European Journal of Medicinal Chemistry 45 (2010) 4562–4569

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Multivariate QSAR study on the antimutagenic activity of flavonoids against 3-NFA on *Salmonella typhimurium* TA98

Eduardo Borges de Melo [a], João Paulo Ataide Martins [b], Teresa Cristina Marinho Jorge [a], Marcelo Couto Friozi [a], Márcia Miguel Castro Ferreira [b,*]

[a] Curso de Farmácia, Centro de Ciências Médicas e Farmacêuticas, Universidade Estadual do Oeste do Paraná – Unioeste, Cascavel, PR, Brazil
[b] Laboratory for Theoretical and Applied Chemometrics[1], Institute of Chemistry, University of Campinas - Unicamp, Campinas, SP 13084-971, P.O.B. 6154, Brazil

ARTICLE INFO

Article history:
Received 16 October 2009
Received in revised form
29 March 2010
Accepted 10 July 2010
Available online 17 July 2010

Keywords:
QSAR
Model validation
Flavonoids
PLS
OPS

ABSTRACT

A quantitative structure–activity relationship (QSAR) study of twenty flavonoid derivatives with antimutagenic activity against 3-nitrofluoranthene (3-NFA) was performed by Partial Least Squares (PLS), using Ordered Predictors Selection (OPS) algorithm for variable selection. Four descriptors (PJI2, Mor27m, G1e and R4u+) were selected and a good model ($n = 19$; $R^2 = 0.747$; SEC = 0.332; $PRESS_{cal} = 1.768$; $F_{(2,27)} = 23.585$; $Q^2_{LOO} = 0.590$; SEV = 0.388; $PRESS_{val} = 2.858$; $R^2_{pred} = 0.591$; SEP = 0.394; $ARE_{pred} = 5.230\%$; $k = 1.005$; $k' = 0.990$; $|R^2_0 - R'^2_0| = 0.109$) was built with two latent variables describing 83.410% of the original information. Leave-*N*-out cross validation (LNO) and **y**-randomization were performed in order to confirm the robustness of the model. The topological descriptors selected indicate that the antimutagenic activity against 3-NFA depends on molecular size, shape and Sanderson electronegativity of flavonoids. The proposed model may provide a better understanding of the antimutagenic activity of flavonoids and can be used as a guidance for proposition of new chemopreventive agents.

1. Introduction

The pharmacological and toxicological properties of nitroarenes have been the subject of several studies for many years. These compounds are generated when polycyclic aromatic hydrocarbons react with nitrogen oxides (NOx) under conditions that might be expected in polluted air or incomplete combustion of organic materials occurs. As a result, nitroaromatic compounds are present in large number of mixtures such as cigarette smoke, coal fly ash, diesel exhaust and grilled foods. In addition, nitroaromatic compounds are also found in the chemical industry, and some nitrofurans and nitroimidazoles are used as drugs. Therefore, human exposure to one or more nitroaromatic compounds could occur by a wide variety of routes [1,2].

2-Nitrofluorene (2-NF) is usually the dominant atmospheric nitroarene, followed by nitrofluoranthenes and nitropyrenes, as for example, 3-nitrofluoranthene (3-NFA) and 1-nitropyrene (1-NP) (Fig. 1). Many nitroarenes have been demonstrated to exert mutagenic activities in bacterial and mammalian test system. Thus, these and other nitroarenes may be involved in the etiology of some human cancers, namely lung and breast [1,2]. The carcinogenic activity of nitroaromatic compounds is usually initiated by an enzymatic nitroreduction. Considerable variation in the enzymes responsible for such nitroreduction has been observed in different organisms. In humans, xantine oxidase and microsomal NADPH-cytochrome c have been identified as the enzymes involved in this process. In *Salmonella typhimurium* TA98, the test strain used in the Ames test (a biological assay to assess the mutagenic potential of chemical compounds designed by the American biologist Bruce Ames), the nitroreduction is carried out by “classical” bacterial nitroreductase [1,3]. It has been proposed that mutagenicity of nitroaromatics involves a redox cycling that creates reactive species causing DNA lesions or formation of DNA adducts derived from the activated forms [1].

The carcinogenicity and mutagenicity of some chemicals may be modulated by other chemicals. It is well known that ingredients in dietary and other plants, fruits and seeds, or some synthetic

Abreviations: 1-NP, 1-nitropyrene; 2-NF, 2-nitrofluorene; 3-NFA, 3-nitrofluoranthene; AM1, Austin Model 1; ARE, average relative error; B3LYP, Becke, three-parameter, Lee-Yang-Parr; DFT, Density Functional Theory; DNA, deoxyribonucleic acid; HF, Hartree-Fock; NADPH, Nicotinamide adenine dinucleotide phosphate; OPS, ordered predictors selection; PLS, partial least squares; QSAR, quantitative structure-activity relationship.

* Corresponding author. Tel.: +55 19 3521 3102; fax: +55 19 3521 3023.
E-mail address: marcia@iqm.unicamp.br (M.M. Castro Ferreira).
[1] (http://lqta.iqm.unicamp.br)

0223-5234/$ – see front matter © 2010 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2010.07.017

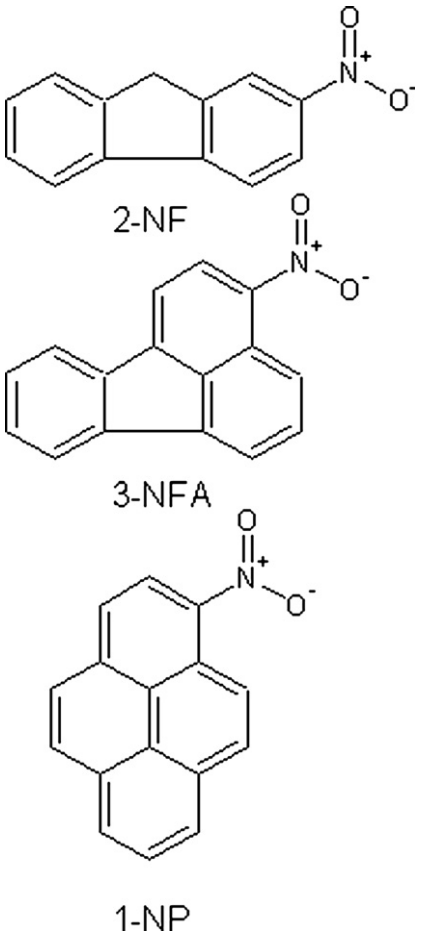


Fig. 1. Structures of nitroarenes 2-NF, 3-NFA and 1-NP.

derivatives, may exert anticarcinogenic and antimutagenic effects [2]. Epidemiological studies have indicated that the ingestion of certain amounts of antioxidants such as C and E vitamins and carotenoids, may retard or prevent cancer appearance [4]. This is the central idea of the chemoprevention therapeutic approach, defined as the use of natural or synthetic chemical agents to reverse, suppress or even prevent progression of invasive cancer [5]. The compounds with this property can act by different mechanisms [6], although in some cases the specific mechanism of antimutagenic effect of a compound (or compounds) is not well known.

Phenols and polyphenols are among the potent chemoprotective agents. Regarding these compounds, plant flavonoids are of outstanding importance. These non-toxic substances, found in several foods, have been demonstrated to posses protective properties, for instance, antioxidative, anticarcinogenic, antimutagenic, antiallergic, antiinflammatory and antiviral activities [2,7,8].

Considering the increasing interest in anticarcinogenicity and antimutagenicity of natural and synthetic phenolic compounds, especially flavonoids, a quantitative structure–activity relationship (QSAR) study was carried out in this work with the aim to obtain mathematical models that could aid in understanding and be used for prediction of the antimutagenic activity of flavonoids against the nitrofluoranthene 3-NFA.

2. Chemistry

The flavonoids of interest were selected from a study performed by Edenharder and Tang [2] on the antimutagenic activity of several compounds in relation to the mutagenicity induced by 2-NF, 3-NFA and 1-NP. In that study, 41 compounds are flavonoids, but only subsets of 12, 20 and 15 compounds presented antimutagenic activity quantitatively determined against, respectively, the three cited mutagenics.

For this study, the 20 flavonoids (the largest subset containing 10 flavones, 8 flavonols and 2 flavanones), listed in Table 1, that inhibited the mutagenic activity induced by 3-NFA were selected for this study. The other compounds (21 compounds) were described as inactive (ID_{50} value not provided) and they are not appropriate for a quantitative study. The histogram in Fig. 2 shows that the distribution of biological activities (pID_{50}) of the twenty compounds follows fairly well a normal distribution, indicating that the biological activities are well spread inside the considered range (pID_{50} 4.967–7.000). From the values of pID_{50} presented by these compounds, it can be seen that four of them present activity around 5.000, nine of them in the range of 5.100–6.100, six compounds have their activities between 6.100 and 6.990 and one compound has activity around 7.000.

3. Pharmacology

The selected training set was assayed as the antimutagenic effect on *S. typhimurium* TA98 by means of the Ames test. The biological activity, ID_{50}, (the dose of a compound in μmol/plate required to inhibit the activity of a given mutagen by 50%, as calculated from the corresponding dose–response curves) was quantitatively determined relative to 3-NFA [2]. The ID_{50} values were converted into $-\log ID_{50}$, or pID_{50}, and are listed in Table 1.

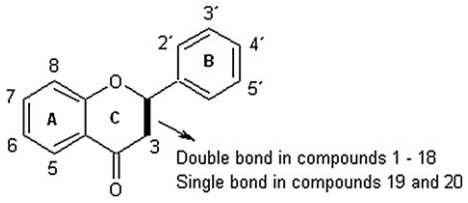
4. Results

Four descriptors (PJI2, R4u+, G1e and Mor27m) (Table 2) were selected out of 1221, applying a pre-selection followed by the OPS algorithm [9] and a refinement by Pirouette [10]. One outlier was detected (**14**) by analyzing the leverage *versus* studentized residual plot. Quercetin (**14**) is structurally similar to derivatives **6** (apigenin), **13** (kaempferol) and **18** (myricetin), meaning that these compounds have similar values for the selected descriptors, what can be seen in the dendrogram of Fig. 3. However, reasonable difference in the pID_{50} values are observed between quercetin and the other analogues (pID_{50} = 5.153 for **14**, and 7.000, 6.538 and 6.222 for **6**, **13** and **18**, respectively) what may be caused by an error in the experimental measurement. In the original paper, Edenharder and Tang [2] commented that quercetin (compound **14**) was the only one showing mutagenic activity in the absence of the mutagens (2-NF, 1-NP and 3-NFA) and its antimutagenic activity against the mutagens had to be corrected. This fact may lead to an error in the presented antimutagenic activity, what is an indication that compound **14** is really an outlier.

The training set used in this work presents a reasonable structural variability, showing substitutions in almost all carbon atoms forming rings A, B and C, including even sugar among them. However, its size is still small when the universe of existing flavonoidic compounds is considered, especially with the removal of compound **14**, detected as an outlier. Thus, a rigorous statistical validation process is necessary to assure the reliability of the model.

The best PLS model equation (1) was obtained with two latent variables describing 83.410% of original information (61.150% in the first latent variable and 22.260% in the second one). The descriptors in the model are capable to explain 74.670% and predict 59.050% of variance. The *F* value, obtained from the *F*-test, was higher than the corresponding critical-*F* ($p = 2$ and n-p-1 = 16) with 95% confidence interval ($\alpha = 0.05$), and the values of $PRESS_{val}$ were smaller than SS_Y, what confirms the statistical significance of the model.

Table 1
Selected training set from literature[2] and the observed antimutagenic effects (in pID_{50}) on mutagenicity induced by 3-NFA in *S. typhimurium* TA98.



Compound	Name	3	5	6	7	8	2′	3′	4′	5′	pID_{50}
1	5-Hydroxyflavone	H	OH	H	H	H	H	H	H	H	5.357
2	6-Hydroxyflavone	H	H	OH	H	H	H	H	H	H	6.699
3	7-Hydroxyflavone	H	H	H	OH	H	H	H	H	H	5.456
4	2′-Methoxyflavone	H	H	H	H	H	OCH_3	H	H	H	5.046
5	Chrysin	H	OH	H	OH	H	H	H	H	H	6.000
6	Apigenin	H	OH	H	OH	H	H	H	OH	H	7.000
7	Apigenin-7-glucoside	H	OH	H	O-Glc[a]	H	H	H	OH	H	5.620
8	Luteolin	H	OH	H	OH	H	H	OH	OH	H	6.523
9	Luteolin-7-glucoside	H	OH	H	O-Glc[a]	H	H	OH	OH	H	5.092
10	Tangeretin	H	OCH_3	OCH_3	OCH_3	OCH_3	H	H	OCH_3	H	4.967
11	Flavonol	OH	H	H	H	H	H	H	H	H	6.538
12	6-Methoxyflavonol	OH	H	OCH_3	H	H	H	H	H	H	5.620
13	Kaempferol	OH	OH	H	OH	H	H	H	OH	H	6.538
14	Quercetin	OH	OH	H	OH	H	H	OH	OH	H	5.143
15	Isorhamnetin	OH	OH	H	OH	H	H	OCH_3	OH	H	6.097
16	Rutin	O-Rut[b]	OH	H	OH	H	H	OH	OH	H	5.022
17	Morin	OH	OH	H	OH	H	OH	H	OH	H	6.155
18	Myricetin	OH	OH	H	OH	H	H	OH	OH	OH	6.222
19	Naringenin	H	OH	H	OH	H	H	H	OH	H	5.886
20	Hesperetin	H	OH	H	OH	H	H	OH	OCH_3	H	6.097

[a] O-Glc: O-glucose.
[b] O-Rut: O-rutinose.

$$pID_{50} = 1.039 + 17.516(PJI2) + 0.932(Mor27m) + 3.028(G1e) + 8.218(R4u+) \quad (1)$$

$R^2 = 0.747$; SEC = 0.332; $PRESS_{cal} = 1.768$; $F_{(2,16)} = 23.584$ ($cF = 3.634$); $Q^2_{LOO} = 0.590$; SEV = 0.388; $PRESS_{val} = 2.858$ ($SS_Y = 6.979$).

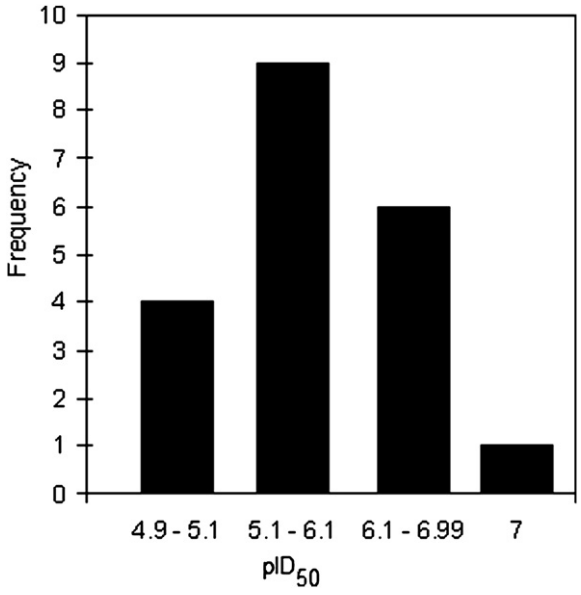


Fig. 2. Histogram presenting the distribution of the compounds in the pID_{50} range.

The results obtained from LNO validation and **y**-randomization analysis are shown in Fig. 4. The **y**-randomization test is useful to verify the possibility that the explained and predicted variances by the obtained model may suffer from chance correlation [11]. It can

Table 2
Values of descriptors used for the formulation of model and LOO cross-validation results.

Compound	PJI2[a]	Mor27m[b]	G1e[c]	R4u+[d]	pIC_{50} obs	pIC_{50} pred	Residues
1	0.800	−0.375	0.172	0.060	5.357	5.629	−0.272
2	1.000	−0.288	0.193	0.065	6.699	6.445	0.254
3	0.800	−0.339	0.172	0.067	5.456	5.812	−0.356
4	0.800	−0.431	0.168	0.061	5.046	5.589	−0.543
5	1.000	−0.384	0.171	0.060	6.000	6.190	−0.190
6	1.000	−0.382	0.169	0.077	7.000	6.370	0.630
7	0.875	−0.434	0.150	0.057	5.620	5.499	0.121
8	1.000	−0.348	0.168	0.075	6.523	6.422	0.101
9	0.875	−0.561	0.149	0.044	5.092	5.179	−0.087
10	0.857	−0.550	0.153	0.028	4.967	4.770	0.196
11	0.800	−0.332	0.172	0.079	6.538	5.753	0.785
12	0.833	−0.403	0.167	0.064	5.620	5.686	−0.066
13	1.000	−0.390	0.168	0.078	6.538	6.436	0.101
14[e]	1.000	−0.420	0.167	0.077	5.143	–	–
15	1.000	−0.419	0.163	0.064	6.097	6.142	−0.045
16	0.889	−0.575	0.139	0.040	5.022	5.041	−0.019
17	1.000	−0.435	0.167	0.089	6.155	6.768	−0.614
18	1.000	−0.386	0.165	0.075	6.222	6.397	−0.176
19	1.000	−0.393	0.167	0.067	5.886	6.287	−0.401
20	0.833	−0.563	0.162	0.066	6.097	5.327	0.770

[a] 2D Petitjean shape index.
[b] R maximal autocorrelation of lag 4/uniweighted.
[c] first symmetry directional component of the Weighted Holistic Invariant Molecular.
[d] 3D-MoRSE — signal 27/weighted by atomic masses.
[e] outlier.

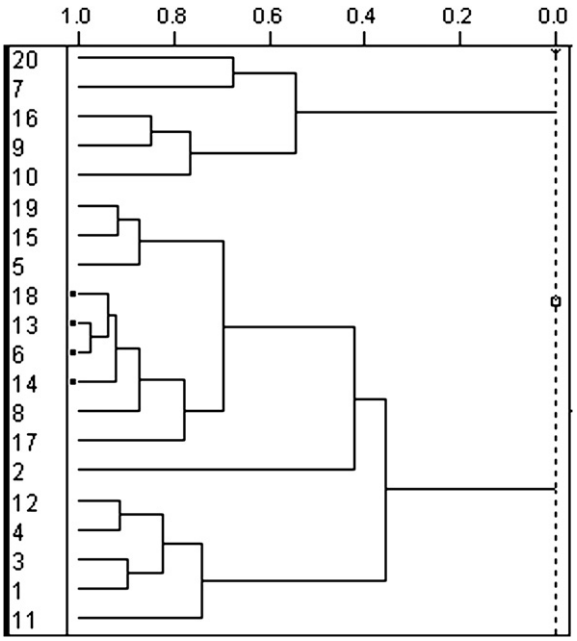


Fig. 3. Dendrogram (autoscaled data) of the training set, with compounds **6**, **14**, **13** and **18** highlighted.

be observed that the results obtained for all randomized models are of bad quality when compared to the real model, and the intercepts (Fig. 4A and B) are inside the acceptable values recommended in the literature, i.e., the intercepts are below the limits 0.3 and 0.05, respectively [12]. Dispersion of data points is observed in the regions around the intercepts, what is reasonable situation for smaller data sets. All obtained values for R^2 and Q^2 test are below 0.4 and 0.05 respectively (Fig. 4C). These results indicate that the variance explained by the model was not due to chance correlation.

LNO cross-validation employs smaller training sets than the LOO procedure and can be repeated several times due to the large number of combinations when leaving many compounds out from the training set once at a time. A QSAR model can be considered robust when its average Q^2_{LNO} values are relatively high and close to the value of Q^2_{LOO} [13]. The model obtained in this study has relatively high average Q^2_{LNO} (0.578), with small variations for each Q^2_{LNO} compared to Q^2_{LOO}. The standard deviation for each "*N*" values is small, with the maximum of 0.040 for L5O.

Another factor that can be evaluated in this model is the coincidence between the signals of *r* (Pearson correlation coefficient) for each descriptor with pID_{50} and the signals of coefficients in the model. According to Kiralj and Ferreira [14], the mismatch between the contributions of these two factors is an indication of lack of self-consistency of the model. As can be seen in Table 3, the model presents descriptors where the signs of their coefficients coincide with the information provided by correlation with biological activity, confirming the self-consistency of the model.

The data set was split into a training set formed by 14 compounds and a test set formed by the compounds **4**, **5**, **7**, **13** and **20**. This test set was used for external validation. These compounds cover well the entire range of pID_{50} values in complete set, as can be seen from the dendrogram in Fig. 5. The model built during the external validation has statistical parameters similar to those found for the model presented in equation (1) ($R^2 = 0.780$, SEC = 0.320, $PRESS_{cal} = 1.128$, $F_{(2,11)} = 19.467$, $Q^2_{LOO} = 0.612$, SEV = 0.376, and $PRESS_{val} = 1.984$). Therefore, they can be considered equivalent.

Many authors argue that only externally validated models (after complete internal validation) may be considered realistic and applicable for drug design [13,15]. Some studies from the literature support this conclusion [16,17]. The external validation results (Table 4) show that the model has high external prediction power,

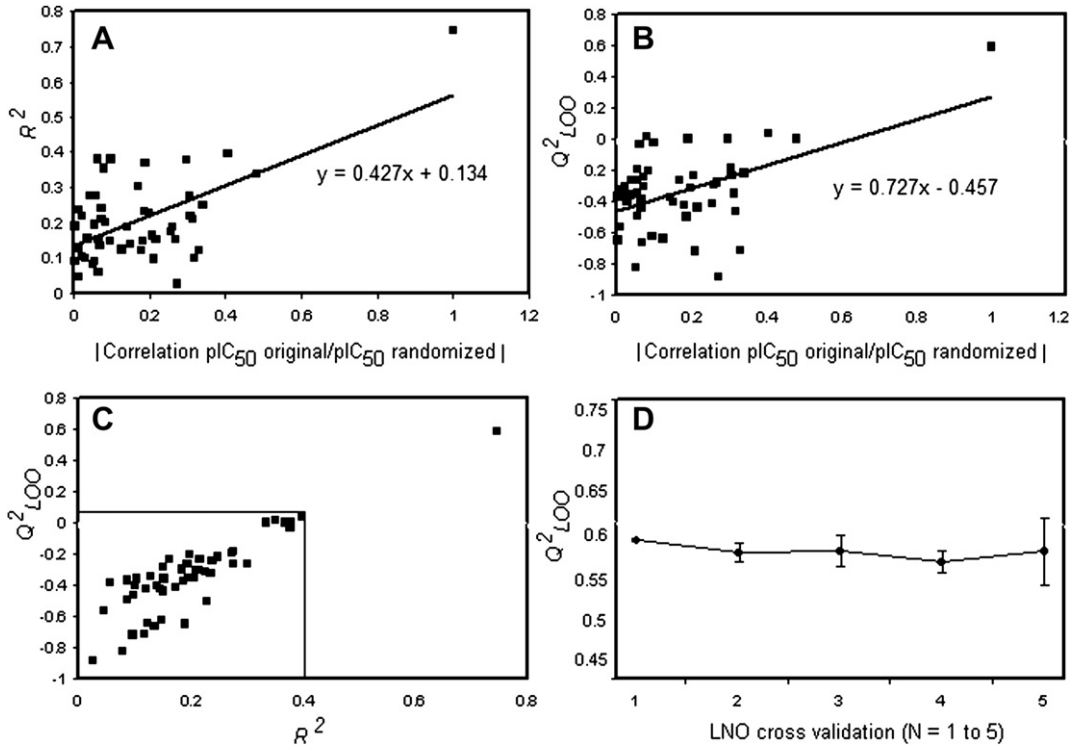


Fig. 4. Plots of **y**-randomization test (A, B e C) and LNO cross-validation (D). In the LNO plot (D), each point refers to the average value from a test in triplicate, and the bars refer to standard deviation.

Table 3
Individual Pearson correlation coefficients (final model, without the outlier) and standardized coefficients of the model.

Descriptor	r	Standardized coefficients
R4u+	0.773	0.411
Mor27m	0.606	0.125
PJI2	0.617	0.427
G1e	0.597	0.150

considering the proposed limits. Both values of k and k' and the relation $|R_0^2 - R'^2_0|$ are inside the acceptable ranges ($0.85 \leq k$ or $k' \leq 1.15$, and $|R_0^2 - R'^2_0| < 0.3$). The SEP and ARE_{pred} values are also considered low, what is an indicative of low prediction error (low deviation compared to the real value) for a synthesized derivative based on this model.

Using the obtained model, it is possible to support the hypothesis that compound **14** can be classified as an outlier. Its predicted pID_{50} value is 6.137, which is 1.006 logarithmic units above the experimental value from the work of Edenharder and Tang [2]. In addition, prediction values for compounds **6**, **13** and **18**, are close to that for compound **14**, what agrees with clustering tendency noticed in the dendrogram from Fig. 3.

Thus, the results of validation steps show that the model can be classified as a good model, since, according to the criteria used, it has good internal quality, it is robust, it does not suffer from chance correlation, it is self-consistent, and it shows a good capacity to external predictions.

5. Model discussion

All selected descriptors were obtained from Dragon 3.0 software [18]. The PJI2 is a topological descriptor based on graph theory, while the others (R4u+, G1e and Mor27m) are descriptors dependent of 3D optimized geometries (in this case, obtained in the molecular modeling step with B3LYP/6-31G theory level). The four descriptors influence positively the pID_{50}. Through the coefficients

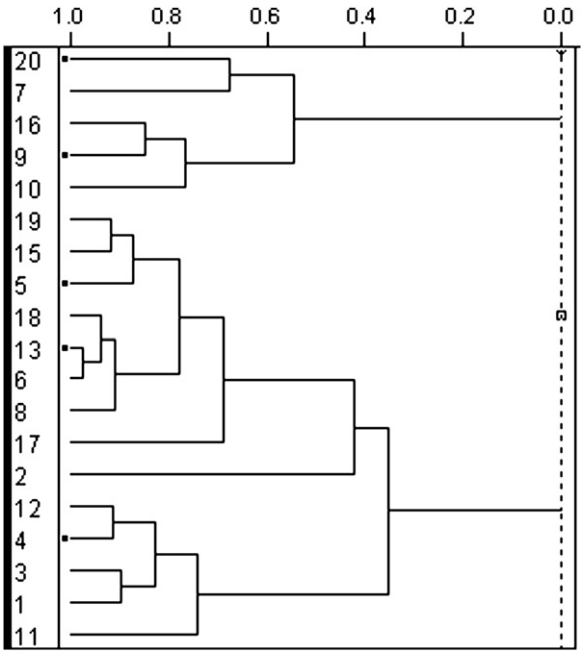


Fig. 5. Dendrogram (autoscaled data) of the training set (without the outlier **14**), with the compounds from the test set highlighted.

Table 4
Predicted values of the test set and results of statistical parameters.

Compound	$pIC_{50\ obs}$	$pIC_{50\ pred}$	Residues
4	5.046	5.517	−0.471
5	6.000	6.194	−0.194
9	5.092	5.091	0.001
13	6.538	6.411	0.127
20	6.097	5.388	0.708
R^2_{pred}		0.591	
SEP		0.394	
ARE_{pred}		5.230%	
k		1.005	
k'		0.990	
$\|R_0^2 - R'^2_0\|$		0.109	

(+0.411 for R4u+, +0.125 for Mor27m, +0.427 for PJI2 and +0.150 for G1e) obtained in the PLS model with autoscaled data, it is possible to see that two of them, R4u+ and PJI2, are the most significant for the model. It is interesting to observe that descriptors related to structural characteristics, usually accepted as important for the activity of flavonoids (e.g. number of OH groups in the molecules, Log P, or the dihedral angle between C and B ring) [19,20], were not selected, but some related characteristics can be encoded in the four selected descriptors.

It can be observed from the model obtained in this study that it has reasonable statistical quality, high prediction capacity and robustness in the desired limits. However, in a QSAR study, it is always desirable to obtain an interpretative model that is able to relate the physicochemical properties represented by the selected molecular descriptors to the mechanism of action of the system under study [15]. However, in this case the mechanism of action is not exactly known. According to Edenharder and Tang [2], antimutagenic flavonoids might modulate the mutagenic response of nitroarenes by: (i) modification of the permeability of bacterial membranes; (ii) physical, chemical or enzymatically catalyzed extracellular interactions between flavonoids and mutagens; (iii) interference with cellular mechanism leading to mutagenicity; or (iv) effects of flavonoids on DNA repair, fixation and expression on DNA damage caused by nitroarenes.

Thus, the information about the mechanism of action of this specific set is based only on the possible encoded information in the selected descriptors *per se* and other similar structure–activity studies on flavonoids [8,18–22]. An important point to be considered is the relative difficulty in the interpretation of the selected descriptors. In general, the literature refers to topological and geometric descriptors with information about shape, size and branching [21]. For a better understanding of the selected descriptors and a possible relation to the mechanism of antimutagenic activity, information about the definition of each selected descriptor was consulted in the literature and is presented in the following text.

The most important selected descriptor is the 2D Petitjean shape index (PJI2), also called graph-theoretical shape coefficient. This molecular shape descriptor describes the degree of deviation from a perfect cyclic topology [23]. The molecular shape descriptors are related to several physicochemical processes, such as transport phenomena as well as entropy contributions, and interaction capability between ligand and receptor [21]. The values of PJI2 vary in a range of 0 (not a circumference) to 1 (perfect circumference). In the training set, it is clear that the most actives compounds (**6**, **2**, **13**, **8**, **18** and **17**) possess PJI2 values equal to 1 and all of them are hydroxylated, not having methoxy or sugar groups. This fact indicates that most compact flavonoidic compounds tend to have a greater antimutagenic activity, maybe because they can bind in a small binding site or penetrate easier through the cell membrane

of *S. typhimurium* which protects the bacterial DNA. This descriptor was selected in another study with flavonoids, carried out by Rasulev and co-workers [24], that studied the structure–activity relationships relative to inhibition of lipids peroxidation, and also presented positive contribution for the activity.

The descriptor R4u+ is the R maximal autocorrelation of lag 4/uniweighted, an R-GETAWAY descriptor. Geometry, Topology and Atom-Weights Assembly (GETAWAY) descriptors are 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 in a chosen conformation. The magnitude of the maximum leverage for a molecule depends on its size and shape, and information about relations between two atoms in the same molecule can also be obtained. This class of descriptors tries to match 3D-molecular geometry, provided by the molecular influence matrix and atom relatedness via molecular topology, with chemical information by using different atomic weights (atomic mass, polarizability, van der Waals volume, electronegativity and unweighted). GETAWAY descriptors are divided into two sets: H-GETAWAY, derived by using only the information provided by the MIM, and R-GETAWAY, that combines this information with interatomic distances in the molecule obtained in a geometry matrix [18,25,26]. In the definition of R4u+, lag is the topological distance, or all contributions of each different path length in the molecular graph. Low terms, as R1 and R2, represent small molecules where they are expected to have low dependence on conformational changes as encoding information on pairs of atoms very close to each other. The higher the lag, the higher is the distance between two atoms [23]. Since R4u+ is unweighted by some chemical property, it probably encodes only geometrical information related to shape, and it is relatively dependent on conformational changes. Similar to PJI2, the tendency of higher values is also obtained for the most active compounds and smaller values for the less active ones. However, compound **14**, one of the less active compounds, has a high value for this descriptor as well, what strengthens the fact that it was identified as an outlier. Also similar to PJI2, the shape information dependent on 3D geometry can be related to a preferred conformation that should be adopted to bind at the binding site or the facility to penetrate through the bacteria membrane.

The descriptor G1e is the first symmetry directional component of the Weighted Holistic Invariant Molecular (WHIM) index weighted by atomic Sanderson electronegativities. The WHIM descriptors are 3D-descriptors based on the calculation of principal component axes calculated from a weighted covariance matrix (the same from GETAWAY plus atomic electrotopological state) obtained by the 3D atomic coordinates. This class of descriptors contains chemical information concerning molecular size, symmetry and shape, and distribution of the constituent atoms [24]. Thus, G1e indicates that the shape of molecules primarily determines the electronic distribution and may be related to the importance of electronegativity (behavior in redox process, electron release and withdraw, etc), in the antimutagenic activity. For instance, in the most active compound (**6**), the first two principal axes are parallel to the plane of the flavonoid skeleton (Fig. 6). In general, the literature describes that the flavonoidic compounds have a planar skeleton, which is an aromatic system with hyperconjugation and it is responsible for antioxidative properties of flavonoids. This π-conjugation system can bind free radicals and others species that damage DNA and other cellular structures [8,19,20]. However, these flavonoid characteristics can be related to the spatial shape descriptors, as PJI2 and R4u+. For instance, one of the less active compounds, **16**, is also a planar flavonoidic system, but the first principal axis is moved out from the system because the big sugar substituent in C3 (Fig. 6).

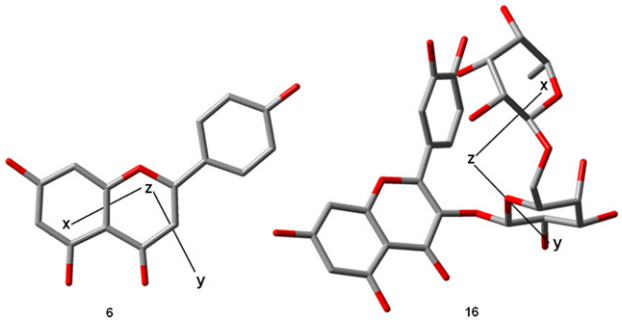


Fig. 6. Principal axes representation for compounds **6** and **16**. x = 1st axis, y = 2nd axis, z = 3rd axis.

Finally, the descriptor Mor27m is a 3D-molecule representation of structures based on electron diffraction (3D-MoRSE). These descriptors are based on the idea of acquiring information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves [24]. A generalized scattering function, called the molecular transform, can be calculated by using 3D atomic coordinates. The function takes into account the 3D arrangement of the atoms without ambiguities as those appearing when using chemical graphs [22]. In this case, Mor27m is the "3D-MoRSE — signal 27/weighted by atomic masses" descriptor, calculated by summing up the atomic weights viewed by angular scattering functions (27 $Å^{-1}$) and weighted by atomic masses. This fact indicates the importance of atomic mass, a steric property, and gives the basic idea that, the larger the molecule, the lower the activity, because the values of Mor27m also decrease when activity decreases. In fact, the less active compounds, **10** and **16**, have Mor27m values of −0.550 and −0.575, and the most active compounds, **6** and **2**, present values of −0.382 and −0.288. It seems clear that small compounds easily penetrate the bacteria through the cell membrane, thus contributing to the antimutagenic effect.

Based on the above discussion, the antimutagenic activity of this training set of flavonoids against 3-NFA is mainly dependent on the size and shape of the molecules. This hypothesis can be related to steric features (flexibility and size) important in the process of binding to the active site. Taking into account that large molecules are more difficult to diffuse through cell membranes, steric properties may also be related to the penetration into the bacteria. Electronic properties, maybe also related to the binding in a specific cellular structure or ability to capture reactive mutagens, are represented by the Sanderson electronegativity used for weighting the WHIM descriptor.

6. Conclusions

In this study, a multivariate QSAR model for a set of twenty flavonoid derivatives (10 flavones, 8 flavonols and 2 flavanones) with ability to inhibit mutagenicity caused by 3-NFA in *S. typhimurium* TA98, has been proposed. The model basic statistics, its internal and external prediction power, performance in LNO cross validation and **y**-randomization have shown that the model is statistically significant, robust and can be used for prediction purposes. The inhibitory activity of these compounds is described based on the topological descriptors PJI2, R4u+, Mor27m and G1e, indicating that the antimutagenic activity of the studied training set is dependent mainly on the molecular size and shape, what agrees with the literature on the activity of flavonoids. The descriptors are related to the flavonoid interaction with a binding site and/or penetration across the bacterial membrane. Therefore,

this study provides deeper insight on important characteristics regarding the antimutagenic activity of flavonoids (in this case, considering specifically the 3-NFA as mutagenic). Thus, it may be helpful for a better understanding of the activity of this class of compounds and useful as a guidance for the proposal of new chemopreventive agents.

7. Methodology

Three-dimensional structures were built based on similar crystallographic structures (codes DUMFAS, KEJBUW and WADRAV) retrieved from Cambridge Structural Database [27]. Necessary modifications of these structures and geometry optimization by molecular mechanics (MM+) and semi-empirical (AM1) quantum mechanical methods were carried out using HyperChem 7 [28]. Through the option potential, a conformational search at AM1 level was performed for all compounds, with increments of 10° at the dihedral angle between the rings B and C. This conformational search was carried out due to the steric hindrance present in compounds with substituent at the positions 2′ (**4**) and 3 (**11** to **18**), and between the flavonoid basic structure and the sugar lateral chain present in three compounds (**7**, **9** and **16**). The most stable geometries obtained by this process were further optimized at Hatree-Fock level (HF/6-31G) followed by Density Functional Theory level (B3LYP/6-31G), using Gaussian 03 [29]. The DFT/B3LYP functional have been chosen because it is known from literature that this method leads to quite satisfactory results for the analysis of geometries and energies [7,30]. The electronic descriptors were obtained after the final optimization. Other descriptors (steric, topological, solubility) were obtained from Parameter Client [31] and ALOGPS 2.1 [32] interfaces and DRAGON 3.0 Web Version [18], leading to a total of 1221 molecular descriptors.

In order to obtain a statistically reliable QSAR model, a three-step procedure was employed. In the first, the 1221 original descriptors were reduced to 840 by eliminating those that presented the absolute value of Pearson correlation coefficient ($|r|$) with pID_{50} lower than 0.3.

In the second step, the Ordered Predictors Selection (OPS) algorithm [9] was used for variable selection. This algorithm is capable of building partial least squares (PLS) [33] models on autoscaled descriptors (preprocessing recommended for this work) by rearranging the columns of the data matrix in such a way that the most important descriptors, classified according to an informative vector, are placed in the first columns. Then, successive PLS regressions are performed with increasing number of descriptors in order to find the best PLS model. In this work, the regression vector was used as the informative vector, and the cross validated prediction error (S_{PRESS}) obtained by the equation $(PRESS_{val})^{1/2}/n\text{-}p\text{-}1$ [34], where n is the number of samples and p is the number of PLS factors, was used as a criterion to classify the models generated by OPS.

At the third step, the set of nine descriptors selected by the OPS method (that presented a $S_{PRESS} = 0.493$) was further refined using the software Pirouette 4 [10], with removal of outliers and five more descriptors, to obtain an optimized model which would fulfill the criteria for being statistically significant, robust and interpretative.

PLS regression was chosen as the regression method because this method projects the original descriptors into a new set of variables, called latent variables, which are orthogonal to each other and then a regression is performed with this new set of variables [35]. Thus, unlike multiple linear regression (MLR), the number of original descriptors and the correlation among them are no longer a problem, since the regression is carried out on a small number of orthogonal latent variables.

Table 5
Statistics parameters analyzed and correspondent equations.

Parameter	Definition
Coefficient of multiple determination of calibration, R^2	$1-\frac{\sum_i (y_i-\hat{y}_{ci})^2}{\sum_i (y_i-\bar{y})^2}$
Standard deviation of calibration model, SEC	$\sqrt{\frac{\sum_i (y_i-\hat{y}_{ci})^2}{n-p-1}}$
Predictive Residual Sum of Squares of Calibration, $PRESS_{cal}$	$\sum_i (y_i-\hat{y}_{ci})^2$
F-test (with 95% confidence interval), F	$\frac{\sqrt{\frac{\sum_i (y_i-\hat{y}_{ci})^2}{k}}}{\sqrt{\frac{\sum_i (y_i-\bar{y})^2}{n-p-1}}}$
Coefficient of multiple determination of cross validation ("leave-N-out", LNO), Q^2_{LNO}	$1-\frac{\sum_i (y_i-\hat{y}_{vi})^2}{\sum_i (y_i-\bar{y})^2}$
Standard error of cross validation, SEV	$\sqrt{\frac{\sum_i (y_i-\hat{y}_{vi})^2}{n}}$
Predictive Residual Sum of Squares of Calibration of Validation, $PRESS_{val}$	$\sum_i (y_i-\hat{y}_{vi})^2$
Coefficient of multiple determination of prediction, R^2_{pred} [a]	$1-\frac{\sum_i (y_i-\hat{y}_{ei})^2}{\sum_i (y_i-\bar{y})^2}$
Average relative error of prediction, ARE_{pred}	$\frac{\sum_i \frac{\lvert y_i-\hat{y}_{ei}\rvert^2}{y_i}\cdot 100}{n}$
Standard error of prediction, SEP	$\sqrt{\frac{\sum_i (y_i-\hat{y}_{ei})^2}{n_{ev}}}$
Slopes of the linear regression lines, k and k'	$\frac{\sum_i (y_i-\hat{y}_{ei})}{\sum_i y_{ei}}$ and $\frac{\sum_i (y_i-\hat{y}_{ei})}{\sum_i y_i}$

y: observed biological activity; $\bar{y}$: average observed biological activity for the training set; $\hat{y}_{ci}$: estimated activity in the regression model; $\hat{y}_{vi}$: estimated activity in the cross-validation; $\hat{y}_{ei}$: estimated activity in the external validation; n: number of samples in the training set; n_{ev}: number of samples in the test set; p: number of latent variables in the model.

[a] For R^2_{pred}, $\bar{y}$ is the average value of observed activities for the training set without the test set.

The final model was thoroughly validated using a set of procedures suggested in the literature [14]. The statistical parameters listed in Table 5 were used to evaluate the quality of the model. For the internal quality, the recommended limits are $R^2 > 0.6$ and $Q^2_{LOO} > 0.5$ [28,36]. The SEC and SEV should be as lower as possible. The $PRESS_{val}$ values should be lower than the sum of squares of the response values (SS_Y) [37]. The F-test value should be higher than the tabled critical-F ($F_{p,n\text{-}p\text{-}1}$, where n is the number of compounds and p is the number of latent variables in the final model) and the higher the difference between them, the more statistically significant is the model [38].

The robustness of the optimized model was examined by leave-N-out cross validation (LNO, $N = 1, \cdots, 5$) procedure. This test was repeated three times for each value of "N", with a randomization of all rows from the data matrix and respective y values in each step of LNO process. The average value of each Q^2_{LNO} is expected to be close to Q^2_{LOO} (coefficient of multiple determination of leave-one-out

cross validation) with standard deviations close to zero [15]. The possibility of chance correlation was tested using **y**-randomization analysis [36], where the **y** vector was scrambled 50 times [37]. The approach suggested by Eriksson and co-workers [12], based on the absolute value of the Pearson correlation coefficient between the original vector **y** and the randomized vectors **y**, was used to quantify chance correlation. In this approach, two regression lines are built using these correlation coefficients (x-axis) and the R^2 and Q^2_{LOO} values (y-axis). The intercepts of the equations obtained in the linear regression should be less than 0.3 for R^2 and 0.05 for Q^2_{LOO}.

Once internally validated, the complete data set was split into training and test sets. The test set was selected using hierarchical cluster analysis in such a way that the entire range of pID_{50} and the structural variations were well represented. The parameter R^2_{pred} was used as a measure of the predictive power of a QSAR model. For this work, it was used the recommended limit of $R^2_{pred} > 0.5$ [39,40]. However, this is not a sufficient condition to guarantee that the model is really predictive. It is also recommended to check: (i) the slopes k or k' of the linear regression lines between the observed activity (y_i) and the predicted activity in the external validation ($\hat{y}_{ei}$) (Table 4), where the slopes should be $0.85 \leq k$ or $k' \leq 1.15$; and (ii) the absolute value of the difference between the coefficients of multiple determination, R^2_0 and R'^2_0, smaller than 0.3 [11,13]. It was also considered adequate to check the SEP and ARE_{pred} values, where the minimum possible values are desirable.

Acknowledgements

EBM acknowledges the Unioeste (http://www.unioeste.br), for supporting his doctoral thesis, and the Institute of Chemistry of the Unicamp (http://www.iqm.unicamp.br). JPAM and MMCF acknowledge FAPESP (2004/04686-5) and CNPQ for research funding. TCMJ and MCF acknowledge Unioeste.

References

[1] V. Purohit, A.K. Basu, Chem. Res. Toxicol. 13 (2000) 673–692.
[2] R. Edenharder, X. Tang, Food Chem. Toxicol. 35 (1997) 357–372.
[3] M.D. Maron, B.N. Ames, Mutat. Res. 113 (1983) 173–215.
[4] C.R.M. Silva, M.M.V. Naves, Rev. Nutr. 14 (2000) 135–143.
[5] V.M. Oliveira, J.M. Aldrighi, J.F. Rinaldi, Rev. Assoc. Med. Bras. 52 (2006) 453–459.
[6] A.S. Tsao, E.S. Kim, W.K. Hong, CA Cancer, J. Clin. 54 (2004) 150–180.
[7] J. Lameira, I.G. Medeiros, M. Reis, A.S. Santos, C.N. Alves, Bioorg. Med. Chem. 14 (2006) 7105–7112.
[8] M.Y. Heo, S.J. Sohn, W.W. Au, Mutat. Res. Fundam. Mol. Mech. Mutagen. 488 (2001) 135–150.
[9] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, J. Chemometr. 23 (2009) 32–48.
[10] Pirouette Software Version 4. Informetrix Inc., USA, 2007.
[11] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2003) 69–77.
[12] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Environ. Health Perspect. 111 (2003) 1361–1375.
[13] G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, J. Markopolus, O. Igglessi-Markopoulou, J. Comput. -Aided Mol. Des. 21 (2007) 251–267.
[14] R. Kiralj, M.M.C. Ferreira, J. Braz. Chem. Soc. 20 (2009) 770–787.
[15] Organization for Economic Co-Operation and Development Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q) SAR] Models pp. 154. OECD, Paris, 2007. http://appli1.oecd.org/olis/2007doc.nsf/linkto/env-jm-mono (2007)2 (accessed: october 2009).
[16] A. Golbraikh, A. Tropsha, J. Mol. Grap. Modell. 20 (2002) 269–276.
[17] A.O. Aptula, N.G. Jeliazkova, T.W. Schultz, M.T.D. Cronin, QSAR Comb. Chem. 24 (2005) 385–396.
[18] DRAGON Software Web Version 3.0. Talete Srl., Italy, 2003.
[19] O. Farkas, J. Jakus, K. Héberger, Molecules 9 (2004) 1079–1088.
[20] F.T. Hatch, F.C. Lightstone, M.E. Colvin, Envirom. Mol. Mutagen. 35 (2000) 279–299.
[21] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, 2000, pp. 667.
[22] A.H. Morales, P.R. Duchowicz, M.Á.C. Pérez, E.A. Castro, M.N.D.S. Cordeiro, M.P. González, Chemom. Intell. Lab. Syst. 81 (2006) 180–187.
[23] R. Put, Q.S. Xu, D.L. Massart, Y.V. Heyden, J. Chromatogr. A 1055 (2004) 11–19.
[24] B.F. Rasulev, N.D. Abdullaev, V.N. Syrov, J. Leszczynski, QSAR Comb. Chem. 24 (2005) 1056–1065.
[25] V. Consonni, R. Todeschini, M. Pavan, J. Chem. Inf. Comp. Sci. 42 (2002) 682–692.
[26] M.P. González, C. Terán, M. Teijeira, M.J. González-Moa, Eur. J. Med. Chem. 40 (2005) 1080–1086.
[27] Cambridge Structural Database Software Version 5.29-2007 + 1 update. Cambridge Crystallographic Data Centre, England, November 2007.
[28] HyperChem Software Version 7.1. Hyper Co., USA, 2002.
[29] Gaussian 03W Software Version 6.0. Gaussian Inc., USA, 2003.
[30] F.A. Molfetta, A.T. Bruni, F.P. Rosseli, A.B.F. Silva, Struct. Chem. 18 (2007) 49–57.
[31] Parameter Client Interface. Virtual Computational Chemistry Laboratory, 2005.http://www.vcclab.org/lab/pclient (accessed: october 2009).
[32] ALOGPS Interface Version 2.1. Virtual Computational Chemistry Laboratory, 2005.http://www.vcclab.org/lab/alogps (accessed: october 2009).
[33] M.M.C. Ferreira, J. Braz. Chem. Soc. 13 (2002) 742–753.
[34] S. Wold, E. Johansson, M. Cocchi, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design, Kluwer/Escom, Dordrecht, 2000, pp. 523–550.
[35] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[36] C. Rücker, G. Rücker, M. Meringer, J. Chem. Inf. Model. 47 (2007) 2345–2357.
[37] S. Wold, L. Eriksson, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, Wiley-VCH, Weinheim, 1995, pp. 309–318.
[38] A.C. Gaudio, E. Zandonade, Quim. Nova 24 (2001) 658–671.
[39] P.P. Roy, J.T. Leonard, K. Roy, Chemom. Intell. Lab. Syst. 90 (2008) 31–42.
[40] P.P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302–313.