Romanian Reports in Physics, Vol. 66, No. 2, P. 411–426, 2014

MULTIVARIATE STUDY OF FLAVONOIDS ACTIVE AGAINST CACO-2 COLON CARCINOMA

STELUTA GOSAV[1,2], MIHAIL LUCIAN BIRSA[1]

[1]"Al.I.Cuza" University, Chemistry Department, 11 Carol I Bldv. RO-700506, Iasi, Romania
[2]"Dunarea de Jos" University of Galati, Department of Chemistry, Physics and Environment, Domneasca St. 47, 800008 Galati, Romania, stelagosav@yahoo.com

Received December 4, 2012

Abstract. A multivariate study involving flavonoids with activity against human colon carcinoma Caco-2 is reported. The data set is composed of 26 flavonoids which can be separated, depending on their EC50 value *i.e.* the half maximal effective concentration, into active and less active compounds. Our purpose was to transform the chemical structure of each compound into a set of numbers, and to correlate them with the biological activity establishing a qualitative/quantitative relationship between calculated molecular descriptors and antiproliferative activity. The geometries of the studied flavonoids were fully optimized employing the Density Functional Theory (DFT) with the hybrid functional B3LYP in conjunction with the 6-31G(d) basis set. For each optimized structure we computed a set of parameters (1,356 descriptors) which characterizes the molecule from structural, topological, steric, electronic, hydrophobic, etc. points of view. Using the Fisher's weight and the correlation matrix, we reduced the number of descriptors from 1,356 to 15.

Aiming to investigate which molecular descriptors would be more efficient in classifying flavonoid compounds according to their degree of anticancer activity, we applied two unsupervised learning methods, Principal Component Analysis (PCA) and Cluster Analysis (CA), and a supervised learning method, Stepwise Discriminant Analysis (SDA). Using the 15 selected descriptors as input, the PCA and SDA techniques supplied us with the following parameters as common relevant descriptors: C-16, EN, Mor08e, HATS8m. The reliability of the structure-activity relationship model is verified using the cross-validation technique, the percentage of correct classifications being of 92.31 %.

Key words: Caco-2 cell; PCA; CA; SAR; SDA

1. INTRODUCTION

Flavonoids are an important class of polyphenolic compounds that are present in all vascular plants and play an important role in the human diet. They possess a remarkable spectrum of biochemical and pharmacological activities, suggesting that they significantly affect basic cell functions such as growth, differentiation and/or programmed cell death [1–5]. In recent years, many studies have provided evidence for the beneficial action of flavonoids on the human body. Flavonoids can influence carcinogen bioactivation, cell signaling, cell-cycle regulation, angiogenesis, oxidative stress and inflammation [6–11].

The rational search for new drugs is a very efficient strategy to obtain more specific and potent compounds without side effects. An interesting approach for the pharmaceutical industry is the using *in silico* techniques [12]. These techniques try to build relationships between a data set with known values of the property of interest and a set of calculated molecular descriptors. The calculated or theoretical descriptors are derived from a molecular structure representation of the compounds, so no experimental set-up is necessary. The advantage of these techniques is that they can be used in the very first stages of the drug development, even before the molecules are synthesized [13–22].

In this paper we present a SAR (Structure-Activity Relationship) model which uses molecular descriptors for the modelling of some flavonoids with different degrees of antiproliferative activity *i.e.*, intense and low activity. The data set is composed of 26 flavonoids with antiproliferative activity against human colon carcinoma Caco-2. Fifteen of them are more *active* (EC50 < 100 μM) representing the class A, and the rest of compounds are *less active* (100 μM < EC50 < 200 μM) representing the class LA. The studied compounds and their EC50 activity (the concentration that caused 50 % inhibition of cell proliferation), are selected from the literature [23]. The studied molecules were optimized employing the DFT with hybrid functional B3LYP in conjunction with the 6-31G(d) basis set. The optimized structures were then transformed into a set of parameters (1,356 descriptors) which characterizes the molecules from structural, topological, steric, electronic, hydrophobic, etc. points of view. Using the Fisher's weight and the correlation matrix we reduced the number of descriptors from 1,356 to 15. PCA, CA and SDA methods were employed to classify the molecules into the A and LA classes. The PCA and SDA methods used the 15 selected descriptors as input and supplied us with the most relevant descriptors for the separation of the compounds into more and less active flavonoids. Finally, a discussion on the relevant molecular descriptors is presented.

2. METHODOLOGY

2.1. COMPOUNDS

The database is formed of 26 flavonoids *i.e.*, 12 flavones, 6 flavanones, 4 flavonols, and 4 isoflavones (Table 1). The biological evaluation of the studied compounds in this work was determined by using the numerical indicator for activity EC50. All studied *flavonoid* compounds have anti-invasive properties in the case of colon cancer (Caco-2 cell) [23], fifteen of them having a growth-inhibitory activity (EC50 value) of less than 100 µM and the rest of the compounds having an EC50 between 100 and 200 µM. The first category of flavonoids forms the class of flavonoids with intense anti-invasive activity *i.e.* class A (*active* compounds) and the latter category of compounds is composed by the flavonoids with low anti-invasive activity *i.e.* class LA (*less active* compounds). The general chemical structures, the structural details and the class (A or LA) of these compounds are given in Table 1.

Table1

The molecular structure, the compound code and the class (A/LA) of each studied flavonoid

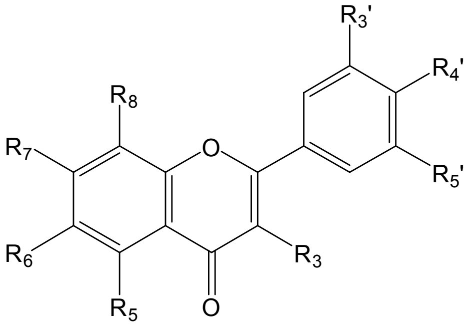
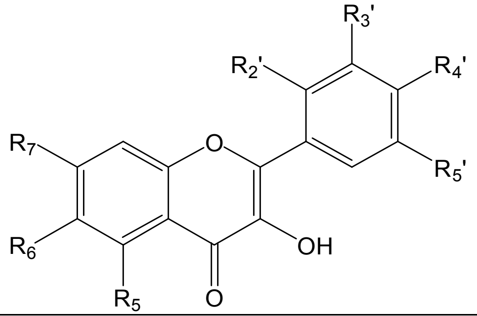
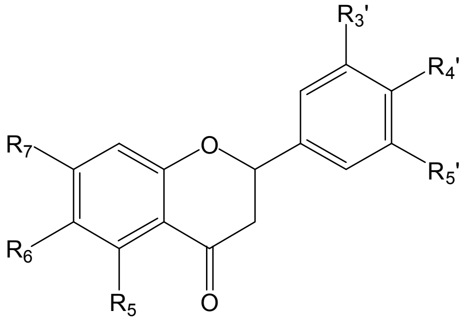
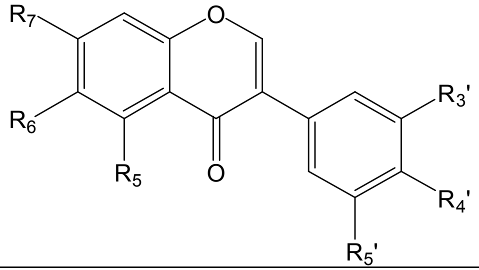
Subclass of flavonoids	Compound	Substituent	Class
Flavones 	C1	5-OH; 3,6,7,8,3’,4’,5’-H	LA
	C2	7-OH; 3,5,6,8,3’,4’,5’-H	LA
	C3	7,8-OH; 3,5,6,3’,4’,5’-H	LA
	C4	5,7-OH; 4’-OCH$_3$; 3,6,8,3’,5’-H	LA
	C5	5,7-OH; 3,6,8,3’,4’,5’-H	LA
	C6	7-sugar; 5,3’-OH; 4’-OCH$_3$; 3,6,8,5’-H	LA
	C7	7-apiosyl glucoside; 5,4’-OH; 3,6,8,3’,5’-H	LA
	C8	5,7,4’-OH; 3,6,8,3’,5’-H	LA
	C9	4’-OCH$_3$; 5,7,3’-OH; 3,6,8,5’-H	LA
	C10	4’-OCH$_3$; 3,5,7-OH; 6,8,3’,5’-H	A
	C11	5,6,7,8,4’-OCH$_3$; 3,3’, 5’-H	A
	C12	3- manno/gluco-pyranosyl; 5,7,3’,4’-OH; 6,8,5’-H	LA
Flavonols 	C13	5,6,7,2’,3’,4’,5’-H	A
	C14	7,3’,4’-OH; 5,6,2’5’-H	A
	C15	5,7,3’,4’-OH; 6,2’,5’-H	A
	C16	5,7,3’,4’,5’-OH; 6,2’-H	A

Table 1 (continued)

Flavanones 	C17	5,6,7,3’,4’,5’-H	A
	C18	6- methyl-butenyl; 7-OCH$_3$; 4’-OH; 5,3’,5’-H	A
	C19	7-rutinosid; 4’-OCH$_3$; 5-OH; 6,3’,5’-H	A
	C20	7- manno/gluco-pyranosyl; 4’-OCH$_3$; 5,3’-OH; 6,5’-H	A
	C21	4’-OCH$_3$; 5,7,3’-OH; 6,5’-H	A
	C22	7- manno/gluco-pyranosyl; 5,4’-OH; 6,3’,5’-H	LA
Isoflavones 	C23	4’-OCH$_3$; 5,7-OH; 6,3’,5’-H	A
	C24	7,4’-OH; 5,6,3’,5’-H	A
	C25	7-glucoside; 5,4’-OH; 6,3’,5’-H	A
	C26	5,7,4’-OH; 6,3’,5’-H	A

2.2. CALCULATION OF DESCRIPTORS

In the present work, the density functional three-parameter hybrid model (DFT/B3LYP) at the 6-31G(d) basis set level was adopted to calculate the properties of the studied molecules. All the calculations were performed using the Gaussian 03W program package [24]. After the optimisation process, only the chemical structures of compounds which have a hydroxyl group in position 3 *i.e.*, fisetin (C16), 3-OH-flavone (C1), kaempferide (C22), myricetin (C26) and quercetin (C28), get to be planar, their three rings being coplanar [25]. The DFT/B3LYP functional has been chosen because it is known from literature that this method leads to quite satisfactory results for the analysis of geometries and energies [26, 27].

The total number of theoretical descriptors computed for the optimized molecular structures is 1,356. The majority of molecular descriptors (1,349 parameters) were computed using Dragon 5.5 software [28] and the category which they belong to is presented in Table 2. The descriptors, like steric and quantum-electronic descriptors, were obtained using the Hyperchem 8.03 software [29] and from the files .log of each studied compound which are supplied to us by the Gaussian software at the end of the optimisation process. The steric and quantum-electronic descriptors are the following: the solvent accessible surface area

(SASA), the volume (Vol), the lowest unoccupied molecular orbital energy (E_{LUMO}), the highest occupied molecular orbital energy (E_{HOMO}), the dipol moment (DM), the Mulliken electronegativity (EN) and the difference between E_{HOMO} and E_{LUMO}, known as the gap energy (E_{gap}). The Mulliken electronegativity (EN) was calculated using the following equation: $EN = (I + EA)/2$, where $I = – E_{HOMO}$ and $EA = – E_{LUMO}$.

Table 2

Computed theoretical groups of descriptors and the number of descriptors corresponding to each group

No.	Type of descriptors	Number of descriptors
1	Constitutional descriptors	26
2	Topological descriptors	76
3	Walk and path counts	43
4	Connectivity indices	29
5	Information indices	47
6	2D autocorrelation indices	96
7	Edge adjacency indices	106
8	Burden eigenvalues	64
9	Topological charge indices	20
10	Eigenvalue-based indices	44
11	Randic molecular profiles	41
12	Geometrical descriptors	38
13	RDF descriptors	150
14	3D-MoRSE descriptors	160
15	WHIM descriptors	99
16	GETAWAY descriptors	191
17	Functional group counts	17
18	Atom-centred fragments	21
19	Charge descriptors	14
20	Molecular properties	26
21	2D binary fingerprints	12
22	2D frequency fingerprints	29
	Total	1,349

2.3. SELECTION OF RELEVANT VARIABLES

Before applying the statistical methods, all variables were auto-scaled so that they could be compared to each other on the same scale and have the same importance in the analysis. The Fisher's weight, $W_{A,\,LA}$, of each variable was then calculated, with the purpose of finding the parameters with the best discrimination power between the two classes, the active compounds class and the less active compounds class. The Fisher's weight for the X_i variable and for the compounds belonging to the *A* and *LA* classes was calculated using the following equation:

$$W_{A,LA}(i) = \frac{[\bar{X}_i(A) - \bar{X}_i(LA)]^2}{S_i^2(A) - S_i^2(LA)}, \tag{1}$$

where $\bar{X}_i(A)$ and $\bar{X}_i(LA)$ are the means of variable X_i for the flavonoids of the *A* and *LA* classes; $S_i(A)$ and $S_i(LA)$ are the standard deviations of variable X_i in the case of compounds of the *A* and *LA* classes.

After determining Fisher's weight for each variable, we retained only those with a value above 0.4, for a total of 28 descriptors. These descriptors present the best ability of discriminating between the more active and the less active flavonoids. The 28 selected descriptors belong to the following types: electronic descriptors (E_{HOMO} and EN), 2D autocorrelations (MATS7m, MATS7v, MATS7e, MATS7p, GATS5m, GATS7m, GATS5v, GATS7v, GATS5e, GATS7e, GATS5p, GATS7p), edge adjacency indices (EEig01d), topological charge indices (JGI4), 3D-MoRSE (Mor28m, Mor32v, Mor08e), GETAWAY (HATS8m, H_{0v}, HATSv, H_{0p}, R_{1v}), functional group counts (Ct, Cs) and atom-centred fragments (C-16, C-17).

In order to determine the contribution of each variable in the building of a model, it is necessary to have relatively small correlations among studied variables. For this reason, we have computed pair-wise correlation among all 28 parameters and removed those parameters which formed any linear dependence with $R > 0.8$. As a result of this procedure, we retained 15 descriptors for future analysis. The correlation matrix and the Fisher's weights for the selected descriptors are showed in Tables 3 and 4.

2.4. PRINCIPAL COMPONENT ANALYSIS

PCA [12] is a projection method which decomposes the original variables (molecular descriptors), into a "structure" part and a "noise" part. Thus, the original data matrix (*X*) of *m* rows (compounds) and *n* columns (variables) is decomposed as follows:

$$X = ML^T + R\,, \tag{2}$$

where *X* is a data matrix (*m*, *n*), *M* is a score matrix (*m*, *f*), L^T is a (transposed) loadings matrix (*f*, *n*), *R* is a matrix of residuals (*m*, *n*) and *f* is the number of significant principal components (PCs).

Table 3

The correlation matrix between the 15 selected descriptors

	E_{HOMO}	EN	MATS7m	GATS5m	GATS7m	Eeig01d	JGI4	Mor28m	Mor32v	Mor08e	HATS8m	H_{0v}	Cs	Ct	C-16
E_{HOMO}	1.00	-0.75	0.40	0.00	0.05	0.52	0.75	0.10	0.14	0.08	0.16	-0.50	0.07	0.22	-0.42
EN		1.00	-0.34	0.09	-0.09	-0.52	-0.48	-0.01	0.07	-0.47	-0.12	0.15	0.44	-0.50	0.43
MATS7m			1.00	0.58	-0.76	0.12	0.34	-0.23	0.25	-0.19	-0.29	-0.11	-0.38	0.60	-0.60
GATS5m				1.00	-0.77	-0.19	0.16	-0.63	0.56	-0.07	-0.62	-0.27	-0.05	0.47	-0.21
GATS7m					1.00	0.00	-0.01	0.33	-0.39	0.13	0.6	0.03	0.27	-0.39	0.45
EEig01d						1.00	0.42	0.18	-0.13	0.32	0.16	-0.26	-0.17	-0.07	-0.56
JGI4							1.00	-0.17	0.08	-0.01	0.08	-0.62	0.21	-0.04	-0.45
Mor28m								1.00	-0.57	-0.23	0.37	0.19	0.13	-0.24	0.15
Mor32v									1.00	0.08	-0.43	-0.32	0.08	0.17	-0.12
Mor08e										1.00	-0.27	-0.07	-0.25	0.15	0.06
HATS8m											1.00	0.40	0.10	-0.25	0.05
H_{0v}												1.00	-0.27	0.06	0.24
Cs													1.00	-0.55	0.58
Ct														1.00	-0.22
C-16															1.00

Table 4

The Fisher's weights for the 15 selected descriptors

Variable	$W_{A,LA}$	Variable	$W_{A,LA}$
E_{HOMO}	0.44	Mor28m	0.47
EN	**0.83**	Mor32v	0.43
MATS7m	**0.87**	Mor08e	0.44
GATS5m	0.54	HATS8m	0.46
GATS7m	0.54	H_{0v}	0.46
EEig01d	0.44	Cs	0.67
JGI4	0.51	Ct	0.47
		C-16	**1.63**

The score matrix represents the position of the samples in the new coordinate system named principal components (PCs) system. The PCs are a linear combination of the original variables and have the important property of being uncorrelated among each other. The loading matrix gives the importance of the original variables in the PCs. We applied this method with the aim to find the original variables which are the most important for the separation of studied flavonoids in more active and less active compounds.

2.5. CLUSTER ANALYSIS

CA is a method used for preliminary data analysis and it is very useful for examining data sets for expected or unexpected clusters. This technique displays data in such a way that natural clusters and patterns can be shown in a two-dimensional plot called dendrogram [12]. In fact, the dendrogram is used to provide information on chemical behaviour and verifies the results obtained by PCA technique. In order to obtain the dendrogram we have used as a joining rule the complete linkage and as input variables those six selected descriptors pointed out by the PCA method. The statistical analysis (PCA and CA) was performed using the Statistica 8.0 program [30].

2.6. STEPWISE DISCRIMINANT ANALYSIS

SDA is a statistical method where the main goal is to find the discriminant functions that can divide the classes of compounds as directly as possible. This method is useful for selecting variables with the highest relevance regarding the separation of the compounds into different groups and for assignment of new compounds to previously defined classes. The SDA technique [12] involves determining a linear equation that will predict which group the compound belongs to. The form of the equation or function is:

$$D = V_1 X_1 + V_2 X_2 + \cdots + V_i X_i + a , \tag{3}$$

where D is the discriminant function, V_i is the discriminant coefficient or the weight for predictor variable X_i, a is a constant and i is the number of predictor variables (independent variables). The number of discriminant functions is equal to the number of classes minus one. In order to evaluate the SAR model, the Wilks' Lambda, a standard statistic method, is used.

The Wilks' Lambda is used to denote the statistical significance of the discriminatory power of a model. Its value can vary from 1.0 representing no discriminatory power to 0.0 which signifies perfect discriminatory power. The Wilks' Lambda statistic for the overall discrimination is computed as the ratio of the determinant of the within-groups variance (det(W))/covariance matrix over the determinant of the total variance covariance matrix (det(T)):

$$\text{Wilk's Lambda} = \det(W)/\det(T). \tag{4}$$

We have applied this technique to all studied flavonoids and to 15 selected descriptors (Table 4) with the aim to find the most important descriptors in the discrimination of the compounds which belong to classes A and LA. Also, our purpose was to verify if the relevant descriptors obtained by this method are the same with those supplied by the PCA method. As in our case we used solely two classes of compounds consequently the SDA method computes one discrimination function. SDA was performed using the SPSS Statistic 20.0 program [31].

3. RESULTS AND DISCUSSION

3.1. RESULTS OF STATISTICAL METHODS

We have begun our PCA analysis by choosing the first 3 descriptors (Table 4) with the highest values of the Fisher's weights as input. Then, by adding other descriptors one by one, we have found that the best separation of the compounds was obtained with six of them: C-16, MATS7m, EN, Mor28m, Mor08e and HATS8m. The first three PCs cumulated a total variance of 80.94 %. By plotting of the first principal component (PC_1 = 34.69 %) *versus* the second principal component (PC_2 = 24.28 %) we can see that the all studied compounds are separated into two groups: class A (active compounds) and class LA (less active compounds) (Fig. 1). Also, we can observe that the first principal component is responsible for the discrimination between more and less active compounds, PC_1 having positive values for the compounds which belong to class A and negative values for the compounds which belong to class LA.

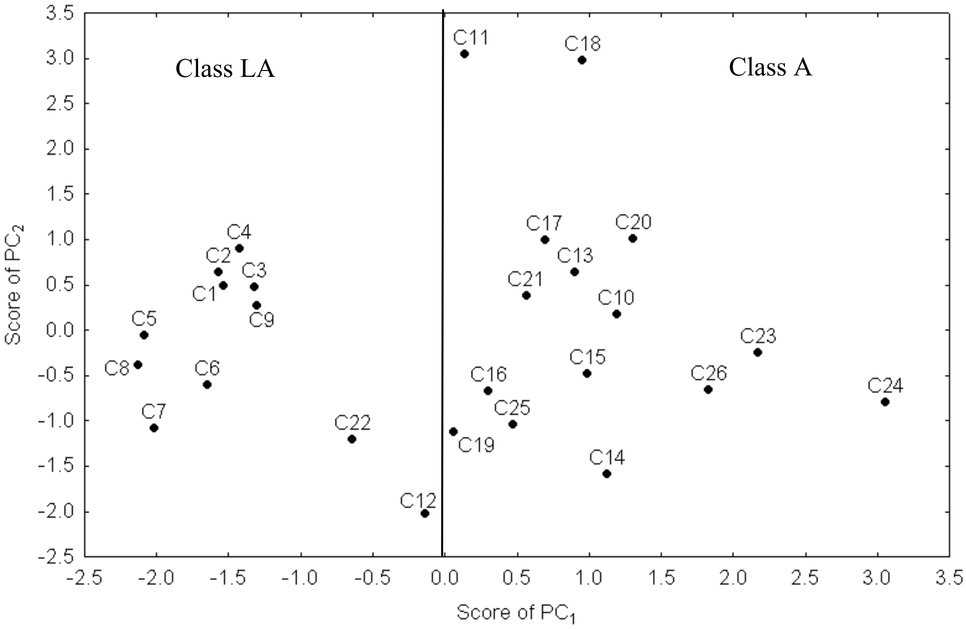


Fig. 1 – Plot of the PC_1 score in function of PC_2 score for studied compounds.

In order to analyse which variables bring the most important contribution to the components PC_1 and PC_2, their loading values are showed in Table 5. Analysing the loading values for PC1 (Table 5) we can see that the most significant descriptors are the following: MATS7m, C-16 and EN, but the contribution of the other three selected descriptors *i.e.* Mor28m, Mor08e and HATS8m, is not negligibile either. We should also notice that five of the six PCA selected descriptors cover all categories of descriptors from the dimensionality point of view, the 1D, 2D and 3D descriptors. Thus, the C-16 is a one-dimensional descriptor, MATS7m is 2D descriptor and Mor28m, Mor08e and HATS8m are 3D descriptors.

Table 5

Loading values of PC_1 and PC_2

Descriptor	Loading of PC_1	Loading of PC_2
EN	-0.634	0.088
MATS7m	0.795	-0.352
C-16	-0.768	0.442
Mor28m	-0.466	-0.508
Mor08e	0.280	0.730
HATS8m	-0.402	-0.581

The results of CA method, qualitative in nature and usually presented in a dendogram form, permit the visualization of clusters and correlations among samples. In our case, the dendrogram obtained for studied flavonoids using the six selected descriptors is presented in Fig. 2. The dendrogram shows a good discrimination between the A and LA classes. Only two compounds, C12 and C22, which are less active compounds, are misplaced in the A cluster but quite near the separation limit between the clusters.

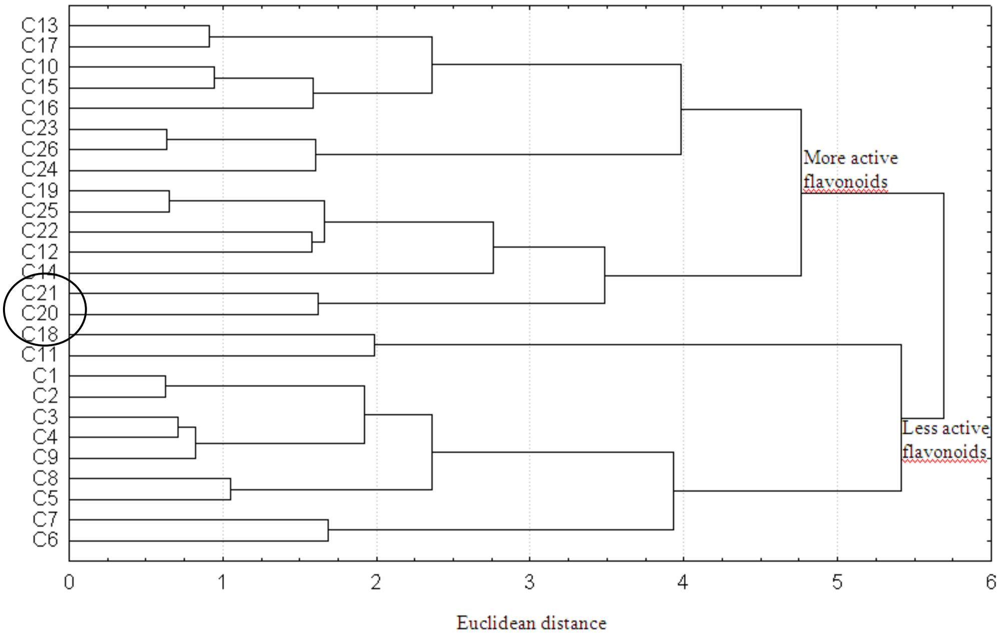


Fig. 2 – Dendrogram obtained for the studied compounds using the set of six selected descriptors.

SDA is a linear discriminant method based on the Fisher test (F-test) used to compute the significance (Sig.) of variables which participate in the discrimination process between classes (A/LA). A smaller significance value means a discriminant function which performs a better separation of compounds into the classes previously defined. After each step one variable is included in the SAR model on the basis of its significance. In our case, after six steps, the more significant variables *i.e.* C-16, EN, Mor08e, HATS8m, Mor32v and GATS5m, are extracted from the set of 15 selected variables. It is worth emphasizing the fact that after six steps, the Wilks' Lambda attains a good value (0.146) which means that the SAR model has only 14.6% unexplained variance (Table 6).

Table 6

The Wilks'Lambda, the canonical correlation, the eigenvalue, the degrees of freedom (df) and the significance of the SAR

Discriminant function	Wilks'Lambda	Canonical correlation	Eigenvalue	df	Sig.
1	0.146	0.924	5.865	6	0.00

The canonical correlation characterizes the overall SAR model from the point of view of the proportion of the explained variance retained by the model. In our case, the canonical correlation of 0.924 suggests that the model explains 85.4% of the variation in the grouping variables, *i.e.* whether a flavonoid belongs to class A or LA (Table 6).

The SDA technique supplies us the canonical discriminant function coefficients concerning the relevant predictor variables (C-16, EN, Mor08e, HATS8m, Mor32v and GATS5m). These non standardized coefficients are used to build the discriminant function, D. In our case we have obtained the following equation:

$$\begin{aligned} D = 1.475[\text{C-16}] + 0.682[EN] + 0.530[\text{HATS8m}] - \\ -0.747[\text{Mor08e}] - 0.585[\text{Mor32v}] - 0.413[\text{GATS5m}] \end{aligned}. \quad (5)$$

After the validation process, we observed that the discriminant score for all studied compounds has a negative value for the more active flavonoids and a positive value for the less active flavonoids. Analysing the selected descriptors, we noticed that the C-16 descriptor is the strongest predictor the fact standing out by its large coefficient (1.475). The other important descriptors (EN, Mor08e, HATS8m, Mor32v and GATS5m) as predictors have the values of the coefficients comprised between 0.4 and 0.8. We should mention that the sign of the coefficient indicates the direction of the relationship between the discriminant function, *D*, and the variable.

The results of the classification obtained using the discrimination function, *D*, are shown in Table 7. In order to verify the reliability of the SAR model, the cross-validation technique was applied. In the first case, all compounds were correctly classified, the percentage of correct classifications being 100 %, and in the second case, when using the cross-validation method, we obtained a good percentage of 92.31 % which means that only two less active compounds were incorrectly classified.

It is important to emphasize that four of the six selected descriptors by the SDA method – C-16, EN, Mor08e, HATS8m – were selected by the PCA method, as well. Concerning the other two descriptors, we can observe that the PCA method selected the Mor28m and MATS7m descriptors while the SDA method chose the Mor32v and GATS5m descriptors. It is worth to notice the fact that this pair of descriptors belongs to the same categories of descriptors: Mor28m and Mor32v are 3D-MoRSE descriptors, and MATS7m and GATS5m are 2D autocorrelation descriptors. In other words, the SDA method replaces the Mor28m descriptor selected by PCA with the Mor32v descriptor and the MATS7m descriptor with GATS5m, respectively. These replacements seem not to be very suitable because after the cross-validation process the SDA method gives two compounds incorrectly classified.

Table 7

The results of classification obtained with SDA method

Class		Classification		Classification with cross-validation	
		A	LA	A	LA
A		15	0	15	0
LA		0	11	2	9
Percentage of correct classifications	A	100	0	100	0
	LA	0	100	0	81.82
Total		100 %		92.31 %	

3.2. DISCUSSION ON RELEVANT DESCRIPTORS

In general, there is a relative difficulty in the interpretation of the molecular descriptors therefore we consulted the literature [32, 33] for information about each relevant descriptor.

A very important task in SAR modelling is to search within the biologically active molecules, the active substructures which give the most part of measured biological answer. All carbon atom centered fragment descriptors, from C-1 to C-44, are defined by Ghose-Crippen and describe each carbon atom through its bonding and neighboring atom types in the molecule. The neighbours of a carbon atom can by hydrogen atoms (H), carbon atoms (represented as R) and heteroatoms (X) in various combinations. The C-16 represents the number of chemical groups of the type: = CHR. In our case, the C-16 descriptor takes only two values, 1 or 0.

The majority of flavonoids from class A has the value 0 for this descriptor, meaning that this chemical group is absent from their molecular structures. The exception is given by two compounds: bavachinin (C18) and tangeretin (C11), which have a value of 1 for this descriptor. The absence of the = CHR chemical group for the majority of the more active flavonoids is confirmed by the fact that the compounds belonging to class A are part of different subclasses of flavonoids, like isoflavones, flavanones, flavonols and flavones, which either do not have the double bond between the C2 and C3 atoms, or if they have it, the substituent R_3 is not a hydrogen atom, but it is one of the following groups: OH, C_6H_4-OCH_3 or C_6H_4-OH. The reason why bavachinin has a value of 1 for the C-16 descriptor is due to the presence in position 6 of the methyl-butenyl group, while tangeretin (C11) is a flavone for which R_3=H. As the majority of compounds of class LA are flavones, it follows that the C-16 descriptor takes the value 1, the compounds having the double bond between the C2 and C3 atoms and the substituent R_3=H.

Generally, MATS descriptors (2D autocorrelation descriptors) are based on the Moran algorithm and they are calculated from lag 1 to lag 8 by using different atomic weights (atomic mass, van der Waals volume, Sanderson electronegativity and atomic polarizability). These descriptors are calculated from the molecular graph by summing the products of atomic weights of the terminal atoms for all the paths of the considered path length (the lag). In other words, a lag is a topological distance or all contributions of each different path length in the molecular graph. We mention that the higher the lag, the higher the distance between two atoms. The MATS7m descriptor is the Moran autocorrelation of the lag 7/weighted by atomic masses. In our case MATS7m descriptor takes negative values for all less active compounds and positive values for the majority of more active compounds.

An important descriptor in the discrimination of the compounds between the more and less active flavonoids is the Mulliken electronegativity, a quantum-electronic descriptor. This descriptor characterizes the charge transfer process which occurs between a compound and the biological receptor. We should notice that the mean value of the Mulliken electronegativity for the compounds belonging to class A (3.61 eV) is slightly smaller than that for the compounds belonging to class LA (3.79 eV).

The HATS8m descriptor is the leverage-weighted autocorrelation of lag 8 weighted by atomic masses, an H-GETAWAY descriptor. The Geometry, Topology and Atom-Weights Assembly (GETAWAY) descriptors [33] account for the influence of individual atoms on the shape of the molecule and are often combined with atomic properties such as atomic mass, polarizability, van der Waals volume, and electronegativity. This class of descriptors is based on a leverage matrix named molecular influence matrix (MIM), proposed as a molecular representation easily calculated from the spatial coordinates of the molecule atoms. For the studied flavonoids the values of the HATS8m descriptor are positive. In addition, we want to mention that the mean value of this descriptor in the case of more active flavonoids (0.092) is lower than that of less active flavonoids (0.12).

The Mor28m and Mor08e are 3D-MoRSE descriptors based on the idea of acquiring information from the 3D atomic coordinates by the transform used in electron diffraction studies. 3D-Molecule Representation of Structures based on Electron diffraction (3D-MoRSE) descriptors are calculated by summing atomic weights viewed by a different angular scattering function [32]. In this case, Mor 28m is the 3D-MoRSE – signal 28/weighted by atomic masses descriptor and Mor08e is the 3D-MoRSE – signal 08/weighted by atomic Sanderson electronegativity descriptor. The selection of these two descriptors by PCA method indicates the importance of atomic masses, steric property and of atomic Sanderson electronegativities, an electronic property for the discrimination of the studied flavonoids in more and less active compounds. In the case of the Mor08e descriptor, we generally obtained greater values for the more active flavonoids than for the less active flavonoids. This fact indicates the importance of the Sanderson electronegativity which may be related to the binding in a specific cellular structure. Analyzing the Mor28m descriptor we observed that for all studied compounds the values of Mor28m are positive, the majority of more active compounds (mean value of Mor28m = 0.23) having smaller values than less active compounds (mean value of Mor28m = 0.36). Allowing for the values obtained for the HATS8m and Mor28m descriptors and for the fact that a small molecule easily diffuses through cell membranes we may conclude that the smaller is the molecule, the higher is its antiproliferative activity.

4. CONCLUSIONS

The statistical methods (PCA, CA and SDA) used in the present paper proved that the discrimination between more and less active flavonoids using molecular descriptors is possible. The PCA and SDA techniques point to six important molecular descriptors each, four of which are common for both methods (C-16, EN, Mor08e, HATS8m). Analysing the relevant descriptors, we may conclude that the degree of antiproliferative activity (intense and low activity) is related to the structural (C-16), quantum-electronic (EN), steric (HATS8m) and combined steric/electronic (Mor08e) properties of studied flavonoids. These properties are related to the flavonoids interaction with a binding site and/or penetration of the cell membrane. Finally, we may add that this paper provides a deeper insight into important characteristics regarding the antiproliferative activity against human colon carcinoma Caco-2 of some flavonoids.

Acknowledgments. The authors are grateful for the financial support of POSDRU/89/1.5/S/49944 Project. Dr. Steluta Gosav would like to thank Maître de Conférence dr. Hélène Jamet (Department of Molecular Chemistry, University Joseph Fourier of Grenoble, France) for the computational facilities (Gaussian 03W software).

REFERENCES

1. C.A. Rice-Evans, L. Packer, *Flavonoids in Health and Disease*, Second Edition revised and expanded, New York, 2003.
2. L.F. Ibrahim, S.A. Kawashty, A. R. Baiuomy, M.M. Shabana, W.I. El-Eraky, S.I. El-Negoumy, *A Comparative study of the flavonoids and some biological activities of two 'Chenopodium' species*, Chem. Nat. Comp., **43**, 24–28 (2007).
3. D. Atmani, N. Chaher, D. Atmani, M. Berboucha, N. Debbache, H. Boudaoud, *Flavonoids in Human Health: From Structure to Biological Activity*, Curr. Nutr. & Food Sci., **5**, 225–237 (2009).
4. M. Imai, H. Kikuchi, T. Denda, K. Ohyama, C. Hirobe, 363 H. Toyoda, *Cytotoxic effects of flavonoids against a human colon cancer derived cell line, COLO 201: A potential natural anti-cancer substance*, Cancer Lett., **276**, 74–80 (2009).
5. L.M. Erhart, B. Lankat-Buttgereit, H. Schmidt, U. Wenzel, H. Daniel, R. Goke, *Flavone initiates a hierarchical activation of the caspase-cascade in colon cancer cells*, Apoptosis, **10**, 611–617 (2005).
6. K.A. Steinmetz, J.D. Potter, *Vegetables, fruit, and cancer. II. Mechanisms, review*, Cancer Caus. and Contr., **2**, 427–442 (1991).
7. M. Cardenas, M. Marder, V.C. Blank, L.P. Roguin, *Antitumor activity of some natural flavonoids and synthetic derivatives on various human and murine cancer cell lines*, Bioorg. & Med. Chem., **14**, 2966–2971 (2006).
8. K. Katayama, K. Masuyama, S. Yoshioka, H. Hasegawa, J. Mitsuhashi, Y. Sugimoto, *Flavonoids inhibit breast cancer resistance protein-mediated drug resistance: transporter specificity and structure-activity relationship*, Cancer Chemother. Pharm., **60**, 789–797 (2007).
9. A. Garcia-Lafuente, E. Guillamon, A. Villares, M.A. Rostagno, J.A. Martinez, *Flavonoids as anti-inflammatory agents: implications in cancer and cardiovascular disease*, Inflamm. Res., **58**, 537–552 (2009).
10. D.M. Brown, G.E. Kelly, A.J. Husband, *Flavonoid compounds in maintenance of prostate health and prevention and treatment of cancer*, Mol. Biotech., **30**, 253–270 (2005).
11. K.H. Shen, S.H. Hung, L.T. Yin, C.S. Huang, C.H. Chao, C.L. Liu, Y.W. Shih, *Acacetin, a flavonoid, inhibits the invasion and migration of human prostate cancer DU145 cells vi inactivation of the p38 MAPK signaling 386 pathway*, Mol. Cell Biochem., **333**, 279–291 (2010).
12. D. Livingstone, *A practical guide to scientific data analysis*, John Wiley, 2009.
13. C. Sarbu, D. Casoni, M. Darabantu, C. Maiereanu, *Quantitative structure-retention and retention-activity relationships of some 1,3-oxazolidine systems by RP-HPTLC and PCA*, J. Pharm. and Biomed. Analysis, **35**, 213–219 (2004).
14. A. Lauria, M. Ippolito, A.M. Almerico, *Principal component analysis on molecular descriptors as an alternative point of view in the search of new Hsp90 inhibitors*, Comput. Biol. and Chem., **33**, 386–390 (2009).
15. J. Olivero-Verbel, L. Pacheco-Londono, *Structure-activity aelationships for the anti-HIV activity of flavonoids*, J. Chem. Inf. Comput. Sci., **42**, 1241-1246 (2002).
16. A. Karawajczyk, V. Drgan, N. Medic, G. Oboh, S. Passamonti, M. Novic, *Properties of flavonoids influencing the binding to bilitranslocase investigated by neural network modeling*, Biochem. Pharm., **73**, 308–320 (2007).
17. A. Mantas, E. Deretey, F.H. Ferretti, M.R. Estrada, I. G. Csizmadia, *Structural analysis of flavonoids with anti-HIV activity*, J. Mol. Struct. (Theochem), **504**, 171–179 (2000).
18. K. Jaiswal, P.K. Naik, *Distinguishing compounds with anticancer activity by ANN using inductive QSAR descriptors*, Bioinform., **2**, 441–451 (2008).
19. D. Amić, D. Davidović-Amić, D. Beslo, V. Rastija, B. Lucić, N. Trinajstić, *SAR and QSAR of the antioxidant activity of flavonoids*, Curr. Med. Chem., **14**, 827–845 (2007).

20. K.C. Weber, K.M. Honorio, A.T. Bruni, A.B. Ferreira da Silva, *The use of classification methods for modeling the antioxidant activity of flavonoid compounds*, J. Mol. Model, **12**, 915–920 (2006).
21. J. Souza, Jr, R.H de Almeida Santos, M.M.C. Ferreira, 409 F.A. Molfetta, A.J. Camargo, K.M. Honorio, A.B. Ferreira da Silva, *A quantum chemical and statistical study of flavonoid compounds (flavones) with anti-HIV activity*, Eur. J. Med. Chem., **38**, 929–938 (2003).
22. F.A. Molfetta, K.M. Honorio, C.N. Alves, A.B. Ferreira da Silva, *A study on the anti-HIV activity of biflavonoid compounds by using quantum chemical and chemometric methods*, J. Mol. Struct. (Theochem), **674**, 191–197 (2004).
23. S. Kuntz, U. Wenzel, H. Daniel, *Comparative analysis of the effect of flavonoids on proliferation, cytotoxicity and apoptosis in human colon cancer cell lines*, Eur. J. Nutr., **38**, 133–142 (1999).
24. Gaussian 03W software, Gaussian Inc., USA, 2003.
25. Santiago Aparicio, *A Systematic Computational Study on Flavonoids*, Int. J. Mol. Sci., *11*, 2017–2038 (2010).
26. J. Lameira, I.G. Medeiros, M. Reis, A.S. Santos, C.N. Alves, *Structure–activity relationship study of flavone compounds with anti-HIV-1 integrase activity: A density functional theory study*, Bioorg. Med. Chem., **14**, 7105–7112 (2006).
27. F.A. Molfetta, A.T. Bruni, F.P. Rosseli, A.B.F. Silva, *A partial least squares and principal component regression study of quinone compounds with trypanocidal activity*, Struct. Chem., **18**, 49–57 (2007).
28. *** Dragon software, Evaluation Version 5.5. Talete Srl., Italy, 2007.
29. *** Hyperchem software, Version 8.0.3. Hyper Co., USA, 2007.
30. *** Statistica software, Version 8.0. StatSoft Inc., USA, 2008.
31. *** SPSS Statistic software, Version 20.0., IBM, USA, 2011.
32. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
33. E. Borges de Melo, J.P.A. Martins, T.C.M. Jorge, M.C. Friozi, M.M.C. Ferreira, *Multivariate QSAR study on the antimutagenic activity of flavonoids against 3-NFA on Salmonella typhimurium TA98*, Eur. J. Med. Chem., **45**, 4562–4569 (2010).