J Mol Model (2006) 12: 915–920
DOI 10.1007/s00894-005-0083-x

ORIGINAL PAPER

Karen C. Weber · Káthia M. Honório · Aline T. Bruni · Albérico B. F. da Silva

The use of classification methods for modeling the antioxidant activity of flavonoid compounds

Received: 28 January 2005 / Accepted: 21 November 2005 / Published online: 7 April 2006

Abstract A study using two classification methods (SDA and SIMCA) was carried out in this work with the aim of investigating the relationship between the structure of flavonoid compounds and their free-radical-scavenging ability. In this work, we report the use of chemometric methods (SDA and SIMCA) able to select the most relevant variables (steric, electronic, and topological) responsible for this ability. The results obtained with the SDA and SIMCA methods agree perfectly with our previous model, in which we used other chemometric methods (PCA, HCA and KNN) and are also corroborated with experimental results from the literature. This is a strong indication of how reliable the selection of variables is.

Keywords Flavonoid compounds · Antioxidant activity · Classification methods

Introduction

In the last decade, the role of free radicals in several diseases has resulted in intense research. Recent studies continue reporting that the action of free radicals in certain cellular targets can cause oxidative damage, leading to a series of illnesses, such as carcinogenesis [1], mutagenesis [2], cardiovascular disease [3] and premature cellular aging [4].

Flavonoid compounds are well known as potent antioxidant agents, since they have the ability to scavenge these damaging free radicals. Previous studies have reported some structural requirements necessary for the antioxidant activity of flavonoids. In general, the free-radical-scavenging ability is greater when the flavonoid compound possesses an *ortho*-hydroxylation in the positions 3′ and 4′ of ring B (see Fig. 1), a 2, 3-double bond in conjugation with a 4-oxo function and/or an OH group at position 3 of the ring C [5–9]. The importance of ring A was not considered until some studies reported that hydroxyl groups at positions 5 and 7 could influence the scavenging ability of flavonoids [10, 11]. However, the results of these studies present conflicting aspects and no further study has dealt with the interactions between different groups.

In a previous study [12], we used three chemometric methods (PCA—Principal Component Analysis, HCA—Hierarchical Cluster Analysis, and KNN—Kth-Nearest Neighbor) in order to find molecular descriptors among steric, electronic, lipophilic and topological ones, that could be related to the free radical scavenging activity presented by a set of flavonoid compounds. Reliable SAR (Structure-Activity Relationship) models were obtained employing this methodology, based on the discrimination of the more antioxidant and less antioxidant flavonoid compounds using four electronic properties [12]. Nevertheless, the unsupervised pattern recognition PCA and HCA are not appropriate for the prediction of unknown compounds, and KNN also has its limitations for this purpose [13]. Thus, in this work we applied two more robust classification methods (SDA—Stepwise Discriminant Analysis and SIMCA—Soft Independent Modeling by Class Analogy) with the aim of using the classification models in the activity prediction of new flavonoid compounds and validate our previous results [12].

K. C. Weber · A. B. F. da Silva (✉)
Instituto de Química de São Carlos, Universidade de São Paulo,
C.P. 780, 13566-590,
São Carlos, SP, Brazil
e-mail: alberico@iqsc.usp.br

K. M. Honório
Instituto de Física de São Carlos, Universidade de São Paulo,
C.P. 369, 13560-570,
São Carlos, SP, Brazil

A. T. Bruni
Departamento de Física, IBILCE, Instituto de Biociências,
Letras e Ciências Exatas, Universidade Estadual Paulista,
15054-000, São José do Rio Preto, SP, Brazil

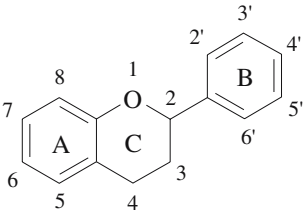

Fig. 1 General structure and numbering of flavonoid compound

Methodology

Table 1 shows the 22 compounds selected to compose the training set used in this work. The flavonoid compounds shown in Table 1 were chosen from the literature [7, 8, 10] and they have very similar structures, which belong to the classes of *flavones*, *flavonols* and *flavanones*. Flavonoids from the subclasses of *anthocyanidins*, which appear as positively charged molecules, and *isoflavones* (where the linkage between rings C and B is at the *meta* position of the former) were not taken into account in composing the training set. All compounds studied here were tested by using the TEAC (Trolox Equivalent Antioxidant Capacity) assay, expressed as the milimolar concentration of Trolox (TEAC=1) equivalent to the activity of a 1 mM solution of the compound being tested [8, 10]. This assay is based on the generation and detection of the radical $ABTS^{\cdot+}$ (2,2′-azinobis-(3-ethylbenzothiazoline)-6-sulfonic acid). The concentration of this radical is measured at 734 nm and the

Table 1 Chemical structure, numbering and antioxidant activity of the 22 flavonoid compounds studied

Subclass	Compound	Hydroxyl positions	Other groups	log TEAC
Flavones	1	3′, 4′	3-OCH_3	0.64[a]
	2	3′, 4′	3-$O(CH_2)_2OH$	0.64[a]
	3	7, 3′, 4′	–	0.63[a]
	4	3′, 4′	7-$O(CH_2)_2OH$	0.63[a]
	5	7, 3′, 4′	3-$O(CH_2)_2OH$	0.53[a]
	6	3′, 4′	7-OCH_3	0.52[a]
	7	5, 6, 7, 3′, 4′	–	0.37[b]
	8	5, 6, 7, 3′	–	0.33[b]
	9	5, 7, 3′, 4′	–	0.32[c]
	10	5, 7, 4′	–	0.16[c]
	11	5, 7	–	0.15[c]
	12	5, 6, 7	–	0.09[d]
Flavonols	13	3, 3′, 4′	7-$O(CH_2)_2OH$	0.69[a]
	14	3, 3′, 4′	–	0.65[a]
	15	3, 7, 3′, 4′	–	0.45[d]
	16	3, 5, 7, 2′, 4′	–	0.41[c]
	17	3, 5, 7	–	0.32[d]
	18	3, 5, 7, 4′	–	0.13[c]
Flavanones	19	3, 5, 7, 4′	–	0.25[e]
	20	5, 7, 4′	–	0.18[c]
	21	3, 5, 7, 3′, 4′	–	0.14[e]
	22	5, 7, 3′	4′-OCH_3	0.14[c]

[a] [10]
[b] [14]
[c] [8]
[d] [32]
[e] [7]

antioxidant-induced reduction of $ABTS^{\cdot+}$ is related to the scavenging capacity of the compound under investigation [8, 10].

Quantum chemical calculations were performed to obtain some steric and electronic properties of the compounds studied, such as bond distances, torsion angles, frontier orbital energies, atomic charges derived from electrostatic potential, bond orders, bond lengths, heat of formation, total and electronic molecular energies, dipole moment and molecular polarizability. Furthermore, several topological parameters representing different steric features were calculated to describe the molecules. A total set of 224 properties (20 electronic, four steric and 200 topological) was obtained and submitted to a statistical evaluation making use of the Fisher's weight [10, 14] in order to reduce the number of variables, and SDA and SIMCA methodologies with the aim of finding the most relevant properties for the free-radical-scavenging ability of the flavonoid compounds studied and build statistical models that could be able to classify them correctly as more antioxidant or less antioxidant compounds.

Initially, the modeling of the flavonoid structures was carried out by using the molecular mechanics method MM+ [15] in a pre-optimization and a conformational analysis was also performed by using the program ChemPlus [16]. Afterwards, the semiempirical method AM1 [17], implemented in the molecular package Ampac [18], was used for a final optimization of the structures and for the calculation of the molecular properties. As we had in mind to determine the qualitative characteristics of our data set, the AM1 method was chosen as it has proven to be appropriate for calculating most of the molecular properties of the flavonoid compounds studied in this work, as shown in previous studies [7, 19, 20]. The topological descriptors were calculated by using the software Dragon [21]. For the statistical analysis, the programs Pirouette [22] and Minitab [23] were used to perform the SIMCA and SDA analyses, respectively.

Results and discussion

Reduction of the number of variables

After the calculation of the relevant variables, we autoscaled each of them so that each variable had the same importance in the analyses. This kind of treatment ensures that the relative influence of different variables on the calculation is independent of their respective units. Afterwards, we calculated the Fisher's weight, W_{AB}, of each variable in order to select the ones presenting high discriminant power to distinguish the more antioxidant and less antioxidant flavonoid compounds. The Fisher's weight, W_{AB}, for the i variable and for samples belonging to the given classes A and B is calculated by Eq. 1:

$$W_{AB}(i) = \frac{\left[\overline{X}_i(A) - \overline{X}_i(B)\right]^2}{S_i^2(A) + S_i^2(B)} \quad (1)$$

After reducing the initial set of 224 calculated properties, we selected only 20 variables that showed significant weight values, i.e. the variables that presented the Fisher's weight above 1.00. The values of the Fisher's weight for these selected variables are shown in Table 2. These variables are considered as those that possess a higher ability in the discrimination (separation) between the more antioxidant and less antioxidant flavonoid compounds. In order to avoid the inclusion of highly correlated variables, we examined the scatter plots for these descriptors and excluded the correlated ones. After this procedure, we carried out the SDA analysis with this reduced set of variables until a discriminant function with any incorrectly classified sample was found.

SDA results

Stepwise Discriminant Analysis (SDA) is a statistical method where the main goal is to find *discriminant functions* (using the calculated variables) that can divide the groups of compounds as distinctly as possible. This method is useful for selecting variables with the highest relevance for the separation of the compounds into different groups (often referred as *discriminant power* of the variables), since it builds the discriminant functions using one variable at a time until the best discriminant function is obtained, based on the variables that give the best separation of the compounds into distinct groups. After the statistical validation of the model obtained by this procedure, the discriminant functions may be used to make predictions about unknown compounds [24].

Several SDA analyses were performed with different subsets of variables, and the best model was obtained using four molecular properties (with high discriminant power): α (molecular polarizability), *QC3* (charge at carbon 3), *QS5* (total charge at substitutent 5) and *QS3′* (total charge at substitutent 3′). Written as a linear combination of these variables, the discriminant functions obtained by SDA for each group of flavonoid compounds (more and less antioxidant) are:

Table 2 Variables presenting Fisher's weights higher than 1.00

Variable	Fisher's weight (W_{AB})	Variable	Fisher's weight (W_{AB})
α	1.08	L2u	1.37
QC3	1.02	L2e	1.28
QS5	93.84	HGM	1.01
QS3'	2.19	HATS4u	1.42
ATS4p	1.04	HATS5e	1.73
RDF085p	2.55	HATS3v	1.88
RDF090m	2.80	H4m	1.50
RDF090v	1.70	R4u	1.04
Mor25u	3.42	R2m+	3.33
Mor21m	1.79	R3v+	2.51

Group 1 (more antioxidant compounds):

$$f = -41.185 + 0.942(\alpha) + 68.396(QS5) - 0.072(QS3') + 5.850(QS3) \quad (2)$$

Group 2 (less antioxidant compounds):

$$f = -19.739 - 0.652(\alpha) - 47.351(QS5) + 0.050(QS3') - 4.050(QS3) \quad (3)$$

Table 3 shows the summary of classification obtained using these variables for the discriminant functions (Eqs. 2 and 3). The cross-validation technique was applied in order to verify the reliability of the model. Both models resulted in 100% of correct information, which means that no compound was incorrectly allocated in any group. Thus, for classification of unknown compounds, the values of these variables for the compounds must be substituted in the expression of the discriminant functions—Eqs. 2 and 3; the compound will belong to the group for which the discriminant function has the highest value.

It is important to notice that the variables selected by SDA in this work were the same ones as selected in our previous study [12]. The variable α is a measure of the electronic distortion in a molecule, caused by an external electric field. So, the greater the value of α, the greater the electronic distortion suffered by the molecule [25]. The importance of α to our model can be attributed to the fact that it is a good indicator of how the whole charge distribution of the flavonoid molecules studied is affected by the presence of different substitutents in certain positions. Additionally, as reported previously [26, 27], the free-radical-scavenging ability presented by some flavonoids depends on the presence of a free hydroxyl at position 3 (see Fig. 1), as it determines the ring B angle with respect to the rest of the molecule [9]. This feature is related to the stability of the flavonoid radical formed when the parent compound reacts with a damaging free radical, which is a strong indication of how good the free-radical scavenging exhibited by the compound is [8]. Flavonoid compounds possessing a 3-OH group are usually planar, which permits electronic delocalization and, consequently, a higher stability for the flavonoid radical [9].

According to a series of studies [6, 26, 28–30], the *o*-dihydroxyl group in ring B is the main requisite for effective free-radical-scavenging ability, since the reaction with the free radical takes place at ring B when this pattern of substitution is present and this arrangement confers high stability to the flavonoid radical. The most antioxidant compounds of our training set exhibit hydroxyls at positions 3′and 4′ and this configuration results in negative values of QS3′, while most of the less antioxidant compounds of our training set have positive values for this property. For the compounds without OH groups at ring B, an intramolecular rearrangement of the atoms at positions 4 and 5 may occur, yielding an *o*-dihydroxyl structure similar to that at 3′and 4′positions [11] and this explains the selection of the variable QS5 as an important descriptor by the SDA method.

Table 3 Summary of classification obtained with the SDA method

Group	Classification		Classification with cross-validation	
	A	**B**	**A**	**B**
A	9	0	9	0
B	0	13	0	13
Total	9	13	9	13
Percentage of correct information	100	100	100	100

SIMCA results

The chemometric method SIMCA (Soft Independent Modelling of Class Analogy) starts from the building of a PCA (Principal Component Analysis) model for each class according to the position and distribution of the compounds in the row space. In this way, SIMCA represents the variables as linear combinations called *Principal Components* (PCs) in order to represent the information contained in all variables. Afterwards, a geometric structure known as *hyperbox* is built around the samples of each class and the limits of the boxes are defined according to a certain level of confidence. When the values of the independent variables of a new sample is projected into the PC space of each class, the new sample is assigned to the class it best fits [24]. The main advantage of SIMCA over other classification methods is its ability to detect outlier samples [13].

In this work, we performed SIMCA using the autoscaled values of the variables selected as described for SDA. The best SIMCA model found was the one built with the same variables as in the SDA model. Fig. 2 shows the three-dimensional projection of the compounds, obtained with three PCs. Together, these PCs contain around 93% of the total variance of the original data, providing a reliable representation of the data set. The *hyperboxes* for the two classes of flavonoid compounds studied are represented in Fig. 2 by the points around of each class. From Fig. 2 we can see the division of the set of compounds into two well distinct classes, corresponding to **Class 1** (*more antioxidant flavonoid compounds*) and **Class 2** (*less antioxidant flavonoid compounds*). The coordinates of the *hyperboxes* that determine the limits of the classes are based on the standard deviations of the sample scores in the direction of each PC and states a confidence limit of 95% for the distribution of the classes (represented by dotted surfaces in Fig. 2). The rotation of Fig. 2 shows that no compound is allocated out of the confidence limits and that there is no superposition between the two classes.

Figure 3 displays the class distances calculated according to the residuals of the samples when they are adjusted to the classes. This plot is divided by two lines that represent the confidence limits (95%). The compounds lying in the **north–west quadrant** (NW) belong only to the

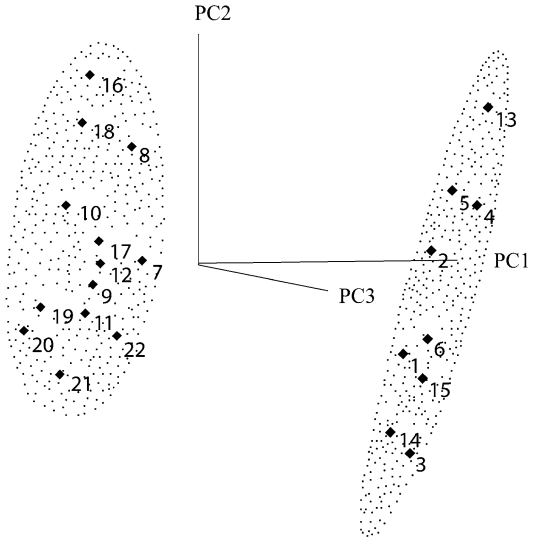

Fig. 2 Three-dimensional projection of the *hyperboxes* for Classes **1** (*right*) and **2** (*left*)

x-axis class, as they are at distances small enough to be considered members of this class. Analogously, the compounds in the **south-east quadrant** (SE) are members of the *y*-axis class only. Compounds positioned in the **south-west quadrant** (SW) may belong to both classes, while the ones in the **north-east** quadrant (NE) belong to none. From Fig. 3, it can be noted that the nine *more antioxidant* flavonoid compounds are in the NW quadrant, therefore they belong to Class 1. Otherwise, the 13 *less antioxidant* flavonoid compounds are in the SE quadrant, therefore belonging to Class 2.

Some important results can be obtained from the SIMCA method: (1) the distance between classes, which is a measure of how separated are the classes in a model; (2) the modeling power (*MP*) of the variables used in the classification model, indicating the influence of each variable in the model; and (3) the discriminant power (*DP*) of the variables, which is an indicative of the importance of each variable in the discrimination of the compounds into different classes [24]. For our SIMCA model, the calculated distance between Classes **1** and **2** is about 13. This allows us to consider the classes as very well separated. In chemometrics, a rule of thumb says that distances above 3 are considered suitable for a good distinction between classes [31]. Table 4 shows the values of MP and DP for the four variables used in our SIMCA model. The results indicate that the most important variable in the model is QS5, which agrees with the fact that the nine more antioxidant flavonoid compounds have positive values of QS5, while the 13 less antioxidant ones have negative values.

Table 4 Modeling and discriminant powers for the four selected variables according to the SIMCA method (α, QC3, QS5 and QS3′)

Variable	Modeling power (MP)	Discriminant power (DP)
α	0.23	2.19
QC3	0.47	67.68
QS5	0.52	734.64
QS3'	0.32	474.91

With these results, we can say that the SDA and SIMCA models presented in this work can help in the design of new antioxidant flavonoid compounds before their synthesis and tests against free radicals, since classification methods are very useful for class prediction of unknown compounds. Particularly, SIMCA is a more powerful method than others, as it has the ability of identifying outlier samples. Furthermore, the use of a multivariational meth-

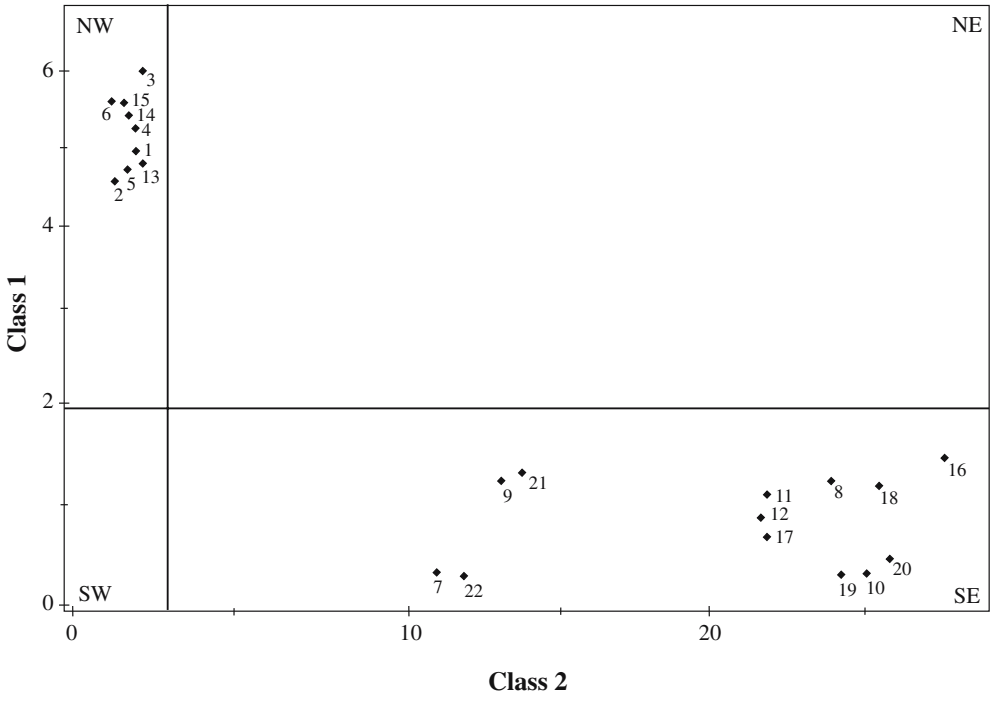

Fig. 3 Class distances obtained by SIMCA for the 22 flavonoid compounds studied

odology to build classification models allows us to consider the influence of all substitutent groups present in the structure of the flavonoid compounds in their reactivity with free radicals, which is better than taking into account only the number or the presence of hydroxyl groups at certain positions.

Conclusions

One of the main problems in most of previous studies on the relationship between free-radical-scavenging ability and the structure of flavonoid compounds lies on the fact that only the number of hydroxyl groups and/or the presence of these groups at certain positions were considered. In the present work, we report an attempt to treat the effect of intrinsic properties of flavonoid compounds caused by not only hydroxyl groups but also by other kinds of substitutents at different positions in the molecular structure of these compounds properly. This was achieved by the use of electronic properties and chemometric methods that were able to select the most relevant variables for the reaction between flavonoid compounds and free radicals.

The results obtained in this work with the SDA and SIMCA methods agree perfectly with our previous model, in which we used the chemometric methods PCA, HCA and KNN, and are coherent with experimental results from the literature, a strong indication that the selection of variables was suitable. The models presented in this work can be used to classify newly designed compounds before synthesis and tests against free radicals, in order to see if they will have the tendency of being good free-radical scavengers or not.

Acknowledgements The authors acknowledge the financial support given by CAPES and FAPESP (Brazilian agencies).

References

1. Takabe W, Niki E, Uchida K, Yamada S, Satoh K, Noguchi N (2001) Carcinogenesis 22:935–941
2. Kawanishi S, Hiraku Y, Oikawa S (2001) Mutat Res 488:65–76
3. Khan MA, Baseer A (2001) J Pak Med Assoc 50:261–264
4. Sastre J, Pallardo FV, Vina J (2000) IUBMB Life 49:427–435
5. van Acker SABE, van den Berg DJ, Tromp MNJL, Griffioen DH, van Bennekom WP, van der Vijgh WJF, Bast A (1996) Free Radical Biol Med 20:331–342
6. Cao G, Sofic E, Prior RL (1997) Free Radical Biol Med 22:749–760
7. Lien EJ, Ren S, Bui HH, Wang R (1999) Free Radical Biol Med 26:285–294
8. Rice-Evans CA, Miller NJ, Paganga G (1996) Free Radical Biol Med 20:933–956
9. van Acker SABE, de Groot MJ, van den Berg, DJ, Tromp MNJL, den Kelder GDO, van der Vijgh WJF, Bast A (1996) Chem Res Toxicol 9:1305–1312
10. van Acker FAA, Hageman JA, Haenen GRMM, van der Vijgh WJF, Bast A, Menge WMPB (2000) J Med Chem 43: 53752–3760
11. Heijnen CGM, Haenen GRMM, van Acker FAA, van der Vijgh WJF, Bast A (2001) Toxicol In Vitro 32:111–121
12. Weber KC, Honório KM, da Silva SL, Mercadante R, da Silva ABF (2005) Int J Quantum Chem 103:731–737
13. Beebe KR, Pell RJ, Seasholtz MB (1998) Chemometrics: a practical guide. Wiley: New York
14. Penga ZF, Strackb D, Baumertb A, Subramaniama R, Goha NK, Chiaa TF, Tana SN, Chiaa S (2003) Phytochem 62: 219–228
15. Allinger NL, Yuh YH, Lin JH (1989) J Am Chem Soc 111:8551–8566
16. Ostlund NS (1995) ChemPlus: program for molecular visualization and simulation. University of Waterloo, Canada
17. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902–3909
18. Ampac 6.5: program for semi-empirical calculations (1997) Semichem Inc, Shawnee
19. Zhang HY (1998) J Am Oil Chem Soc 75:1705–1709
20. Zhang HY (2000) Quant Struct Act Relat 19:50–53
21. Todeschini R, Consonni V, Pavan M (2002) Dragon 2.1. Milan
22. Pirouette 3.11 (2002) Infometrix Inc, Woodinville
23. Minitab Statistical Software (2000) Minitab Inc, State College
24. Sharaf MA, Illman DL, Kowalski BR (1986) Chemometrics. Wiley, New York
25. Kurtz HA, Stewart JJP, Dieter KM (1990) J Comput Chem 11:82
26. Arora A, Nair MG, Strasburg GM (1998) Free Radical Biol Med 24:1355–1363
27. Burda S, Oleszek W (2001) J Agric Food Chem 49:2774–2779
28. Kerry N, Rice-Evans C (1999) J Neurochem 73:247–253
29. Dugas Jr. AJ, Castaneda-Acosta J, Bonin G, Price KL, Fischer NH, Winston GW (2000) J Nat Products 63:327–331
30. Sekher Pannala A, Chan TS, O'Brien PJ, Rice-Evans CA (2001) Biochem Biophys Res Commun 282:1161–1168
31. Brereton RG (1992) Multivariate pattern recognition in chemometrics, illustrated by case studies. Elsevier, Amsterdam
32. Ishige K, Schubert D, Sagara Y (2001) Free Radical Biol Med 30:433–446