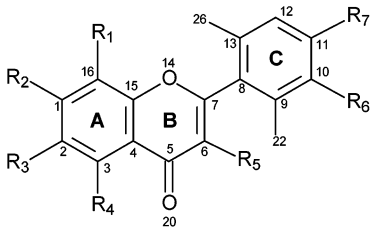
The present work employs the AM1 semiempirical method[9] in order to calculate some molecular properties (descriptors) of 45 flavonoid compounds (reported in the literature as potent and selective antipicornavirus agents[4]) and 9 new ones[10−12] that are used in an activity prediction study. Our objective in this work is to investigate, in a qualitative way, the structure−activity relationship (SAR) of flavone compounds by using quantum chemical descriptors and other molecular properties. Chemometric methods, namely, Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), Stepwise Discriminant Analysis (SDA), and K-Nearest Neighbor (KNN) are employed to analyze the data set. Afterward, these chemometric methods are also used to classify 9 new flavone molecules according to their antipicornavirus activity.

## METHODOLOGY

**Compounds.** The molecular structure of each one of the 45 flavone compounds studied in this work is shown in

* Corresponding author e-mail: alberico@iqsc.usp.br.

10.1021/ci030384n CCC: $27.50 
Published on Web 04/21/2004




| Compound | $R_1$ | $R_2$ | $R_3$ | $R_4$ | $R_5$ | $R_6$ | $R_7$ | $RF_{MNTD}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | H | H | H | H | $OCH_3$ | H | OH | $10^4$ |
| 2 | H | H | $CH_3$ | H | $OCH_3$ | H | OH | $10^{4.5}$ |
| 3 | H | $CH_3$ | H | H | $OCH_3$ | H | OH | $10^{4.5}$ |
| 4 | H | H | $CH(CH_3)_2$ | H | $OCH_3$ | H | OH | $10^{1.5}$ |
| 5 | H | $OCH_3$ | H | H | $OCH_3$ | H | OH | $10^{3.5}$ |
| 6 | $OCH_3$ | H | H | H | $OCH_3$ | H | OH | $10^5$ |
| 7 | H | H | OH | H | $OCH_3$ | H | OH | $10^4$ |
| 8 | H | OH | H | H | $OCH_3$ | H | OH | $10^4$ |
| 9 | H | H | Cl | H | $OCH_3$ | H | OH | $10^{3.5}$ |
| 10 | H | Cl | H | H | $OCH_3$ | H | OH | $10^3$ |
| 11 | H | I | H | H | $OCH_3$ | H | OH | $10^3$ |
| 12 | H | H | $NH_2$ | H | $OCH_3$ | H | OH | $10^3$ |
| 13 | H | H | H | OH | $OCH_3$ | H | OH | $10^3$ |
| 14 | H | OH | H | OH | $OCH_3$ | H | OH | $10^5$ |
| 15 | H | $OCH_3$ | H | OH | $OCH_3$ | H | OH | $10^5$ |
| 16 | H | $OCH_3$ | H | $OCH_3$ | $OCH_3$ | H | OH | $10^4$ |
| 17 | H | OH | H | $CH_3$ | $OCH_3$ | H | OH | $10^4$ |
| 18 | H | $OCH_3$ | H | $CH_3$ | $OCH_3$ | H | OH | $10^4$ |
| 19 | H | OH | $CH_3$ | $CH_3$ | $OCH_3$ | H | OH | $10^5$ |
| 20 | H | OH | H | OH | $OCH_3$ | OH | OH | $10^5$ |
| 21 | H | $OCH_3$ | H | OH | $OCH_3$ | OH | OH | $10^5$ |
| 22 | H | $OCH_3$ | H | OH | $OCH_3$ | $OCH_3$ | OH | $10^5$ |
| 23 | H | OH | $OCH_3$ | OH | $OCH_3$ | H | OH | $10^4$ |
| 24 | H | $OCH_3$ | $OCH_3$ | OH | $OCH_3$ | H | OH | $10^5$ |
| 25 | $OCH_3$ | $OCH_3$ | H | OH | $OCH_3$ | H | OH | $10^5$ |
| 26 | $OCH_3$ | $OCH_3$ | H | OH | $OCH_3$ | OH | OH | $10^4$ |
| 27 | $OCH_3$ | OH | H | OH | $OCH_3$ | OH | OH | $10^5$ |
| 28 | H | OH | $OCH_3$ | OH | $OCH_3$ | $OCH_3$ | OH | $10^3$ |
| 29 | H | H | H | $OCH_3$ | $OCH_3$ | $OCH_3$ | H | 1 |
| 30 | H | H | $OCH_3$ | $OCH_3$ | $OCH_3$ | $OCH_3$ | H | $10^2$ |
| 31 | $OCH_3$ | H | H | H | $OCH_3$ | $OCH_3$ | H | 1 |
| 32 | H | H | OH | H | $OCH_3$ | $OCH_3$ | H | 1 |
| 33 | H | H | Cl | H | $OCH_3$ | $OCH_3$ | H | 1 |
| 34 | H | Cl | H | H | $OCH_3$ | $OCH_3$ | H | 1 |
| 35 | H | $CH_3$ | H | H | $OCH_3$ | Cl | H | 1 |
| 36 | H | H | Cl | H | $OCH_3$ | Cl | H | 1 |
| 37 | H | Cl | H | H | $OCH_3$ | Cl | H | 1 |
| 38 | H | $OCH_3$ | H | OH | $OCH_3$ | $OCH_3$ | H | 1 |
| 39 | H | $OCH_3$ | H | OH | $OCH_3$ | $OCH_3$ | $OCH_3$ | 1 |
| 40 | $OCH_3$ | OH | H | OH | $OCH_3$ | $OCH_3$ | H | 1 |
| 41 | $OCH_3$ | $OCH_3$ | H | OH | $OCH_3$ | $OCH_3$ | H | - |
| 42 | $OCH_3$ | $OCH_3$ | $OCH_3$ | OH | $OCH_3$ | $OCH_3$ | H | 1 |
| 43 | H | $CH_3$ | H | H | $OC_2H_5$ | H | OH | 1 |
| 44 | H | $CH_3$ | H | H | $OCH(CH_3)_2$ | H | OH | 1 |
| 45 | H | $CH_3$ | H | H | $NH_2$ | H | OH | 1 |

**Figure 1.** Chemical structure of the 45 flavone compounds studied (training set) and the antiviral measurement used for the discrimination between active and inactive compounds, which was expressed as the viral strength reduction factor (RF) after 3 days of incubation of the virus in the presence of the maximal nontoxic dose of the compound (MNTD).

Figure 1, and all of the 45 molecules (the training set) can be divided into two groups: Group A, which contains the active compounds (molecules 1, 2, 3, and 5−28 in Figure 1), and Group B, which contains the inactive compounds (molecules 4 and 29−45 in Figure 1). The discrimination between active and inactive compounds was made by using an antiviral measurement, which was expressed as the viral strength reduction factor (RF) after 3 days of incubation of the virus in the presence of the maximal nontoxic dose of the compound (MNTD).[4] The $RF_{MNTD}$ was calculated as the ratio of the viral titer of the virus control to the viral titer in the presence of the MNTD of the test compound.[4] The $RF_{MNTD}$ values for the 45 compounds studied are shown in Figure 1, and compounds with a $RF_{MNTD}$ value $\geq 10^3$ were considered as active ones.[4] The chemical structure of the 9 new compounds used in the activity prediction study is presented in Figure 2.

**Calculation of the Molecular Properties.** The initial molecular geometry of each compound was obtained by using the molecular mechanics method (MM+)[13] of the Hyperchem 4.5 molecular package.[14] Afterward, an additional geometry optimization was carried out for each compound by using the AM1 semiempirical method[9] of the molecular package AMPAC 6.5.[15]

All of the AM1 molecular geometry optimizations were performed by using the Precise keyword. After the conformational study, the molecular properties (variables) of each flavonoid compound studied were calculated by using the more stable structure obtained with the AM1 optimization.

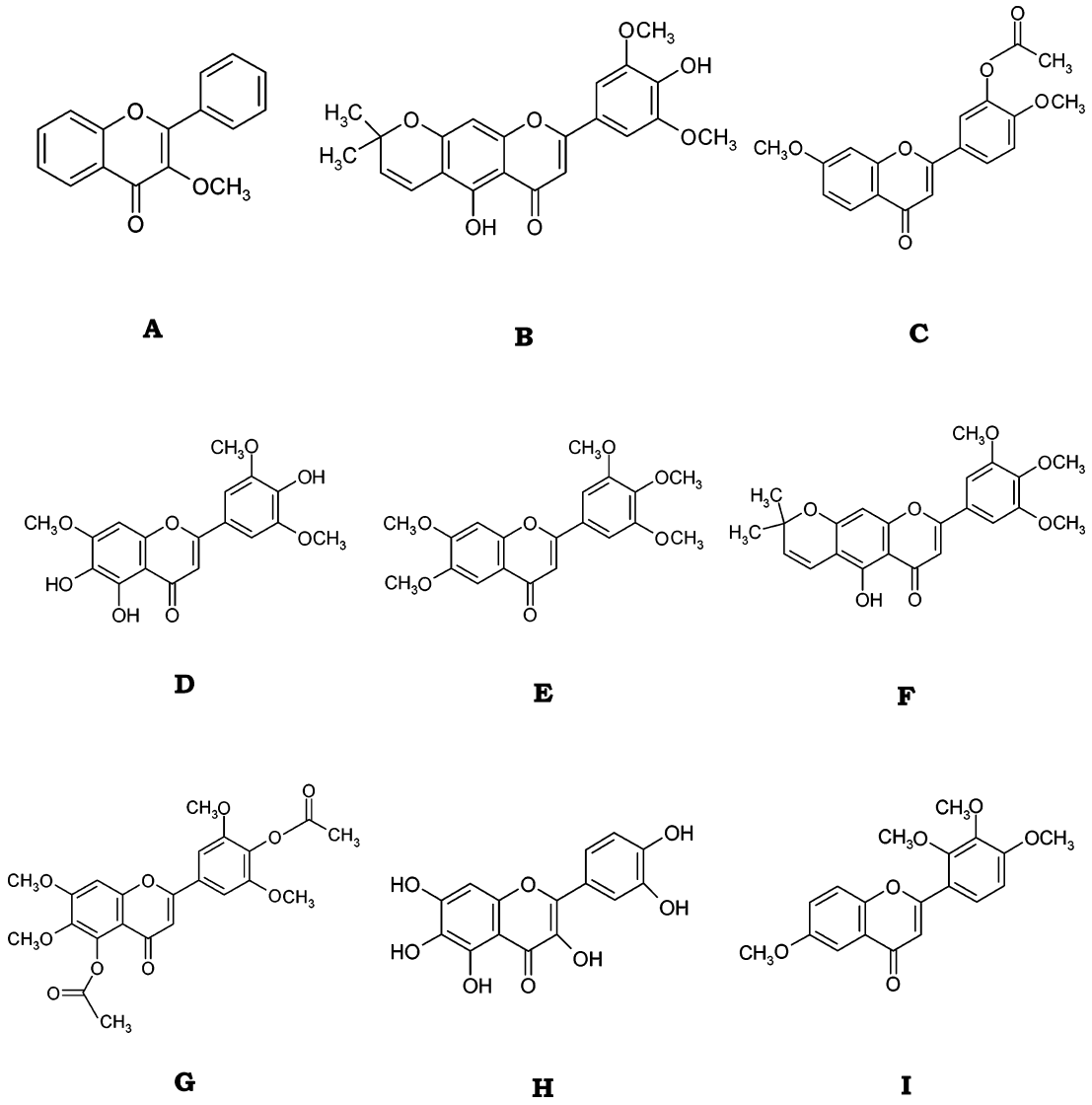



**Figure 2.** Chemical structure of the 9 new flavone compounds used in the activity prediction study.

The biological activity of a drug depends mainly on three different kinds of molecular properties: electronic, steric, and hydrophobic.[16] Our strategy is to calculate as many physicochemical descriptors (parameters) as possible by using available computational packages, as we do not know beforehand which properties are closely related to the biological activity.

In this work the following descriptors were calculated: $E_{HOMO}$ − the highest occupied molecular orbital energy (eV); $E_{LUMO}$ − the lowest unoccupied molecular orbital energy (eV); $\chi$ − Mulliken's electronegativity (eV); $\mu$ − dipole moment (au); $\alpha$ − molecular polarizability (au); $\Delta H_f$ − heat of formation (kcal mol$^{-1}$); MR molar refractivity (Å$^3$); A − molecular surface area (Å$^2$); Vol − volume (Å$^3$); $E_T$ − total energy (eV); $E_{el}$ − electronic energy (eV); E.A. − electronic affinity (eV); log $P$ − partition coefficient; B − bond order; $D_n$ and $A_n$ − dihedral and interatomic angles, respectively; Ln − bond length.

The calculated descriptors were selected so that they represent electronic ($E_{HOMO}$, $E_{LUMO}$, $\chi$, $\mu$, $\alpha$, $\Delta H_f$, MR, E.A., $E_T$, and $E_{el}$), steric (A and Vol), and hydrophobic (log $P$) features of the compounds studied. These features are supposed to be important in the antipicornavirus activity of the flavone compounds studied. The correlation between the molecular properties and biological activity was done by using the pattern recognition methods (PCA, HCA, SDA, and KNN) built-in the computational package Matlab 6.0.[17]

The descriptors $E_{HOMO}$, $E_{LUMO}$, $\mu$, $\alpha$, $E_T$, $E_{el}$, and $\Delta H_f$ were calculated with the AM1 semiempirical method[9] by using the molecular package AMPAC 6.5.[15] The other molecular descriptors were calculated with the HyperChem 4.5 program.[14] The descriptors $\chi$ and E.A. were obtained according to Mulliken's theory,[18] and they are defined as

$$\chi = \frac{1}{2}(\mathrm{IP} + \mathrm{E.A.}) = \frac{1}{2}(-E_{HOMO} - E_{LUMO}) \quad (1)$$

$$\mathrm{E.A.} = -E_{LUMO} \quad (2)$$

## RESULTS AND DISCUSSION

**Principal Component Analysis (PCA).** The principal component analysis (PCA) was first described by Pearson in 1901[19] and by Hotelling in 1933.[20] In the past years, the use of PCA has increased and is often applied in the field of chemometrics.[21] The main purpose of this analysis is

**Table 1.** Values Obtained for the Fisher's Weights for All Calculated Variables

| variable | $W_{Fish}$ | variable | $W_{Fish}$ | variable | $W_{Fish}$ | variable | $W_{Fish}$ | variable | $W_{Fish}$ |
|---|---|---|---|---|---|---|---|---|---|
| A1 | 0.055 | D8 | 0.381 | L15 | 0.002 | B6 | 0.056 | B26 | 0.229 |
| A2 | 0.002 | D9 | 0.039 | L16 | 0.001 | B7 | 0.000 | B27 | 0.000 |
| A3 | 0.197 | D10 | 0.001 | L17 | 0.047 | B8 | 0.047 | B28 | 0.515 |
| A4 | 0.007 | D11 | 0.018 | L18 | 0.025 | B9 | 10.711 | B29 | 0.292 |
| A5 | 0.119 | D12 | 0.018 | L19 | 0.008 | B10 | 0.238 | $E_{el}$ | 0.045 |
| A6 | 0.112 | D13 | 0.038 | L20 | 0.113 | B11 | 0.238 | $E_T$ | 0.017 |
| A7 | 0.004 | L1 | 0.047 | L21 | 0.380 | B12 | 0.031 | $E_{HOMO}$ | 0.000 |
| A8 | 0.005 | L2 | 0.047 | L22 | 0.253 | B13 | 0.118 | $E_{LUMO}$ | 0.211 |
| A9 | 0.051 | L3 | 0.047 | L23 | 0.118 | B14 | 0.020 | $\mu$ | 0.003 |
| A10 | 0.016 | L4 | 0.018 | L24 | 0.079 | B15 | 0.020 | $\alpha$ | 0.155 |
| A11 | 0.001 | L5 | 0.008 | L25 | 0.082 | B16 | 0.238 | $\chi$ | 0.052 |
| A12 | 0.000 | L6 | 0.118 | L26 | 0.022 | B17 | 0.112 | $A$ | 0.362 |
| A13 | 0.075 | L7 | 0.010 | L27 | 0.118 | B18 | 0.153 | Vol | 0.332 |
| D1 | 0.055 | L8 | 0.047 | L28 | 0.130 | B19 | 0.003 | logP | 0.018 |
| D2 | 0.013 | L9 | 0.047 | L29 | 0.112 | B20 | 0.005 | MR | 0.408 |
| D3 | 0.001 | L10 | 0.047 | B1 | 0.176 | B21 | 0.632 | E.A. | 0.205 |
| D4 | 0.036 | L11 | 0.047 | B2 | 0.354 | B22 | 0.020 | | |
| D5 | 0.026 | L12 | 0.047 | B3 | 0.302 | B23 | 0.585 | | |
| D6 | 0.069 | L13 | 0.047 | B4 | 0.585 | B24 | 0.454 | | |
| D7 | 0.004 | L14 | 0.104 | B5 | 0.585 | B25 | 1.373 | | |

grouping correlated variables, generating a new set of variables called principal components (PCs), onto which the data set is projected. These PCs are built as a linear combination of the original variables and have the important property of being completely uncorrelated. In this analysis, the position of the samples in the new coordinate system (PCs) is represented by the score matrix, while the loadings matrix gives the importance (weight) of the original variables in the PCs.

The first new axis, $PC_1$, is chosen in such a direction that maximizes the variance along that axis. The second axis must be chosen orthogonal to the first one and in the direction to describe as much variance left as possible and so on. In this work, before applying the PCA method, each one of the selected variables was autoscaled so that they could be compared to each other on the same scale.

To perform our PCA analysis with the descriptors calculated in this work, it was necessary to reduce the number of descriptors choosing the more relevant ones, as we had a very large number (96) of descriptors (see Table 1), and this amount of information could prejudice our PCA analysis.

This reduction in the number of descriptors was made by using the correlation matrix between the calculated variables and the Fisher's weight.[22,23] This procedure gives the relative importance of each variable. The values of the Fisher's weight for the calculated variables are presented in Table 1.

After reducing the initial set of 96 descriptors (Table 1), we selected only 9 variables that showed significant weight values, i.e., the variables that presented the Fisher's weight above 0.40. These variables were those that possessed a higher ability in the discrimination (separation) between active and inactive molecules.

From the 9 selected variables obtained by using the Fisher's weight, we also attained the correlation matrix among them (see Table 2). From Table 2 we can see that some variables are correlated to each other (we considered correlated variables only those that possess correlation coefficients above 0.65), and, according to the results shown in Tables 1 and 2, only 7 variables can be considered important for the separation between active and inactive compounds.

**Table 2.** Correlation Matrix between the Selected Variables

| | B4 | B5 | B9 | B21 | B23 | B24 | B25 | B28 | MR |
|---|---|---|---|---|---|---|---|---|---|
| B4 | 1.00 | 1.00 | −0.29 | −0.46 | 1.00 | 0.69 | −0.36 | 0.35 | 0.03 |
| B5 | | 1.00 | −0.29 | −0.46 | 1.00 | 0.69 | −0.36 | 0.35 | 0.03 |
| B9 | | | 1.00 | 0.36 | −0.29 | −0.32 | 0.60 | −0.41 | −0.54 |
| B21 | | | | 1.00 | −0.46 | −0.37 | 0.38 | −0.33 | −0.07 |
| B23 | | | | | 1.00 | 0.69 | −0.36 | 0.35 | 0.03 |
| B24 | | | | | | 1.00 | −0.28 | 0.48 | 0.37 |
| B25 | | | | | | | 1.00 | −0.55 | −0.15 |
| B28 | | | | | | | | 1.00 | 0.22 |
| MR | | | | | | | | | 1.00 |

**Table 3.** Calculated Values for the Three Most Important Properties (Variables) that Classify the 45 Flavone Compounds Studied

| compound | MR (Å$^3$) | $B_9$ | $B_{25}$ | compound | MR (Å$^3$) | $B_9$ | $B_{25}$ |
|---|---|---|---|---|---|---|---|
| 1 | 75.11 | 1.44 | 1.07 | 24 | 84.97 | 1.42 | 1.06 |
| 2 | 80.16 | 1.42 | 1.07 | 25 | 89.74 | 1.42 | 1.06 |
| 3 | 80.16 | 1.44 | 1.06 | 26 | 89.74 | 1.42 | 1.06 |
| 4 | 89.31 | 1.37 | 1.06 | 27 | 89.74 | 1.42 | 1.06 |
| 5 | 81.58 | 1.42 | 1.06 | 28 | 86.66 | 1.42 | 1.06 |
| 6 | 81.58 | 1.45 | 1.06 | 29 | 91.43 | 1.42 | 1.06 |
| 7 | 76.81 | 1.45 | 1.06 | 30 | 100.64 | 1.38 | 1.04 |
| 8 | 76.81 | 1.43 | 1.06 | 31 | 86.35 | 1.38 | 1.04 |
| 9 | 79.92 | 1.43 | 1.06 | 32 | 86.35 | 1.38 | 1.04 |
| 10 | 79.92 | 1.43 | 1.06 | 33 | 88.04 | 1.39 | 1.04 |
| 11 | 87.52 | 1.43 | 1.06 | 34 | 84.69 | 1.39 | 1.04 |
| 12 | 79.82 | 1.45 | 1.07 | 35 | 84.69 | 1.39 | 1.04 |
| 13 | 76.81 | 1.45 | 1.06 | 36 | 83.27 | 1.39 | 1.00 |
| 14 | 78.50 | 1.43 | 1.06 | 37 | 83.03 | 1.39 | 1.00 |
| 15 | 83.27 | 1.43 | 1.06 | 38 | 83.03 | 1.39 | 1.00 |
| 16 | 88.04 | 1.42 | 1.06 | 39 | 88.04 | 1.39 | 1.04 |
| 17 | 81.85 | 1.42 | 1.06 | 40 | 94.50 | 1.39 | 1.04 |
| 18 | 86.62 | 1.42 | 1.06 | 41 | 89.74 | 1.39 | 1.04 |
| 19 | 86.89 | 1.42 | 1.06 | 42 | 100.97 | 1.39 | 1.05 |
| 20 | 80.20 | 1.42 | 1.06 | 43 | 107.43 | 1.39 | 1.05 |
| 21 | 84.97 | 1.42 | 1.06 | 44 | 84.90 | 1.39 | 1.06 |
| 22 | 89.74 | 1.42 | 1.06 | 45 | 89.32 | 1.39 | 1.06 |
| 23 | 75.11 | 1.44 | 1.07 | | | | |

After several attempts to obtain a good classification of the compounds by using the 7 selected variables, the best separation was obtained with three of them (MR, $B_9$, and $B_{25}$—see Table 4). The first two principal components explained 91.33% of total variance in the data set as follows: $PC_1 = 62.99\%$ and $PC_2 = 28.34\%$. A number of

**Table 4.** Loading Values in the Two Principal Components

| variable | $PC_1$ | $PC_2$ |
|---|---|---|
| MR | −0.501 | 0.748 |
| $B_9$ | 0.674 | 0.022 |
| $B_{25}$ | 0.543 | 0.663 |

score plots were examined, and the most informative one is presented in Figure 3 (PC1 × PC2). From Figure 3 we can see that the 45 flavone compounds studied are separated into two groups: Group A (active compounds) and Group B (inactive compounds), and that only $PC_1$ is responsible for the separation between active and inactive compounds.

Figure 4 displays the plot of the loading vectors to the first two principal components ($PC_1$ and $PC_2$). Analyzing Figures 3 and 4 together, we can see that the training set is separated into two groups regarding the antipicornavirus activity when we used the variables MR, $B_9$, and $B_{25}$ in the PCA analysis: active (Group A — molecules 1, 2, 3, and 5−28 in Figure 1) and inactive (Group B — molecules 4 and 29−45 in Figure 1) compounds. From Figure 3 we can see that the active compounds present positive values for $PC_1$, while the inactive compounds present negative values for $PC_1$.

According to Table 4 and Figure 4, $PC_1$ can be expressed through the following equation

$$PC_1 = -0.501[MR] + 0.674[B_9] + 0.543[B_{25}] \quad (3)$$

From eq 3 we can see that for a compound to become active it needs to present large values for $B_9$ (bond order between $C_9$ and $C_{10}$ atoms) and $B_{25}$ (bond order between $C_{11}$ and $R_7$ atoms) and negative values for the molar refractivity (MR), since $PC_1$ presents positive value for active flavone compounds.

The MR descriptor can be related to the size and polarizability of a substituent and the larger the polar part of a molecule, the larger the MR value.[24] The role of MR can be ambivalent, i.e., MR can represent dispersive forces (which help the interaction between substituents and the biological receptor) and it can also represent a measure of volume (consequently, MR can measure the capacity of the substituent to distort the receptor's conformation, avoiding the interaction with the substratum). In the first case, it is expected to be a positive coefficient for MR. In the second case, a negative signal for the MR coefficient reflects stereochemical hindrance.[25,26]

For the active compounds studied in this work, the MR must present a negative coefficient in eq 3, and according to Montanari et al.[26] negative coefficients of MR represent a measurement of volume, i.e., a negative signal of MR indicates stereochemical hindrances of the substituents at the flavone compounds. In this way, we can say that some substituents in the active compounds can interact with the biological receptor through steric effects. Analyzing the MR values for the active compounds (see Table 3) we can see that these compounds present smaller MR values than inactive compounds. This indicates that the presence of small substituents at the active compounds avoid the interaction between the flavone compounds and the biological receptor.

Regarding the bond order descriptor, we can define it as half of the difference among electrons in bonding and antibonding molecular orbitals. The greater the bond order, the greater the dissociation energy and the smaller the bond length. Thus, for the active compounds studied, we can

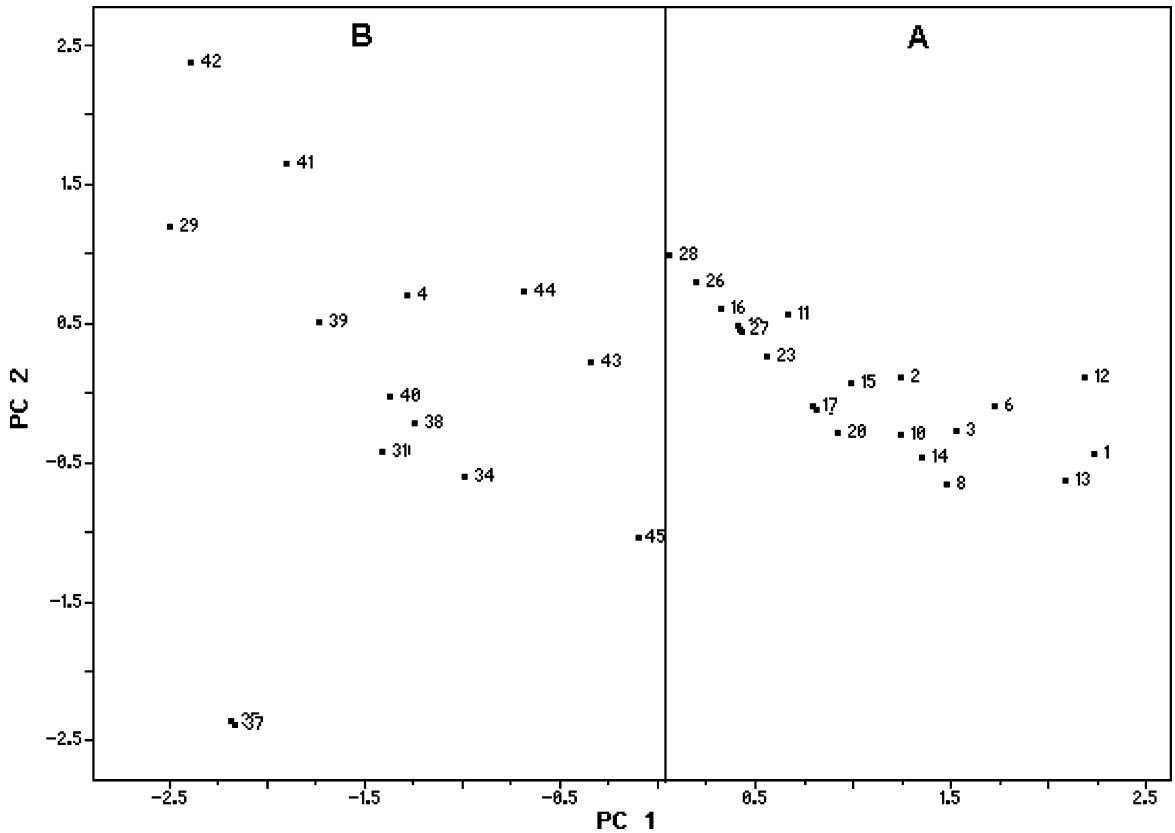



**Figure 3.** Plot of the first two PC score vectors ($PC_1$ and $PC_2$) for the separation of the training set into two groups: active compounds (Group A) and inactive compounds (Group B).

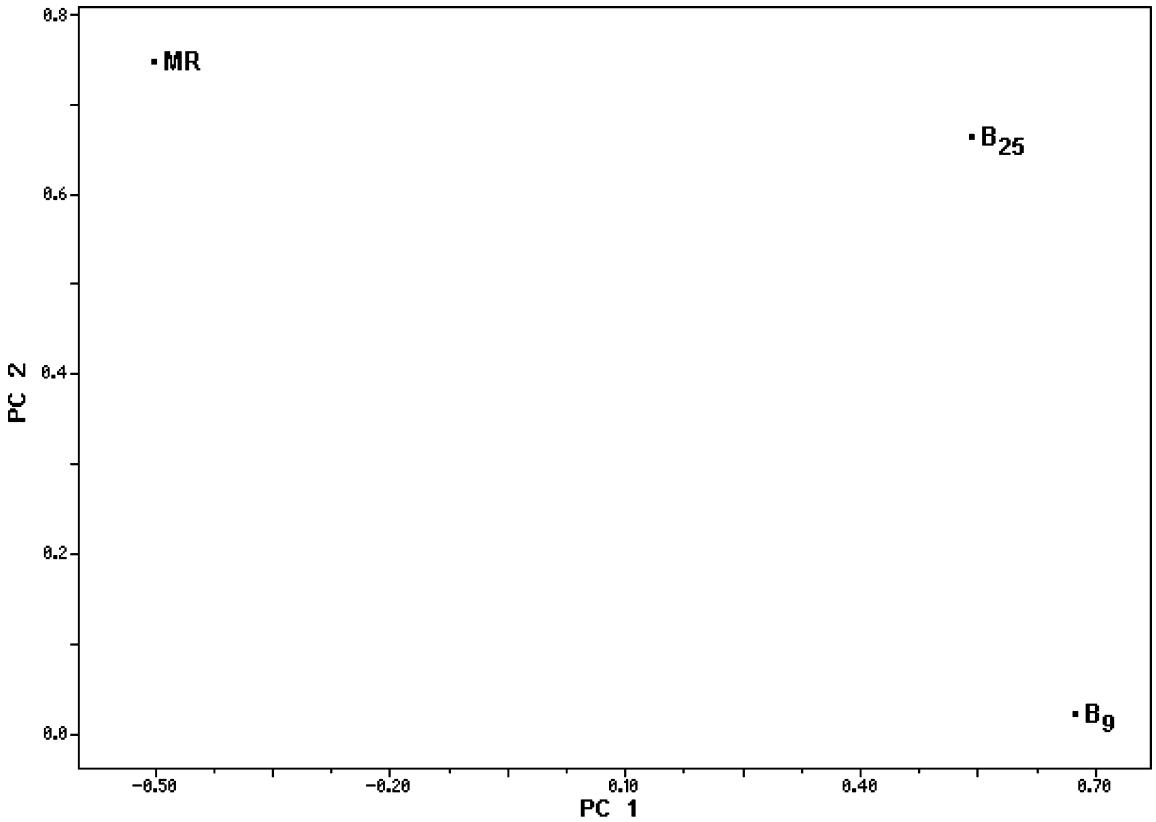



**Figure 4.** Plot of the first two PC loading vectors ($PC_1$ and $PC_2$) for the three variables responsible for the separation of the training set: MR (molar refractivity), $B_9$ (bond order between $C_9$ and $C_{10}$ atoms). and $B_{25}$ (bond order between $C_{11}$ and $R_7$ atoms).

conclude that some groups at the $C_{10}$ position are required so that the electronic density between $C_9$ and $C_{10}$ atoms presents a high value and, consequently, a high bond order ($B_9$).

Considering the bond order between $C_{11}$ and $R_7$ atoms ($B_{25}$) for the active compounds, we can see that some substituents are required at the $C_{11}$ position so that the bond order presents a high value, i.e., the electronic density between $C_{11}$ and $R_7$ atoms must present a high value. So, the greater the bond order between $C_9-C_{10}$ ($B_9$) and $C_{11}-R_7$ ($B_{25}$), the greater the possibility of interaction between the compounds studied and the biological receptor. From these descriptors ($B_9$ and $B_{25}$) we can conclude that the C ring region is very important in the understanding of the antipicornavirus activity of flavone compounds.

**Hierarchical Cluster Analysis (HCA).** The main purpose of the hierarchical cluster analysis (HCA) is to examine the distances between the samples in a data set, and this information is represented as a two-dimensional plot called a dendrogram. The HCA method is an excellent tool for preliminary data analysis, and it is useful to analyze data sets for expected or unexpected clusters, including the presence of outliers. It is advisable to analyze the dendrogram in conjunction with PCA, as they give similar information in different forms.

In HCA each point forms, initially, an only cluster and then the similarity matrix is analyzed. The most similar points are grouped forming one cluster, and the process is repeated until all the points belong to an only cluster.[27]

In the HCA we employed the same variables used in the PCA, i.e., MR, $B_9$, and $B_{25}$, and the results obtained from the HCA are displayed in the dendrogram shown in Figure 5. In fact, the dendrogram can be used to provide information on chemical behavior and verify the results obtained by PCA. In the dendrogram (Figure 5), the vertical lines represent the compounds, and the horizontal lines represent the similarity values between a pair of compounds, a compound and a group of compounds, and between groups of compounds.

The similarity value between the two classes of compounds was 0.136, and this means that these two classes are distinct. From Figure 5 we can see that HCA results are similar to those obtained with the PCA analysis, i.e., the compounds studied were grouped into two categories: active compounds (Group A — molecules 1, 2, 3, and 5−28 in Figure 1) and inactive compounds (Group B — molecules 4 and 29−45 in Figure 1).

Based on the results obtained with the PCA and HCA methodologies, the variables MR, $B_9$, and $B_{25}$ are those responsible for the separation between active and inactive flavone compounds with antipicornavirus activity.

**Stepwise Discriminant Analysis (SDA).** Discriminant analysis is a multivariate technique that has two principal goals: (1) separate objects from distinct populations and (2) allocate new objects to populations previously defined.[28,29] In this work we consider two groups: Group A, which contains the active compounds (molecules 1, 2, 3, and 5−28 in Figure 1), and Group B, which contains the inactive compounds (molecules 4 and 29−45 in Figure 1).

The stepwise discriminant analysis is a linear discriminant method based on the Fischer test (F-test) for the significance of the variables.[29] In each step one variable will be selected on the basis of its significance. After two steps, the more significant variables are extracted from the initial set of

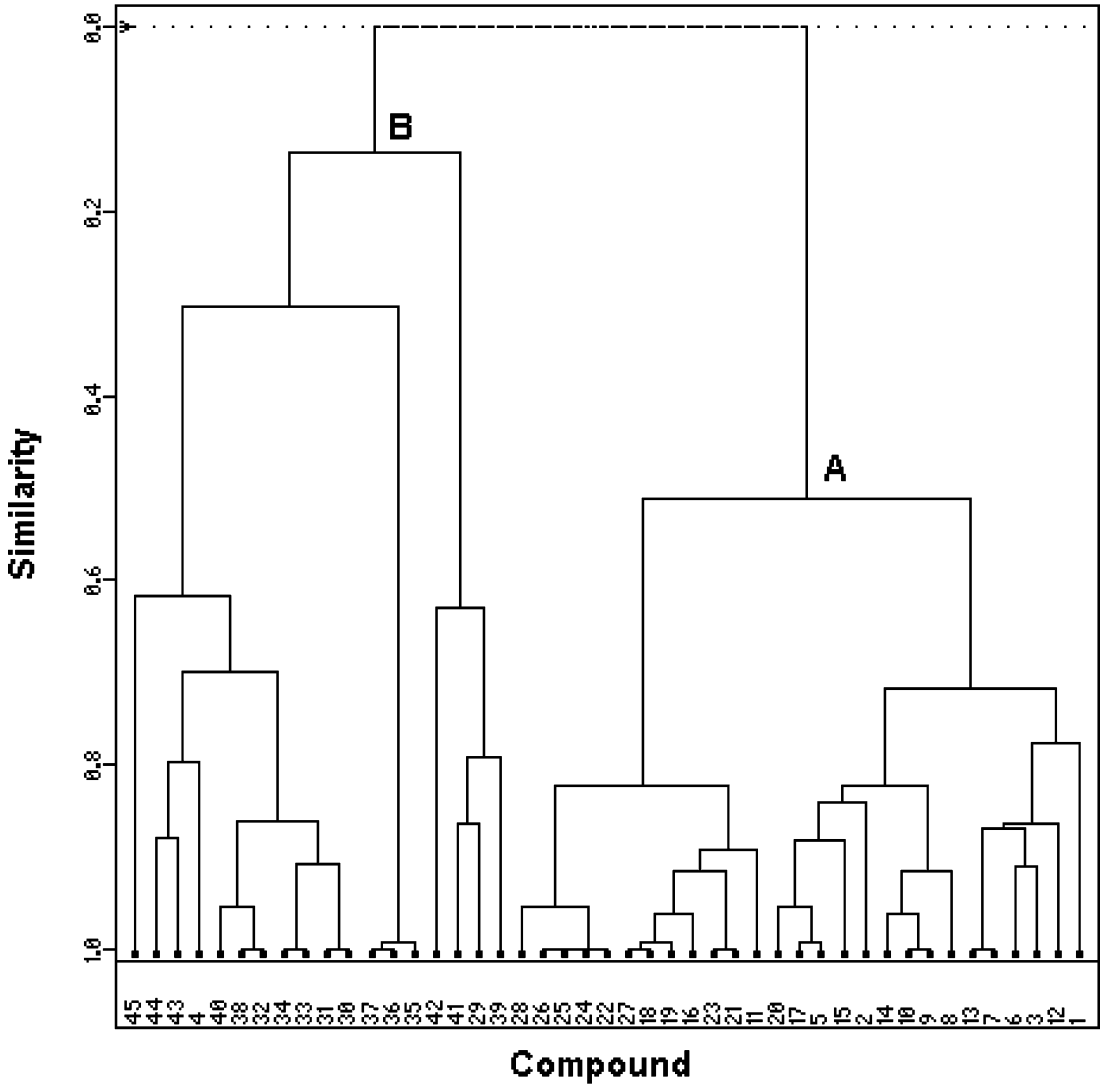



**Figure 5.** Dendrogram obtained for the compounds of the training set. The HCA classifies the compounds into two groups: active compounds (Group A) and inactive compounds (Group B).

**Table 5.** Classification Matrix Obtained with SDA

| | true group | |
|---|---|---|
| classified group | A | B |
| A | 27 | 0 |
| B | 0 | 18 |
| total | 27 | 18 |
| percentage | 100% | 100% |

variables under investigation: MR, $B_9$, and $B_{25}$. The two discriminant functions obtained with the SDA analysis are

$$\text{Group A: } -1.884 + 0.228\ \text{MR} + 4.449\ B_9 + 1.054\ B_{25}$$

$$\text{Group B: } -4.238 + 0.342\ \text{MR} + 6.673\ B_9 + 1.581\ B_{25}$$

The classification error rate was 0% (see Table 5), resulting in a satisfactory separation of the two groups. The allocation rule derived from the SDA results, when the antipicornavirus activity of a new flavone compound is investigated, is as follows: (a) initially one calculates, for the new flavone compound, the value of the more important variables obtained with the SDA methodology (MR, $B_9$, and $B_{25}$); (b) substitute these values in the two discriminant functions obtained in this work; (c) check which discriminant function (Group A − antipicornavirus active compounds or Group B − antipicornavirus inactive compounds) presents the higher value. The new flavone compound is active if the higher value is related to the discriminant function of Group A and vice versa.

Comparing the results obtained with the SDA, PCA, and HCA methods, we can notice that the variables MR, $B_9$, and $B_{25}$ are important in these three methodologies. Thus, combining the results obtained with SDA, PCA, and HCA we can say that MR, $B_9$, and $B_{25}$ are key variables to be taken into account in order to understand the antipicornavirus activity of the flavonoid compounds studied in this work.

To determine if our model is reliable we carried out a cross-validation test which uses the "leave-one-out" methodology. In this procedure, one compound is omitted from the data set, and the classification functions are built based on the remaining compounds. Afterward, 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 goes on until the last compound is removed. The results obtained with the cross-validation methodology are summarized in Table 6, and from these results we can see that the model

**Table 6.** Cross-Validation Matrix Obtained with SDA

| | true group | |
|---|---|---|
| classified group | A | B |
| A | 27 | 0 |
| B | 0 | 18 |
| total | 27 | 18 |
| percentage | 100% | 100% |

**Table 7.** Classification Obtained with the KNN Method

| | number of | compounds incorrectly classified | | |
|---|---|---|---|---|
| category | compounds | 1NN | 3NN | 5NN |
| active | 27 | 0 | 0 | 0 |
| inactive | 18 | 0 | 0 | 0 |
| total | 45 | 0 | 0 | 0 |
| percentage of correct information | | 100 | 100 | 100 |

obtained with PCA, HCA, and SDA is reliable once the cross-validation error is equal to 0%.

**K-Nearest Neighbor (KNN).** The KNN classifies a new object according to its distance to objects in the training set by using the distance matrix among the objects (compounds).[30] The K-nearest neighbors in the training set are found, and the object is assigned to the class of the majority of these K-nearest neighbors.

This method was used for the validation of the initial data set (45 compounds), and Table 7 presents the results obtained with 1, 3, and 5 nearest neighbors. For the case of 1, 3, and 5 nearest neighbors (1NN, 3NN, and 5NN, respectively), the percentage of correct information was 100%. We decided to use 5NN instead of 1NN and 3NN because the greater the number of nearest neighbors, the better the reliability of the KNN method. The result obtained with the KNN method is similar to those obtained with PCA, HCA, and SDA.

Here it is interesting to notice that all four chemometric methodologies we have employed in this work (PCA, HCA, SDA, and KNN) presented similar results indicating that our models, despite being built through theoretical tools only, have been assessed and validated through different methods.

Another important characteristic observed is that the three variables (MR, $B_9$, and $B_{25}$) responsible for the separation between active and inactive compounds are electronic variables, but MR is also related to the size of chemical groups in the molecule (i.e. MR may be also considered a steric variable). Therefore, we can conclude that electronic and steric effects could have an important role when one is trying to understand the activity of flavone compounds against picornaviruses, i.e., these three variables may have an important role in the understanding of the interaction between flavone compounds with antipicornavirus activity and the biological receptor. In fact, getting to know the behavior of these three variables may be useful when one is trying to obtain flavonoid compounds with high antipicornavirus activity.

Knowing the performance of the four pattern recognition methods on the 45 flavone compounds (training set) studied here, we decided to apply them to a series of new compounds whose biological activities are not yet known. The 9 flavone compounds proposed for our activity prediction study (labeled from A to I in Figure 2) were supplied by the organic chemistry group of the Federal University of Pará, Brazil. These 9 new compounds present a similar structure to the ones of our training set (the 45 flavone compounds shown in Figure 1) and biological tests were not performed with them yet, but future antipicornavirus tests with them could be used to validate our models.

**Table 8.** Calculated Values Obtained for the Properties (Variables) of the 9 New Flavone Compounds

| compound | MR (Å$^3$) | $B_9$ | $B_{25}$ |
|---|---|---|---|
| A | 73.42 | 1.42 | 0.95 |
| B | 107.68 | 1.37 | 1.08 |
| C | 91.03 | 1.42 | 1.04 |
| D | 91.44 | 1.34 | 1.04 |
| E | 99.29 | 1.39 | 1.04 |
| F | 112.45 | 1.28 | 1.04 |
| G | 115.09 | 1.36 | 1.00 |
| H | 77.12 | 1.42 | 1.07 |
| I | 92.82 | 1.42 | 1.05 |

**Table 9.** Activity Prediction Results Obtained with the Four Pattern Recognition Methods (PCA, HCA, SDA, and KNN) for the 9 New Flavone Compounds: Active (+) and Inactive (−)

| | activity | | | |
|---|---|---|---|---|
| compound | PCA | HCA | SDA | KNN |
| A | − | − | − | − |
| B | − | − | − | − |
| C | − | + | + | + |
| D | − | − | − | − |
| E | − | − | − | − |
| F | − | − | − | − |
| G | − | − | − | − |
| H | + | + | + | + |
| I | − | + | + | + |

The calculated values obtained with the variables MR, $B_9$, and $B_{25}$ for the new 9 flavone compounds are presented in Table 8. The results obtained with the four methods are summarized in Table 9. The compound H was classified as active by using all methodologies (PCA, HCA, SDA, and KNN), and the compounds C and I were classified as active with HCA, SDA, and KNN methodologies. The compounds A, B, D, E, F, and G were classified as inactive by all four methodologies. Thus, we can consider that compounds C, H, and I are predicted as potentially active against picornaviruses.

## CONCLUSIONS

Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), Stepwise Discriminant Analysis (SDA), and K-Nearest Neighbor (KNN) showed that 45 flavonoid compounds (flavones) studied in this work can be classified into two groups according to their degree of antipicornavirus activity (active and inactive compounds).

The variables MR, $B_9$, and $B_{25}$ were responsible for the separation between active (Group A) and inactive (Group B) compounds. Since MR, $B_9$, and $B_{25}$ are electronic variables and MR is also related to the size of chemical groups in the molecule, we can conclude that electronic and steric properties can be important in the understanding of the interaction between flavone compounds with antipicornavirus activity and the biological receptor. Therefore, the behavior of these three variables may be useful when one is trying to obtain flavonoid compounds with high antipicornavirus activity.

By using our PCA, HCA, SDA, and KNN models obtained with the 45 flavone compounds studied in this work, it was

possible to perform an activity prediction study with 9 new flavone compounds. Three of them were predicted to be potentially active against picornaviruses.

## ACKNOWLEDGMENT

The authors would like to thank CAPES and CNPq (Brazilian Agencies) for the financial support.

## REFERENCES AND NOTES

(1) Rotbart, H. A. Treatment of picornavirus infections. *Antivir. Res.* **2002**, *53*, 83−98.
(2) Tsang, S. K.; Cheh, J.; Isaacs, L.; Joseph-McCarthy, D.; Choi, S.; Pevear, S. C.; Whitesides, G. M.; Hogle, J. M. A structurally biased combinatorial approach for discovering new anti-picornaviral compounds. *Chem. Biol.* **2001**, *8*, 33−45.
(3) Santii, J.; Vainionpää, R.; Hyypiä, T. Molecular detection and typing of human picornaviruses. *Vir. Res.* **1999**, *62*, 177−183.
(4) Demeyer, N.; Haemers, A.; Mishra, L.; Pandey, H.; Pieters, L. A. C.; Berghe, D. A. V.; Vlietinck, A. J. 4′-hydroxy-3-methoxyflavones with potent antipicornavirus activity. *J. Med. Chem.* **1991**, *34*, 736−746.
(5) Alves, C. N.; Macedo, L. G.; Honório, K. M.; Camargo, A. J.; Santos, L. S.; Jardim, I. N.; Barata, L. E. S.; da Silva, A. B. F. A structure−activity (SAR) study of neolignan compounds with anti-schistosomiasis activity. *J. Braz. Chem. Soc.* **2002**, *13*, 300−307.
(6) Camargo, A. J.; Mercadante, R.; Honório, K. M.; Alves, C. N.; da Silva, A. B. F. A structure−activity (SAR) study of synthetic neolignans and related compounds with biological activity against *Escherichia coli*. *J. Mol. Struct. (THEOCHEM)* **2002**, *583*, 105−116.
(7) Alves, C. N.; Pinheiro, J. C.; Camargo, A. J.; Souza, A. J.; Carvalho, R. B.; da Silva, A. B. F. A quantum chemical and chemometric study of flavonoid compounds with anti-HIV activity. *J. Mol. Struct. (THEOCHEM)* **1999**, *491*, 123−131.
(8) Molfetta, F. A.; Alves, C. N.; da Silva, A. B. F. A quantum chemical and chemometric study of biflavonoid compounds with anti-HIV activity. *J. Mol. Struct. (THEOCHEM)* **2002**, *577*, 187−195.
(9) Dewar, M. J.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The development and use of quantum mechanical molecular models. 76. AM1 − a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(10) Wallet, J. C.; Gaydou, E. M.; Fadlane, A.; Baldy, A. Structure of 3-methoxy-2-phenyl-4H-1-benzopyran-4-one (3-methoxyflavone). *Acta Crystallogr.* **1988**, *C44*, 357−359.
(11) Wallet, J. C.; Gaydou, E. M.; Fadlane, A.; Baldy, A. 6-methoxy-2-(2,3,4-trimethoxyphenyl)-4H-1-benzopyran-4-one (6,2′,3′,4′-tetramethoxyflavone). *Acta Crystallogr.* **1990**, *C46*, 1131−1133.
(12) Souza, J., Jr.; Santos, R. H. A. 4′,5-dihydroxy-3′,5′-dimethoxy-6,7-(2",2"-dimethylpirano flavone); 3′acetoxi-4′,7-dimethoxy flavone; 4′′,5-diacetate-3′,5’,6,7-tetramethoxy flavone. *Z. Kristallogr.* Submitted for publication.
(13) Allinger, N. L.; Yuh, Y. H.; Lin, J. H. Molecular mechanics − the MM3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111*, 8551−8566.
(14) Ostlund, N. S. HyperChem 4.5: Program for molecular visualization and simulation, University of Waterloo, Canada, 1995.
(15) Dewar, M. J. S. AMPAC 6.5: Program for semiempirical calculations, University of Texas, U.S.A., 1994.
(16) Martin, Y. C. *Quantitative drug design: a Critical Introduction*; Marcel Dekker: New York, 1978.
(17) Little, J. MATLAB: Program for mathematical computing, MathWorks Inc., U.S.A., 2000
(18) Mulliken, R. S. A new electroaffinity scale; together with data on valence satates and on valence ionization potentials and electron affinities. *J. Chem. Phys.* **1934**, *2*, 782.
(19) Pearson, K. *Biometric series I*; Cambridge University Press: London, 1901.
(20) Hotelling, H. J. Analysis of a complex of chemometric variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 498−520.
(21) Massart, D. L.; Vandeginste, B. G. M.; Dening, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: A textbook*, Elsevier: Amsterdam, 1988.
(22) Bruns, R. E.; Faigle, J. F. G. Quimiometria. *Quim. Nova* **1985**, *84*, 84.
(23) Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*; John Wiley & Sons: New York, 1986.
(24) Kubinyi, H. *QSAR: Hansch analysis and related approaches*; VCH: New York, 1993.
(25) Dunn, W. J. Molar refractivity as an independent variable in quantitative structure−activity studies. *Eur. J. Med. Chem.* **1977**, *12*, 109−112.
(26) Montanari, M. L. C.; Montanari, C. A.; Gaudio, A. C. Validação lateral em relações quantitativas entre estrutura e atividade farmacológica, QSAR. *Quim. Nova* **2002**, *25*, 231−240.
(27) Kowalski, B. R.; Bender, C. F. Pattern recognition − powerful approach to interpreting chemical data. *J. Am. Chem. Soc.* **1972**, *9*, 5632.
(28) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Chemometric Analysis*; Prentice Hall: NJ, 1992.
(29) Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: New York, 1979.
(30) Dillon, W. R.; Goldstein, M. *Multivariate analysis*; John Wiley & Sons: New York, 1984.

CI030384N