Bioorganic & Medicinal Chemistry 20 (2012) 4848–4855

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Rational drug design for anti-cancer chemotherapy: Multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents

Alejandro Speck-Planche[a,*], Valeria V. Kleandrova[b], Feng Luan[a,c], M. Natália D. S. Cordeiro[a,*]

[a] REQUIMTE/Department of Chemistry and Biochemistry, University of Porto, Porto 4169-007, Portugal
[b] Faculty of Technology and Production Management, Moscow State University of Food Production, Volokolamskoe shosse 11, Moscow, Russia
[c] Department of Applied Chemistry, Yantai University, Yantai 264005, PR China

ARTICLE INFO

Article history:
Received 9 February 2012
Revised 21 May 2012
Accepted 25 May 2012
Available online 15 June 2012

Keywords:
Anti-CRC activity
Artificial neural network
CRC cell lines
In silico design
Linear discriminant analysis
mt-QSAR
Quantitative contributions

ABSTRACT

The discovery of new and more potent anti-cancer agents constitutes one of the most active fields of research in chemotherapy. Colorectal cancer (CRC) is one of the most studied cancers because of its high prevalence and number of deaths. In the current pharmaceutical design of more efficient anti-CRC drugs, the use of methodologies based on Chemoinformatics has played a decisive role, including Quantitative-Structure–Activity Relationship (QSAR) techniques. However, until now, there is no methodology able to predict anti-CRC activity of compounds against more than one CRC cell line, which should constitute the principal goal. In an attempt to overcome this problem we develop here the first multi-target (mt) approach for the virtual screening and rational in silico discovery of anti-CRC agents against ten cell lines. Here, two mt-QSAR classification models were constructed using a large and heterogeneous database of compounds. The first model was based on linear discriminant analysis (mt-QSAR-LDA) employing fragment-based descriptors while the second model was obtained using artificial neural networks (mt-QSAR-ANN) with global 2D descriptors. Both models correctly classified more than 90% of active and inactive compounds in training and prediction sets. Some fragments were extracted from the molecules and their contributions to anti-CRC activity were calculated using mt-QSAR-LDA model. Several fragments were identified as potential substructural features responsible for the anti-CRC activity and new molecules designed from those fragments with positive contributions were suggested and correctly predicted by the two models as possible potent and versatile anti-CRC agents.

1. Introduction

Colorectal cancer (CRC) is nowadays ranked the third most common in the world taking into consideration the number of patients affected by cancer in their lifetime.[1] This cancer is the result of the uncontrolled cell growth in the colon, rectum, or appendix, with an incidence of 700 per million of people in western societies each year.[2] The most important element in CRC is that 50% of patients eventually present metastasis which spreads mainly to the liver but also to the lungs, peritoneal cavity and bones.[1] Until now, the current chemotherapy for the treatment of CRC is only used together with surgery in certain cases increasing life expectancy 10–15%.[3] Thousands of compounds have been synthesized and tested for anti-CRC activity.[4] However, the current pharmaceutical design of anti-CRC drugs requires more efficient and rational methodologies, because the serendipitous synthesis and design of these chemotherapeutic agents lead to remarkable consumption of financial resources and time.[5]

Computational approaches have played an important role in rational drug design.[6–8] In the field related with the design of more effective anti-CRC agents, several works have been reported,[9–18] providing essential insights about the future perspectives concerning the drug discovery. However, almost all prior studies reported present two remarkable disadvantages. From one side, the in silico methodologies employed for the design of anti-CRC agents, use small databases of analogous compounds. On the other hand, the studies were carried out usually considering one biomolecular target like protein or cell line associated to CRC. Thus, it is not possible to explore in deeper way, the structural patterns which could be related with development of anti-CRC activity. For this reason, in the area related with Chemoinformatics and techniques concerning Quantitative-Structure–Activity Relationships (QSAR),[5] several researchers have made emphasis on the development of new multi-target QSAR (mt-QSAR) models, which permit the prediction of different biological profiles against more than one biological entity (protein, microorganism, partition system, tissue).[19–24]

Prediction of activity of selected molecules in the field of drug discovery is not always possible with linear models. In some cases there is no strong linear relationship between the inputs (drug structures) and the outputs (biological activities). For this reason,

* Corresponding authors. Fax: +351 220402659.
E-mail addresses: alejspivanovich@gmail.com (A. Speck-Planche), ncordeir@fc.up.pt (M.N.D.S. Cordeiro).

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmc.2012.05.071

non-linear input/output relationships should be considered. In this sense, non-linear Artificial Neural Networks (ANNs) have become promising tools for virtual screening of drug candidates.[21,25]

The previous ideas demonstrate that the discovery of versatile and highly active compounds against several CRC cell lines should constitute a major interest. This fact can help to prevent the appearance of resistance to anti-CRC drugs. In this work, we develop the first unified computational approach by exploring linear and non-linear classification techniques. Here, two mt-QSAR models are created for the virtual screening, prediction and in silico design of potent and versatile anti-CRC compounds against ten of the most important and best studied CRC cell lines.

2. Materials and methods

2.1. Fragment-based descriptors

Nowadays, there is an explosion in the quantity of molecular descriptors.[26] More than 4000 of these structural variables have been reported in the literature for use in Chemoinformatics.[27] In this work, we selected the descriptors known as functional group counts (FGC) and atom-centered fragments (ACF). These descriptors have been successfully applied in several QSAR studies.[28–30] They provide very useful information about specific fragments or functional groups in the molecules,[31] and their abilities to participate in hydrophobic and dispersive interactions,[32] or to exhibit a defined chemical reactivity.[28–30] These fragment-based descriptors mentioned above, have some similarities with the variables used in a Free-Wilson analysis.[5] Other fragment-based descriptors chosen by us were the spectral moments of the bond adjacency matrix (μ_k) which are the essence of the TOPS-MODE (topological **s**ubstructural molecular design) approach. These descriptors have been widely employed in QSAR studies,[33–35] and for the assessment of multiple toxicological profiles.[36–38] In mathematical terms,[39–41] the μ_k is the sum of the main diagonal elements (e_{ii}) of the matrix $\boldsymbol{B}^k$:

$$\mu_k = Tr(B^k) = \sum_{i=1}^{s} (e_{ii})^k \tag{1}$$

where **Tr** means the trace of the matrix, that is, the sum of the diagonal entries of the matrix. The elements (e_{ii}) are bond weights which represent different physicochemical properties which include dipole moments, standard bond distances, or mathematical expressions involving atomic weights such as polarizability, polar surface area, molar refractivity and hydrophobicity. Although μ_k descriptors have topological nature, they can be used along or in combination with other fragment-based descriptors to calculate the quantitative contribution of any fragment to the desired activity.[28–30,33]

2.1.1. Data set: fragment based-descriptors and development of the discriminant model

The complete data set was formed by 571 compounds with anti-CRC activity against ten CRC cell lines.[4] Not all the compounds were tested against all the CRC cell lines. We had also 100 drugs extracted from the Merck Index.[42] These drugs have other profiles other than anti-CRC activity, and for this reason they were considered as inactive. The FGC and ACF were calculated using DRAGON v5.3.[31] The μ_k descriptors (from order 1 to 15), were calculated using MODESLAB v1.5,[43] and they were weighted by properties such as dipole moment, molar refractivity, and atomic weight. Linear discriminant analysis (LDA) was used to construct the classifier model.[44] The prediction of anti-CRC activity of compounds against 10 CRC cell lines is not a classical LDA problem. In this sense, we opted for one of the diverse mt-QSAR methodologies which have been developed by González-Díaz and co-workers in drug design.[19–23] Thus, we used a similar methodology to that developed for the prediction of enzyme classes in *Leishmania infantum*.[20] The same methodology has been generalized to the prediction of GSK-3 inhibitors,[45] and a detailed description of the procedure applied here, was recently published in the work of Speck-Planche et al. for the in silico design of anti-prostate cancer agents.[46] Here, a binary discriminant function was generated for the classification of compounds: those which belonged to a particular active group (inhibitory activity against a specific CRC cell line) and compounds that did not belong to this group (inactive). For this, the following steps were realized:

- First, the 571 compounds which belonged to the group of compounds with anti-CRC activity were divided according to their activity against the 10 CRC cell lines. Each CRC cell line had a cutoff value of anti-CRC activity (Table 1), that is, the values of inhibitory activity from which the compounds were considered as active. The variable selected as measure of cytotoxic activity was IC_{50}, that is, the ability of the compounds to inhibit at 50% the proliferation of the different CRC cell lines.[4] We need to point out that the cutoff values of IC_{50} were chosen in arbitrary way. However, they constitute in general terms, rigorous values which are comparable with IC_{50} of current drugs reported in the literature.[4]
- The data file was created by assigning to each compound 275 structural variables (inputs): 89 FGC, 51 ACF and 135 descriptors like μ_k. We had also; one output variable and one classification variable related with the type of CRC cell line (CRCCL). This last variable was an auxiliary variable, and it was not used to construct the mt-QSAR-LDA model. Thus, the row for each compound (input) in the spreadsheet contained 277 elements in total.
- The output variable is a categorical variable called anti-CRC activity (A_{CRC}); $A_{CRC} = 1$ if the compound has anti-CRC activity against any of the 10 CRC cell lines and −1, otherwise ($A_{CRC} = -1$). In the case of the 100 drugs which were considered as inactive, we could repeat each of these 100 drugs ten times corresponding to the ten CRC cell lines. We should emphasize that CRCCL code was only used to relate each compound with its corresponding type of CRC cell line, and thus, these compounds entered only once. Conversely, the 100 drugs (decoys) have more than one line entry with different CRCCL classes.

With this kind of organization of the data, a problem is generated. The original 45 μ_k descriptors (from order 1 to 15 and weighted by the three physicochemical properties mentioned above) are not able to discriminate the structural information about all the anti-CRC compounds which are active against the different CRC cell lines. For this reason, a LDA model based only on 45 μ_k (or any of the other 89 FGC or 51 ACF) would fail. The point is that we need to predict ten specific probabilities: each of them confirming the real CRCCL class. We solved this problem by introducing mt-descriptors which are characteristics of each CRCCL

Table 1
Cutoff values for anti-CRC activity in different cell lines

CRC cell line	Cutoff[a]	CRC cell line	Cutoff[a]
Caco-2	⩽3.000	LoVo	⩽0.400
COLO-205	⩽4.860	RKO	⩽2.000
DLD-1	⩽2.600	SW-620	⩽0.500
HCT-15	⩽0.200	WiDr	⩽2.000
HT-29	⩽0.005	LS-174T	⩽6.500

[a] Cutoff values are IC_{50} expressed in μM.

class. We used the average value of each μ_k descriptor ($avg\mu_k$) for all the active compounds tested against the same CRC cell line (same CRCCL class). We also calculated the deviation of the μ_k from the respective group ($dif\mu_k$) indicated in CRCCL. These last descriptors were calculated as the difference between the original μ_k descriptor of each compound and $avg\mu_k$. Therefore, we had (45 μ_k values) + (45 $avg\mu_k$ values) + (45 $dif\mu_k$ like deviation values) = 135 input variables like μ_k descriptors mentioned above. It is very important to understand that mt-descriptors were calculated only from μ_k because they are continuous numbers. The FGC and ACF descriptors are discrete variables and the calculation of an average would conduct to decimal numbers with no phenomenological or physical meaning. The names or codes of all the compounds used in this work appear in the Supplementary data 1 file (Supplementary data 1).

Then, we collected 1651 cases (compound/CRC cell line pairs) in total. We used as cross-validation method that known as independent test.[44] In this sense, the database was randomly split into two series: training and prediction sets. Anyway, we took into consideration that the proportion of cases in training/prediction sets should be 3/1, which means that training set would contain approximately 75% of the total cases, while 25% of the whole database would be in prediction set.[47] Thus, the training set was used to construct the model, and it was formed by 1237 cases, 487 of them considered as anti-CRC agents and 750 inactive. The prediction set was used to validate the model and to demonstrate its predictive power. This set was composed by 414 cases, 164 with anti-CRC activity and 250 inactive cases. The general expression for this mt-QSAR-LDA model is presented in the following form:

$$\mathrm{A_{CRC}} = a_0 + \sum_{\mathrm{k}} b_k \cdot D_{\mathrm{k}} + \sum_{\mathrm{k}} c_k \cdot avgD_{\mathrm{k}} + \sum_{\mathrm{k}} d_k \cdot difD_{\mathrm{k}} \tag{2}$$

where $\mathrm{A_{CRC}}$ is the real score which predicts the propensity of a compound to have anti-CRC activity (or not) against a defined CRC cell line. The term described as a_0 is the constant, b_k, c_k and d_k represent the coefficients of the variables in the model. The symbol D_{k} represents the different descriptors (FGC, ACF and/or μ_k), while $avgD_{\mathrm{k}}$ and $difD_{\mathrm{k}}$ are average and deviation values (only referred to μ_k) respectively. The discriminant function was obtained by employing the LDA modules of STATISTICA 6.0.[48] The variables included in the mt-QSAR-LDA model were selected using a forward stepwise procedure as the variable selection strategy. We took into consideration the principle of parsimony. For this reason, the best model was chosen as that with high statistical significance, but having as few parameters as possible. We assessed the quality of the model by calculating some statistical indices which are classically reported.[49] These statistics are: the Wilks' lambda (λ), the square of the Mahalanobis distance (D^2), the chi-square (χ^2), the Fisher's test (F) and the p-level. Wilk's λ is used to assess the statistical significance of the discriminatory power of any LDA model under study. Its value ranges from 0.0 (perfect discriminatory power) to 1.0 (no discriminatory power). At the same time the statistic D^2 measures the separation between two groups or classes. The larger D^2 the greater will be the separation (discrimination) between groups. The index χ^2 is employed to have an idea about the independence of two criteria of classification of qualitative data. A large value of χ^2 is associated with a great independency between the two groups or populations. The F value indicates the statistical significance of the model to fit well the data. The p-level is usually associated to the F value, and it is considered to be less than 0.05.[49] We ensured the quality and predictive power of the model by considering the percentages of correct classification of compounds, Mathew's correlation coefficient (MCC)[50] and the areas under the ROC curves[51] in both, training and prediction sets.

2.2. Global 2D descriptors

Molecular descriptors are constructed to encode different kinds of physicochemical and/or structural informations.[27] In order to have more complete idea about the potential relationships between the structural patterns and the development of the anti-CRC activity, we also considered global 2D descriptors.[26] Unlike fragment-based descriptors, which consider in some way the molecular structure as the contributions of its component parts, global 2D descriptors consider the molecules as a whole and important geometric (3D) elements.[26] The global 2D descriptors (100 in total) were calculated using DRAGON v5.3.[31] These include blocks of variables such as topological descriptors, atom connectivity indices, information indices, 2D-autocorrelations, Burden eigenvalues, topological charge indices, and eigenvalues-based indices. We also applied the previous mt-methodology to these descriptors. Thus, were able to collect 300 variables like global 2D descriptors. In order to seek the best mt-QSAR model, both, LDA and ANN techniques were analyzed using STATISTICA 6.0.[48]

3. Results and discussion

3.1. mt-QSAR-LDA model

The best model found using fragment-based descriptors and LDA, contained 13 descriptors which best described the anti-CRC activity in the whole database:

$$\begin{aligned}
\mathrm{A_{CRC}} &= 0.985(Car) - 0.972(CbH) - 3.110(RCOOH) \\
&-4.424(RCONH_2) + 10.890(RCHO) - 3.030(SO_2N) \\
&-1.682(ArX) + 9.268(Oxet) - 0.149(H\text{-}052) \\
&+3.063(N - 070) + 1.178 \cdot 10^{-9} \mu_{13}^{(Dip)} + 1.476 \cdot 10^{-3} avg\mu_3^{(MR)} \\
&+1.533 \cdot 10^{-10} dif\mu_5^{(Ato)} - 8.232
\end{aligned}$$

$$\begin{aligned}
&N = 1237 \; \lambda = 0.376 \; D^2 = 6.935 \; \chi^2 = 1201.00 \\
&F_{(13,1223)} = 155.99 \; p < 0.001
\end{aligned} \tag{3}$$

The symbology of the different descriptors in the Eq. 3, together with their corresponding meanings, appears summarized in Table 2. As we can see, the small value of λ, high value of D^2, large value χ^2, and the large value for the F statistics, demonstrates the quality of the model. An important element is the examination of the

Table 2
Molecular descriptors used in the mt-QSAR-LDA model

Descriptor	Definition[a]
Car	Number aromatic carbons
CbH	Number of unsubstituted benzene carbons
RCOOH	Number of aliphatic carboxylic groups
$RCONH_2$	Presence or absence of non-substituted aliphatic amide groups
RCHO	Presence or absence of aliphatic carbonyl groups (aldehydes)
SO_2N	Presence or absence of sulfonamide groups
ArX	Number of halogens on aromatic ring
Oxet	Number of oxetane rings
H-052	Number of hydrogen atoms attached to C^0(sp3) with 1X attached to next C
N-070	Number of fragments of type ArNHR
$\mu_{13}^{(Dip)}$	Spectral moment of order 13 weighted by the dipole moment
$avg\mu_3^{(MR)}$	Average spectral moment of order 3 weighted by the molar refractivity
$dif\mu_5^{(Ato)}$	Deviation of the spectral moment of order 5 weighted by the atomic weight

[a] **R** represents an aliphatic group; **Ar** represents aromatic group; **X** represents any electronegative atom (O, N, S, P, Se, halogens); the superscript represents the formal oxidation number. The formal oxidation number of a carbon atom equals the sum of the conventional bond orders with electronegative atoms.

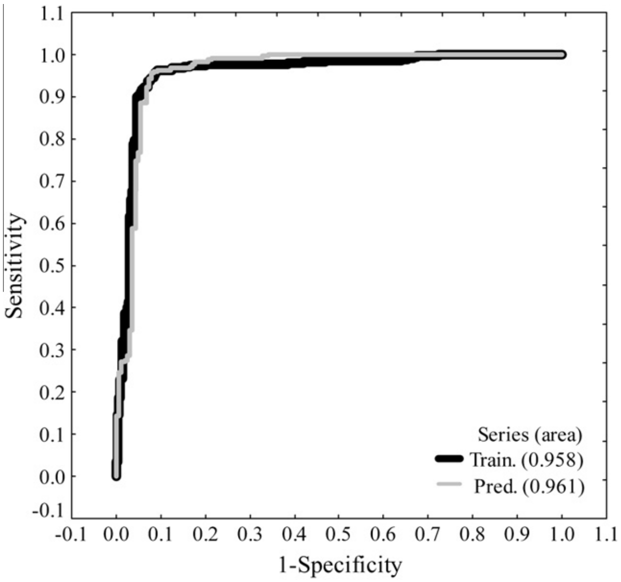

Figure 1. ROC curve for mt-QSAR-LDA.

p-level associated with the *F* value. Our model exhibits *p*-level less than 0.001, which means that the hypothesis of groups overlapping with a 5% error can be rejected.

The mt-QSAR-LDA model could correctly classify 1158 out of 1237 cases (93.61%) in the training sets, while 388 out of 414 cases (93.72%) were correctly classified in prediction sets. We took also as determinant statistical indices for the good performance of the model, the areas under the ROC curves. These values were 0.958 and 0.961 for training and prediction sets, respectively (Fig. 1), and by this mean we proved that our model is not a random classifier because the areas under the ROC curves are different and statistically significant from those obtained by random classifiers (area = 0.5).

3.2. mt-QSAR-ANN model

Although for the case of the global 2D descriptors both, LDA and ANN techniques were analyzed, the best model found by us was an ANN using radial basis function (RBF). The profile of the best mt-QSAR ANN was as follows: RBF 11:11–311–1:1. The symbology of the different global 2D descriptors appears depicted in Table 3. The mt-QSAR-ANN could correctly classify 1187 out of 1237 cases (95.96%) in the training sets, while 382 out of 414 cases (92.75%) were correctly classified in prediction sets (see Supplementary data 2). The values of area under the ROC curves for mt-QSAR-ANN were 0.995 and 0.969 for training and prediction sets, respectively (Fig. 2). Thus, these values prove that mt-QSAR-ANN model is not a random classifier. The classification results which permit to ensure the quality and predictive power of both, mt-QSAR-LDA and mt-QSAR-ANN models, appear summarized in Table 4. We need to point out an extremely important element. We obtained two mt-QSAR models with different statistical classification techniques, using different kinds of variables. For this reason, the same training and prediction sets were used for both mt-QSAR models. From one side, this aspect ensures the correct and rigorous comparison between the models. On the other hand, the mt-QSAR models can be used in combination for future discovery and virtual screening of anti-CRC agents. The names or codes, predicted anti-CRC activities, and posterior probabilities (only for the case of mt-QSAR-LDA) for each compound expressed as percentages, are recorded in the Supplementary data 3 file (**Suppl. Inf. 3**). According to the statistical indices in Eq. 3, the results from Table 4 and the values of areas under the ROC curves, we can say that mt-QSAR-ANN model has better statistical quality than mt-QSAR-LDA. Also, the mt-QSAR-ANN model employs smaller number of variables to assess the anti-CRC activity (11 descriptors against 13 used by the mt-QSAR-LDA). Anyway, both mt-QSAR models display excellent qualities and predictive powers which are comparable with other reports in the literature related to the use of the mt-methodologies combined with LDA and/or ANN techniques.[19–25,28–30]

Table 3
Molecular descriptors used in the mt-QSAR-ANN model

Descriptor	Definition
CIC3	Complementary information content (neighborhood symmetry of 3-order)
MATS5v	Moran autocorrelation—lag 5/weighted by atomic van der Waals volumes
MATS6v	Moran autocorrelation—lag 6/weighted by atomic van der Waals volumes
GATS5m	Geary autocorrelation—lag 5/weighted by atomic masses
avgGATS3e	Average Geary autocorrelation—lag 3/weighted by Sanderson electronegativities
avgGATS7e	Average Geary autocorrelation—lag 7/weighted by Sanderson electronegativities
difMATS2v	Deviation of Moran autocorrelation—lag 2/weighted by atomic van der Waals volumes
difGATS1v	Deviation of Geary autocorrelation—lag 1/weighted by atomic van der Waals volumes
difGATS4v	Deviation of Geary autocorrelation—lag 4/weighted by atomic van der Waals volumes
difJGI2	Deviation of mean topological charge index of order 2
difSEigv	Deviation of the eigenvalue sum from van der Waals weighted distance matrix

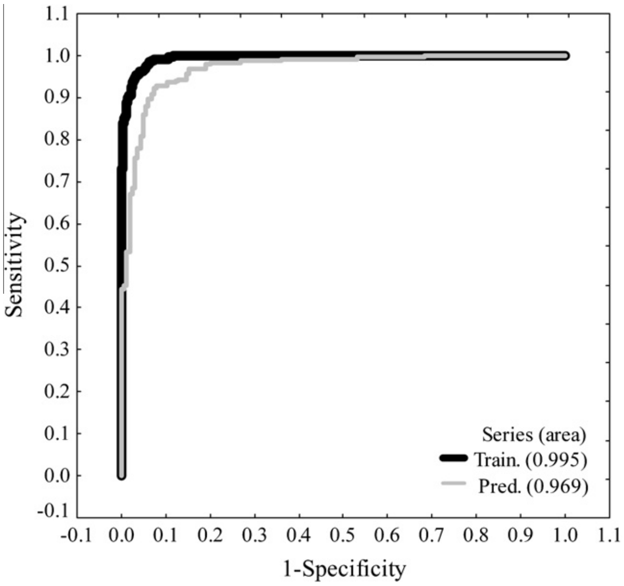

Figure 2. ROC curve for mt-QSAR-ANN.

3.3. Fragment contributions

The calculation of quantitative contributions of fragments to the desired biological activity is an important concept in the fields of pharmaceutical design, medicinal chemistry and drug discovery.[28–30,33] If any fragment is selected and its corresponding contribution to activity is calculated, we can have essential information about the potential relationship between this fragment and the appearance of the pharmacological profile under study.[22] Also, the calculation of quantitative contributions can offer a very useful

Table 4
Results of classification

Classification	Training series				Prediction series			
	mt-QSAR-LDA		mt-QSAR-ANN		mt-QSAR-LDA		mt-QSAR-ANN	
	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive
Total	487	750	487	750	164	250	164	250
Correct[a]	450	708	467	720	152	236	152	230
Wrong	37	42	20	30	12	14	12	20
Correct[b] (%)	92.40	94.40	95.89	96.00	92.68	94.40	92.68	92.00
Wrong (%)	7.60	5.60	4.11	4.00	7.32	5.60	7.32	8.00
Acc[c] (%)	93.61		95.96		93.72		92.27	
MCC	0.866		0.916		0.869		0.840	

[a] Compounds which were correctly classified by the model.
[b] Formally known as sensitivity for active and specificity for inactive.
[c] Referred to the accuracy as total percentage of correct classification.

guide to pharmaceutical and medicinal chemists for the automatic generation of 2D pharmacophores for drug discovery.[34] As stated in previous works, the calculation of quantitative contributions should be carried out only with the use of linear regression techniques.[28–30,33–38] With this aim, we employed our mt-QSAR-LDA model. All the descriptors employed in this model encode useful information about fragments and at the same time they comply with the rule of linear additivity. Then, we were able to calculate the quantitative contribution of any fragment to anti-CRC activity. In this sense, some fragments were represented (Fig. 3) and their quantitative contributions to the anti-CRC activity against the 10 CRC cell lines were calculated (Table 5). The following procedure was realized:

All descriptors for each fragment according to the Eq. 3, were calculated. After, all the scores of contributions of each fragment against the ten CRC cell lines were obtained by substituting the descriptors into the mt-QSAR-LDA equation using the Microsoft Excel application. Finally, the quantitative contributions of all molecular fragments were standardized using the score of contribution less the total average and divided by the standard deviation. This procedure was applied for González-Díaz and co-workers for the better interpretation of quantitative contributions of fragments to the biological activity in mt-QSAR models.[22] An important aspect about the calculation of fragment contributions to anti-CRC activity is that we can select in fast and efficient way, some suitable fragments which could be considered as potential substructural patterns for the discovery of new and versatile anti-CRC agents. For example, **F3**–**F7**, **F12**, **F16**, and **F19** could comply with the desirability for development of compounds with anti-CRC profile due to their remarkable high positive contributions against the 10 CRC cell lines. Thus, in principle, new molecules can be generated from these fragments mentioned above.

3.4. In silico generation of compounds with anti-CRC activity

The mt-chemoinformatic approach introduced in this study focuses on the construction of two mt-QSAR models. While one of them will be used for fast generation of substructural patterns related with the anti-CRC activity (mt-QSAR-LDA), the other will be used to confirm if the design was correctly performed (mt-QSAR-ANN). We need to emphasize that the two mt-QSAR models should be simultaneously employed to predict the anti-CRC activity of compounds against the ten CRC cell lines. For this reason, only when a compound is predicted as anti-CRC agent (against a defined CRC cell line) by both, mt-QSAR-LDA and mt-QSAR-ANN, the compound will be really considered as possible anti-CRC agent.

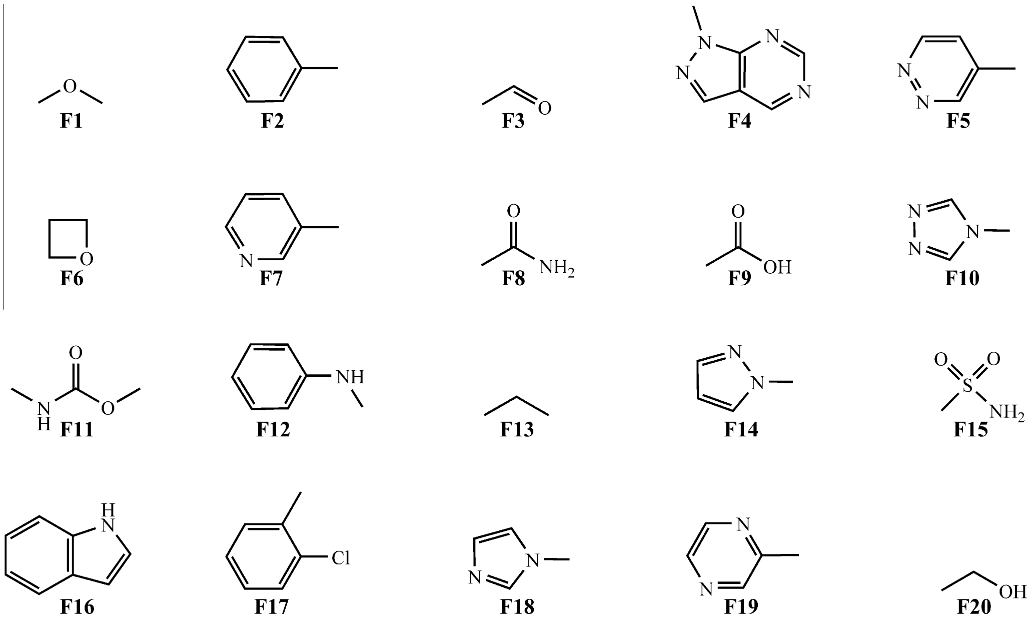

Figure 3. Different fragments which were found in the molecules.

Table 5
Quantitative contributions of different fragments to the anti-CRC activity

ID	Caco-2	COLO-205	DLD-1	HCT-15	HT-29	LoVo	RKO	SW-620	WiDr	LS-174T
F1	−0.478	−0.715	−0.670	−0.710	−0.515	−0.646	−0.696	−0.511	−0.661	−0.751
F2	−0.190	−0.428	−0.382	−0.422	−0.227	−0.359	−0.408	−0.223	−0.373	−0.464
F3	**2.508**	**2.270**	**2.315**	**2.275**	**2.471**	**2.339**	**2.290**	**2.475**	**2.325**	**2.234**
F4	**0.873**	**0.635**	**0.680**	**0.641**	**0.836**	**0.704**	**0.655**	**0.840**	**0.690**	**0.599**
F5	**0.603**	**0.365**	**0.410**	**0.370**	**0.565**	**0.434**	**0.384**	**0.569**	**0.419**	**0.329**
F6	**2.063**	**1.825**	**1.870**	**1.831**	**2.026**	**1.894**	**1.845**	**2.030**	**1.880**	**1.789**
F7	**0.873**	**0.635**	**0.680**	**0.640**	**0.835**	**0.704**	**0.654**	**0.839**	**0.689**	**0.599**
F8	−1.687	−1.925	−1.880	−1.920	−1.724	−1.856	−1.905	−1.720	−1.870	−1.961
F9	−1.325	−1.563	−1.518	−1.558	−1.363	−1.494	−1.544	−1.358	−1.509	−1.599
F10	0.062	−0.176	−0.130	−0.170	0.025	−0.107	−0.156	0.029	−0.121	−0.211
F11	−0.473	−0.711	−0.666	−0.706	−0.511	−0.642	−0.692	−0.507	−0.657	−0.747
F12	**0.650**	**0.412**	**0.457**	**0.417**	**0.612**	**0.481**	**0.432**	**0.617**	**0.467**	**0.376**
F13	−0.478	−0.716	−0.670	−0.710	−0.515	−0.647	−0.696	−0.511	−0.661	−0.752
F14	0.332	0.095	0.140	0.100	0.295	0.163	0.114	0.299	0.149	0.059
F15	−0.791	−1.029	−0.983	−1.023	−0.828	−0.960	−1.009	−0.824	−0.974	−1.065
F16	**0.618**	**0.380**	**0.425**	**0.385**	**0.580**	**0.449**	**0.399**	**0.585**	**0.434**	**0.344**
F17	−0.380	−0.618	−0.572	−0.612	−0.417	−0.549	−0.598	−0.413	−0.563	−0.654
F18	0.332	0.095	0.140	0.100	0.295	0.163	0.114	0.299	0.149	0.059
F19	**0.603**	**0.365**	**0.410**	**0.370**	**0.565**	**0.434**	**0.384**	**0.569**	**0.419**	**0.329**
F20	−0.559	−0.797	−0.752	−0.792	−0.597	−0.728	−0.777	−0.592	−0.743	−0.833

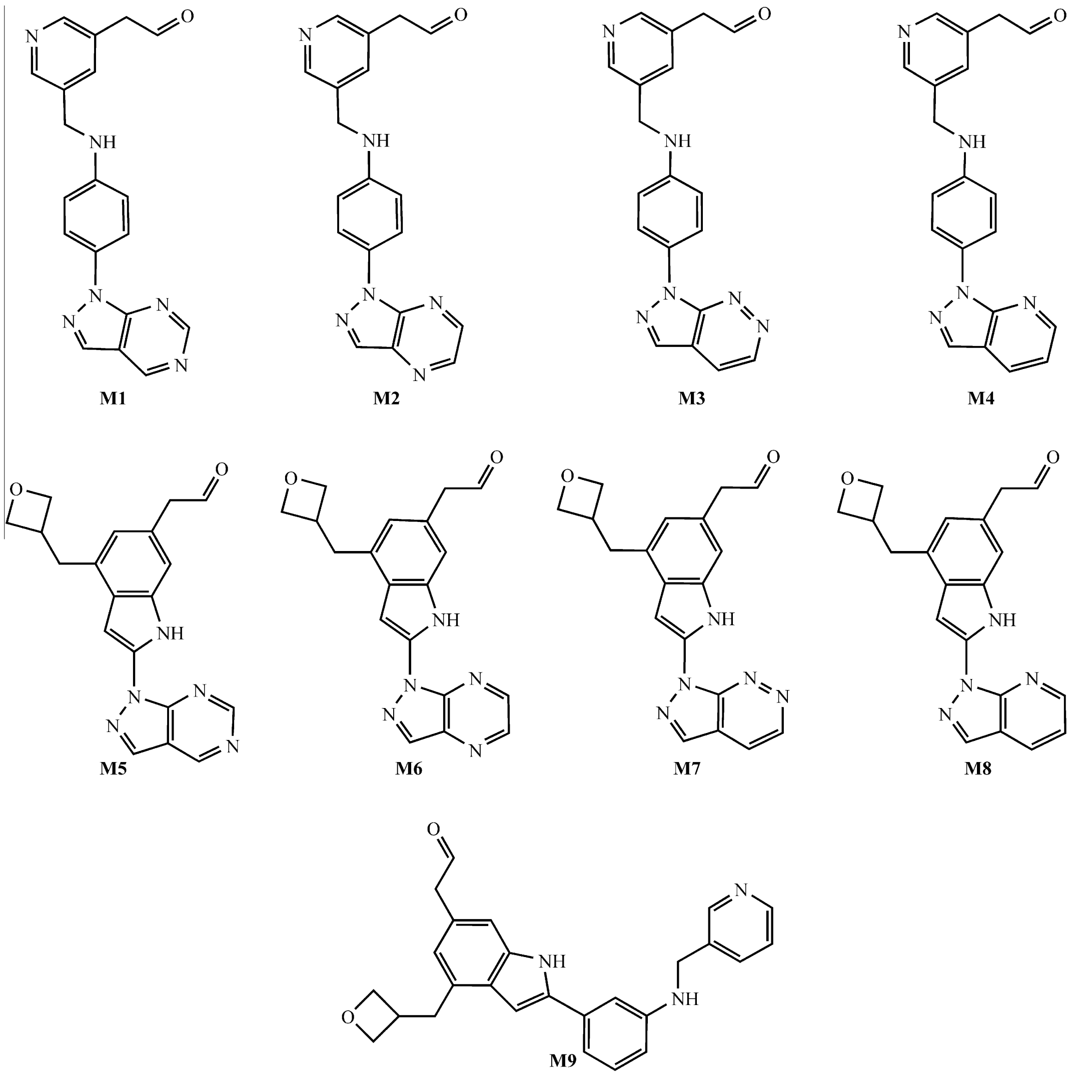

Figure 4. New molecules suggested as possible anti-CRC agents.

This is the determinant element because the mt-QSAR-LDA will assess the biological activity under study, taking into consideration the molecules as contributions of their component parts, while the mt-QSAR-ANN model, which has better quality, will be focused

Table 6
Results of the predictions for the mt-QSAR-ANN model

	Predicted activity[a]									
ID	Caco-2	COLO-205	DLD-1	HCT-15	HT-29	LoVo	RKO	SW-620	WiDr	LS-174T
M1	1	1	1	1	1	1	1	1	-1	1
M2	1	1	1	1	−1	1	1	1	1	1
M3	1	1	1	1	1	1	1	1	1	1
M4	1	1	1	1	−1	1	1	1	−1	1
M5	1	1	1	1	1	1	1	1	1	1
M6	1	1	1	1	1	1	1	1	−1	1
M7	1	1	1	1	1	1	1	1	1	1
M8	1	1	1	1	−1	1	1	1	−1	1
M9	1	1	1	1	1	1	1	1	1	1

[a] The value 1 is used when the compounds were predicted as active and −1, when the compounds were predicted as inactive.

on the prediction of the same biological activity considering global properties of the molecules, that is, the molecular entities as a whole as we commented above. With all these ideas in mind, we designed nine molecules (Fig. 4), which were the result of analyzing the structural information contained in the descriptors in the Eq. 3, and the quantitative contributions of the suitable fragments discussed above. We calculated the probabilities of these nine molecules to have anti-CRC activity against the ten CRC cell lines. In the case of the mt-QSAR-LDA model, the probabilities for all the molecules to be anti-CRC agents were 100% against all CRC cell lines under study. The results of the predictions for the mt-QSAR-ANN model are depicted in Table 6. These results show a strong agreement between both mt-QSAR models in their abilities to predict anti-CRC agents against the 10 CRC cell lines. We predicted nine molecules against 10 CRC cell lines. So, we predicted 90 cases. The mt-QSAR-ANN model confirms that in 83 out of 90 (92.22%), the predictions could be correct.

Some divergences are observed in several compounds against two CRC cell lines: **HT-29** and **WiDr**. For example, compounds **M1**–**M3** are isomers, where the only difference is the ring fused to pyrazole moiety, that is, pyrimidine, pyrazine, and pyridazine, respectively. Only **M3** is active against all the CRC lines, which means that pyridazine ring is essential. The substitution of this ring by any of the other two will inactivate the compounds against **HT-29** or **WiDr**, while the replacement of the same ring by a pyridine will cause possible lack of activity against both CRC cell lines mentioned above (see structure of compounds **M4**). Something similar take place in the isomers **M5**–**M7** with the difference that now, pyrimidine ring is as desirable as pyridazine ring. Thus, **M5** and **M7** will be active against all CRC cell lines while **M6** will have the same profile with the exception of the CRC cell line **WiDr**. The substitution of pyrimidine by pyridine will cause also, inactivation against **HT-29** and **WiDr**. Finally, **M9** which take into consideration substructures present in the two families of isomers discussed above, is active also against all CRC cell lines. All results explained here, are clearly dependent on the cutoff values of activity selected by us. Anyway, we are demonstrating that the molecular descriptors used in the mt-QSAR-ANN model are very sensitive to small changes in the structure of the molecules, even by simple replacement of carbon by nitrogen. Then, considering the information given previously, all the nine molecules could be considered as potential and versatile molecular entities with possible anti-CRC activity, with special priority in the compounds **M3**, **M5**, **M7** and **M9**.

4. Conclusions

Chemoinformatics has contributed in considerable way to a better understanding of the relationships between the structures of the molecules and their anti-cancer profiles. Our chemoinformatic approach reported here, pretends to assess the anti-CRC activity of large and heterogeneous databases of compounds against ten different CRC cell lines by employing two mt-QSAR models. Both models predict more efficiently the biological profile under study and in more general situations than other classical QSAR methodologies which use very restricted groups of analogous compounds against only one biological receptor or CRC cell line. By analyzing both, the mt-QSAR-LDA and mt-QSAR-ANN models, is possible to have a deeper insight and knowledge about the structural features which can be directly related with the appearance of anti-CRC activity. The present unified chemoinformatic approach based in the creation of two mt-QSAR models discussed above, can serve as a guide for the fast and rational in silico generation of anti-CRC leads, opening new horizons in the medicinal chemistry of anti-CRC drugs, and more generally, in the search for more efficient anti-cancer chemotherapies.

Acknowledgments

The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) and the European Social Found for financial support (SFRH/BPD/63666/2009). Moreover, this work has been further supported by FCT through project PTDC/QUIQUI/113687/2009 and Grant no. Pest-C/EQB/LA0006/2011.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bmc.2012.05.071.

References and notes

1. *Metastasis of Colorectal Cancer*; Beauchemin, N., Huot, J., Eds.; Springer Science+Business Media B.V.: Dordrecht, Heidelberg, London, New York, 2010.
2. Georgios, T.; Manousos-Georgios, P. In *Current Cancer Treatment—Novel Beyond Conventional Approaches*; Özdemir, Ö., Ed.; InTech: Rijeka, 2011; p 41.
3. Cunningham, D.; Atkin, W.; Lenz, H. J.; Lynch, H. T.; Minsky, B.; Nordlinger, B.; Starling, N. *Lancet* **2010**, *375*, 1030.
4. EBI-Team. ChEMBL Database. http://www.ebi.ac.uk/chembldb/, 2010.
5. Kubinyi, H. *QSAR: Hansch analysis and related approaches*; VCH Publishers: Weinheim, New York, Basel, Cambridge, Tokyo, 1993.
6. Oprea, T. *Chemoinformatics in Drug Discovery*; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, 2005.
7. Gasteiger, J. *Handbook of Chemoinformatics*; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, 2003.
8. Larson, R. S. *Bioinformatics and Drug Discovery*; Humana Press Inc.: Totowa, New Jersey, 2006.
9. Sharma, S. K.; Kumar, P.; Narasimhan, B.; Ramasamy, K.; Mani, V.; Mishra, R. K.; Majeed, A. B. *Eur. J. Med. Chem.* **2012**, *48*, 16.
10. Girgis, A. S.; Stawinski, J.; Ismail, N. S.; Farag, H. *Eur. J. Med. Chem.* **2012**, *47*, 312.
11. Drakulic, B. J.; Stanojkovic, T. P.; Zizak, Z. S.; Dabovic, M. M. *Eur. J. Med. Chem.* **2011**, *46*, 3265.
12. Verma, R. P.; Hansch, C. *Eur. J. Med. Chem.* **2010**, *45*, 1470.
13. Nolan, K. A.; Doncaster, J. R.; Dunstan, M. S.; Scott, K. A.; Frenkel, A. D.; Siegel, D.; Ross, D.; Barnes, J.; Levy, C.; Leys, D.; Whitehead, R. C.; Stratford, I. J.; Bryce, R. A. *J. Med. Chem.* **2009**, *52*, 7142.

14. Zheng, X.; Ekins, S.; Raufman, J. P.; Polli, J. E. *Mol. Pharm.* **2009**, *6*, 1591.
15. Parra-Delgado, H.; Compadre, C. M.; Ramirez-Apan, T.; Munoz-Fambuena, M. J.; Compadre, R. L.; Ostrosky-Wegman, P.; Martinez-Vazquez, M. *Bioorg. Med. Chem.* **2006**, *1889*, 14.
16. Niculescu-Duvaz, D.; Niculescu-Duvaz, I.; Friedlos, F.; Martin, J.; Lehouritis, P.; Marais, R.; Springer, C. J. *J. Med. Chem.* **2003**, *46*, 1690.
17. Munteanu, C. R.; Magalhaes, A. L.; Uriarte, E.; Gonzalez-Diaz, H. J. *Theor. Biol.* **2009**, *257*, 303.
18. Vilar, S.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. J. *Theor. Biol.* **2009**, *261*, 449.
19. Vina, D.; Uriarte, E.; Orallo, F.; Gonzalez-Diaz, H. *Mol. Pharm.* **2009**, *6*, 825.
20. Concu, R.; Dea-Ayuela, M. A.; Perez-Montoto, L. G.; Bolas-Fernandez, F.; Prado-Prado, F. J.; Podda, G.; Uriarte, E.; Ubeira, F. M.; Gonzalez-Diaz, H. J. *Proteome Res.* **2009**, *8*, 4372.
21. Gonzalez-Diaz, H.; Prado-Prado, F.; Sobarzo-Sanchez, E.; Haddad, M.; Maurel Chevalley, S.; Valentin, A.; Quetin-Leclercq, J.; Dea-Ayuela, M. A.; Teresa Gomez-Munos, M.; Munteanu, C. R.; Jose Torres-Labandeira, J.; Garcia-Mera, X.; Tapia, R. A.; Ubeira, F. M. *J. Theor. Biol.* **2011**, *276*, 229.
22. Prado-Prado, F. J.; Garcia-Mera, X.; Gonzalez-Diaz, H. *Bioorg. Med. Chem.* **2010**, *18*, 2225.
23. Cruz-Monteagudo, M.; Gonzalez-Diaz, H.; Aguero-Chapin, G.; Santana, L.; Borges, F.; Dominguez, E. R.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **1909**, *2007*, 28.
24. Prado-Prado, F. J.; Gonzalez-Diaz, H.; de la Vega, O. M.; Ubeira, F. M.; Chou, K. C. *Bioorg. Med. Chem.* **2008**, *16*, 5871.
25. Gonzalez-Diaz, H.; Muino, L.; Anadon, A. M.; Romaris, F.; Prado-Prado, F. J.; Munteanu, C. R.; Dorado, J.; Sierra, A. P.; Mezo, M.; Gonzalez-Warleta, M.; Garate, T.; Ubeira, F. M. *Mol. Biosyst.* **1938**, *2011*, 7.
26. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; WILEY-VCH Verlag GmbH: Weinheim, New York, Chichester, Brisbane, Singapore, Toronto, 2000.
27. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, 2009.
28. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. *Eur. J. Med. Chem.* **2011**, *46*, 5910.
29. Speck-Planche, A.; Kleandrova, V. V.; Rojas-Vargas, J. A. *Mol. Divers.* **2011**, *15*, 901.
30. Speck-Planche, A.; Scotti, M. T.; García-López, A.; Emerenciano, V. P.; Molina-Pérez, E.; Uriarte, E. *Mol. Divers.* **2009**, *13*, 445.
31. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. DRAGON for Windows (Software for Molecular Descriptor Calculations), v5.3; Milano Chemometrics and QSAR Research Group: Milano, 2005.
32. Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163.
33. Estrada, E.; Peña, A. *Bioorg. Med. Chem.* **2000**, *8*, 2755.
34. Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. *J. Med. Chem.* **1975**, *2000*, 43.
35. Perez Gonzalez, M.; Gonzalez Diaz, H.; Molina Ruiz, R.; Cabrera, M. A.; Ramos de Armas, R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1192.
36. Estrada, E.; Patlewicz, G.; Chamberlain, M.; Basketter, D.; Larbey, S. *Chem. Res. Toxicol.* **2003**, *16*, 1226.
37. Morales Helguera, A.; Perez Gonzalez, M.; Cordeiro, M. N. D. S.; Cabrera Perez, M. A. *Chem. Res. Toxicol.* **2008**, *21*, 633.
38. Helguera, A. M.; Gonzalez, M. P.; Cordeiro, M. N. D. S.; Perez, M. A. *Toxicol. Appl. Pharmacol.* **2007**, *221*, 189.
39. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 844.
40. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 320.
41. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 23.
42. O'Neill, M. J.; Heckelman, P. E.; Koch, C. B.; Roman, K. J. *The Merck Index, An Encyclopedia of Chemicals, Drugs and Biologicals*; Merck & Co., Inc.: New Jersey: Whitehouse Station, NJ, 2006.
43. Estrada, E.; Gutiérrez, Y. MODESLAB, v1.5; Santiago de Compostela, 2002–2004.
44. van de Waterbeemd, H. *Chemometrics methods in molecular design*; VCH Publishers: Weinheim, New York, Basel, Cambridge, Tokyo, 1995.
45. Garcia, I.; Fall, Y.; Gomez, G.; Gonzalez-Diaz, H. *Mol. Divers.* **2011**, *15*, 561.
46. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. *Bioorg. Med. Chem.* **2011**, *19*, 6239.
47. Munteanu, C. R.; Gonzalez-Diaz, H.; Borges, F.; de Magalhaes, A. L. *J. Theor. Biol.* **2008**, *254*, 775.
48. StatSoft. STATISTICA. Data analysis software system, v6.0; Tulsa, 2001.
49. Huberty, C. J.; Olejnik, S. *Applied MANOVA and discriminant analysis*; John Wiley & Sons, Inc.: Hoboken, New Jersey, 2006.
50. González-Díaz, H.; Pérez-Bello, A.; Cruz-Monteagudo, M.; González-Díaz, Y.; Santana, L.; Uriarte, E. *Chemometr. Intell. Lab. Syst.* **2007**, *85*, 20.
51. Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E. R. *Bioinformatics* **2010**, *26*, 822.