http://dx.doi.org/10.5935/0103-5053.20150172

J. Braz. Chem. Soc., Vol. 26, No. 9, 1936-1946, 2015.
Printed in Brazil - ©2015 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

ALK-5 Inhibition: A Molecular Interpretation of the Main Physicochemical Properties Related to Bioactive Ligands

Sheila C. Araujo,[a] Vinicius G. Maltarollo,[a] Danielle C. Silva,[a] Jadson C. Gertrudes[b] and Kathia M. Honorio,[a,b]*

[a]*Centro de Ciências Naturais e Humanas, Universidade Federal do ABC,*
R. Santa Adélia 166, 09210-170 Santo André-SP, Brazil

[b]*Escola de Artes, Ciências e Humanidades, Universidade de São Paulo,*
Av. Arlindo Bettio 1000, 03828-000 São Paulo-SP, Brazil

Activin-like kinase 5 (ALK-5) receptor represents an attractive object to treat cancer. Analyses on the quantitative structure-activity relationship were performed to explore the relationship between the molecular structure of 1,5-naphthyridine, pyrazole and quinazoline derivatives and the inhibition of the activin-like kinase 5. From a data set containing 59 compounds, various electronic descriptors were calculated using density functional theory (DFT) method; stereochemical descriptors (as molecular volume and area), polar surface area (PSA), log P and dragon descriptors were also calculated. The ordered predictor selection (OPS) algorithm, weighted principal component analysis (PCA) and Fisher's weights (FW), combined with sequential forward selection, were employed to select the most relevant descriptors to be employed in all partial least square regressions. Using this procedure, we selected the most informative descriptors and significant correlation coefficients were achieved ($r^2 = 0.74$, $q^2 = 0.83$). Additional validation tests were carried out, indicating that the obtained model is robust and reliable and, consequently, it can be used to predict the biological activity of new compounds.

Keywords: ALK-5, cancer, molecular modeling, feature selection, QSAR

Introduction

Cancer is a global problem and is cause of death in all countries. It is estimate that the number of cancer cases will increase worldwide due to the growth and aging of the population, particularly in less developed countries, in which about 82% of the world's population resides. In Brazil, an estimate performed by National Institute of Cancer in Brazil (INCA) for 2014, also valid for 2015, predicted an increase of 75% in new events of cancer.[1,2] Already, Global Cancer Statistics indicated that in 2014, about 580,350 Americans were expected to die of cancer, almost 1,600 people *per* day.[1-4] Cancer is the second most common cause of death in the United States, exceeded only by heart diseases, accounting for nearly one of every four deaths. There are many cases of cancer in population and the mortality level is expected to rise globally.[3,4] The global estimates are very concern because cancer is generally caused by genetic mutations, which provide some specific characteristics to the affected cell such as, high levels of proliferation, including neighboring tissues (metastasis) and evasion to apoptosis. Thus, it is extremely important to find out new drug candidates that target the cancer progression, invasion and metastasis.[1,2]

In this scenario, there is an interesting target protein known as transforming growth factor β (TGF-β).[3] The role of TGF-β in the cancer biology was described in the literature recently.[5-8] The complex function of TGF-β depends on the activation of two highly conserved single *trans*-membrane serine/threonine kinases: type I (TβRI or ALK-5 activin-like kinase 5) and type II receptors (TβRII). The mechanism related to the TGF-β binding involves the following steps: TβRII phosphorylates the threonine residues in the GS (repeated series of serine-glycine) domain of the ligand-occupied ALK-5 (or TβRI). The ALK-5 receptor, on the other hand, phosphorylates the cytoplasmic proteins SMAD2 and SMAD3 at two carboxyl terminals of serine residues. The phosphorylated SMAD proteins form heteromeric complexes with SMAD4; this complex translocates inside the nucleus to affect the gene transcription. It is known that changes in the DNA expression are important to evolution and adaptation of living organisms. Thus, TGF-β and its

*e-mail: kmhonorio@usp.br

receptors (ALK-5 and TβRII) are able to control the cellular growth and to promote several biological responses. In summary, these receptors can be considered as important targets to treat complex diseases such as cancer and fibrosis. Furthermore, considering the incidence of a large number of side effects in the cancer treatment, the discovery of new small-molecule inhibitors against the kinase activity of the ALK-5 receptor represents an attractive way to the combat of cancer.[6,7,9-12] Some compounds targeting ALK-5 receptor are in the preclinical evaluation, such as LY364947/HTS-466284 (4-[3-(2-pyridinyl)-1*H*-pyrazol-4-yl]-quinoline)[13] and LY2157299 (4-[2-(6-methyl-pyridin-2-yl)-5,6-dihydro-4*H*-pyrrolo[1,2-b]pyrazol-3-yl]-quinoline-6-carboxylic acid amide),[14] which are developed by the company Eli Lilly, perform the main interactions in the important hinge region for the inhibitory activity.[15-17]

In vitro assays on the activity of ALK5 inhibitors remain an intensive labor and time consuming operation. In this context, more efficient and economical alternative methods should be employed, such as *in silico* molecular modeling approaches, which are used in virtual screenings to predict and prioritize chemicals for subsequent *in vitro* and *in vivo* screenings.

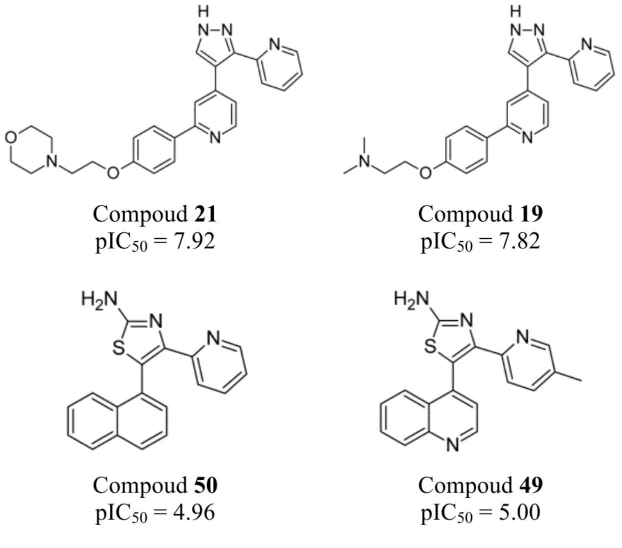
Quantitative Structure-Activity Relationship (QSAR) studies have been widely used to help in predicting and designing new bioactive compounds. In this way, QSAR methodology was employed in this study to explain how the molecular properties of a compound series are associated to biological activity. So, QSAR models can be used to understand the possible mechanisms of interaction between ligands and receptors, as well as helping the development of new lead-like drugs.[18-20] There are some QSAR models of ALK-5 inhibitors such as benzimidazoles,[21] 4-(quinolone-4-yl)-substituted, 1,5-naphthyridine, pyrazole and quinazoline derivative series reported on literature.[22,23] These models were statistically validated and showed common physicochemical features, for example, the importance of interaction with HIS283 at the hinge region.[21-23] However, these previous studies did not take into account all compound classes analyzed here and the authors employed other QSAR techniques. This study presents a different point of view on the main interactions that can be occurring between the compound classes selected and the biological target (ALK-5).

QSAR studies, along with the extracted information from the available X-ray crystallographic structure of ALK-5, have shown to be useful tools in the lead compound optimization in order to obtain potential therapeutic agents for the treatment of cancer and to understand the role of ALK-5 in the pathology of this disease. For this, our study constructed a series of models in order to elucidate the most relevant relationship between the molecular properties of the ALK-5 inhibitors studied in this work and their biological activity. Another objective of this study is evaluating the ability of various methodologies used for an efficient variable selection and, consequently, constructing a statistical model independent of the molecular alignment aiming the construction of a simple, effective and innovative model that could be employed in further virtual screening protocols.

Experimental

Data set

It was selected a dataset of 59 compounds, synthesized and tested at the same experimental conditions by Gellibert *et al.,*[24-26] to construct robust and reliable statistical models. The biological activity (IC_{50} values) of all compounds was tested by using a transcriptional assay in HepG2 (hepatocellular cells).[24-26] These compounds comprised three different classes of diverse structures: 1,5-naphthyridine, pyrazole and quinazoline derivatives, whose IC_{50} values were converted in pIC_{50} ($-logIC_{50}$, see Figure 1 and Table S1).


Compoud **21**
$pIC_{50} = 7.92$

Compoud **19**
$pIC_{50} = 7.82$

Compoud **50**
$pIC_{50} = 4.96$

Compoud **49**
$pIC_{50} = 5.00$

Figure 1. The most and the least active compounds of the data set.

Generation of 3D structures

In order to generate a bioactive conformation of all compounds, we performed several docking analyses employing the GOLD 5.0 software,[27] which uses genetic algorithm to generate the ligand conformation and GOLDScore as scoring function. All steps and details of the docking protocol, as well as the pose generation were shown in previous study.[23] The good quality of the selected poses/conformations can be noted from the analyses of all statistical parameters for the 3D model, which is highly affected by the tridimensional alignment of the data set.[28] In addition, the docking analyses of the most and the least active compounds corroborate the binding mode proposed in the literature. Therefore, the docking poses could be considered a good model for the bioactive conformations of the studied ALK-5 inhibitors.

Physicochemical properties and calculation of descriptors

After the generation of all 3D conformations from the docking analyses, it was calculated several electronic properties (for example, molecular orbital energies, dipole moment and atomic charges), as well as other descriptors obtained from density functional theory (DFT) method with B3LYP functional[24] and 6-311g(d) basis set,[29,30] implemented in Gaussian09 package.[27] Stereochemical descriptors (such as molecular volume and area), polar surface area (PSA), log P, molecular weight and others were calculated using the software Spartan'08,[31] HyperChem 8.1[32] and Sybyl 8.1.[33] Topological descriptors were calculated employing E-dragon 2.1 available at Virtual Computational Chemistry Laboratory (VCCLAB),[34] which are considered valuable information about several aspects of the molecular structure.[35,36]

Feature selection

The selection of features that mathematically represent the compound set and the relationships with the biological activity are not a trivial task. The methods employed to generate the molecular descriptors are able to provide a large number of variables (some of them may have up to thousands of descriptors). Furthermore, there is an ideal condition in QSAR studies: in general, each descriptor can explain five chemical compounds.[18,36] This proportion (one descriptor for each five compounds), called parsimony principle or Occam's razor, was proposed to facilitate the physicochemical interpretation of QSAR model and also to avoid the overfitting in QSAR modeling (a condition when the excess of information improves randomly the quality of the model).[37] For this reason, we tested the ability of various methods used for an efficient variable selection with the aim to select only chemical descriptors able to generate a robust model. All methods employed in this study for the variable selection will be described below.

Fisher's weight (FW)

Fisher's weight (FW) is a very used method in pattern recognition studies. It selects the variables that characterize or separate in two or more groups a given data set.[38,39] The main idea of FW is finding a subset of variables such that on the data space generated by the selected variables, the distances between observations in different classes are as large as possible, while the distances between the observations in the same class are minimal as possible. The variable selection occurs by maximizing the trace criterion, an optimization function that can be applied to several methods of dimensionality reduction because it directly holds the distance between the observations within or between the classes of data. To simplify this problem, the most used heuristic is computing a weight for each variable of the set $X(x^j, j = 1,\ldots,n)$ according to the criterion F.[40] Considering μ_k^j the average of the k_{th} class corresponding j_{th} variable and that σ_j the standard deviation of the j_{th} variable, it is computed the weight of each variable by equation 1:

$$F\left(x^j\right) = \frac{\sum_{i=1}^{k} t_i \left(u_k^j - u^j\right)}{\left(\sigma_j\right)^2} \qquad (1)$$

After the calculation of the Fisher's weight for each variable, variables with the highest weight are selected.

Ordered predictor selection (OPS)

Ordered predictor selection (OPS) is an algorithm employed to select the most relevant descriptors that will be employed in regression analyses.[41] The OPS method generates a vector (informative vector) that contains information about the location of the best chemical descriptors for prediction. The vectors can be directly obtained from calculations performed with information about responses and dependent variables or combinations of different vectors obtained with the same purpose. Afterwards, the original variables are differentiated according to the corresponding absolute values of the informative vector obtained in the previous step. The higher the absolute value, more important is the response variable, which enables its sorting in descending order of magnitude.[42] The multivariate regression models are built and evaluated using a cross validation strategy. An initial subset of variables (window) is selected to build and evaluate the model. Then, this matrix is expanded by the addition of a fixed number of variables (increment) and a new model is built and evaluated. New increments are added until all or some percentage of variables are taken into account. Quality parameters of the models are obtained for every evaluation and stored for a future comparison. The evaluated variable sets (initial window and its extensions) are compared using the quality parameters calculated during the validations. The model with the best quality parameters should contain the variables with the best predictive capability and so these will be the selected variables.

Weighted principal components analysis (WPCA)

Weighted principal components analysis (WPCA) is a method that uses the matrix of loadings obtained by PCA technique to perform the variable selection.[43] In WPCA, there are combinations of the weighted principal components with a threshold algorithm. Specifically, the contribution of each feature is represented by a loading value in a weighted principal component, and a threshold algorithm based on a moving range-based control chart evaluates the significance of its contribution.[43]

In WPCA, the weight of each variable is obtained from the sum of the loading values that represent the importance of each feature in the formation of a PC (for example, a_{ij} indicates the degree of importance of j_{th} feature for the i_{th} PC). For the case where a loading value of the j_{th} original feature is initially computed m PC's, the importance of the j_{th} feature can be represented by equation 2, where a_{ij} (i,j = 1,2,…,n) represents the loading values of each variable in each PC after the application of PCA[44] and b_i represents the weight of the i_{th} PC. A way to determine b_i is computing the total variance explained by the i_{th} PC; w_j is called a weighted PC loading for the feature j.

$$w_j = \sum_{i=1}^{m} \left|a_{ij}\right| b_i, = j = 1, \ldots, n \qquad (2)$$

After obtaining the weighted PC's, it is performed the moving range-based threshold algorithm as a way to identify the significant features from the weighted PC loadings. The threshold algorithm comes from a moving average control chart widely used in quality control.[41] A feature is considered as significant if the corresponding weighted PC loading exceeds the threshold γ.

Sequential forward selection (SFS)

Sequential forward selection (SFS) is another method used for variable selection, which selects a subset of variables that have the best result in the generation of a regression or classification model. This search is carried out as following: (*i*) the algorithm starts its execution looking for a single variable that generates a regression model that satisfies a certain value (i.e., low calibration error), (*ii*) after, these new variables are sequentially grouped to the initial selected variable, since the value obtained will be better than the value obtained from the previous subset, or until a certain number of variables is reached. More information about this method can be found in other studies.[45] All described methods (OPS, WPCA and FW) were employed in combination to sequential forward selection to achieve a defined final number of descriptors that better describe our system.

Splitting of training and test sets

Training and test sets are important to determine the quality of the statistical models obtained from regression methods. The composition of training and test sets is important to obtain an internally consistent model and to test its external ability of prediction using an equally representative set. Kennard-Stone is a rational method that is very employed to split training and test sets.[46,47] This method was developed to produce a division when no standard experimental design can be applied.[44] The Kennard-Stone algorithm selects the objects so that they are divided evenly throughout the descriptor space of the original data set. This technique is applied as follows: (*i*) initially, select the first two molecules of the dataset are selected by choosing the two ones that are farthest apart in terms of Euclidean distance; (*ii*) to select the compound that has the maximum dissimilarity from each one of the previously selected molecules and place this molecule in the training set; (*iii*) to repeat the step (*ii*) until the desired number of molecules has been added to the training set.

Outlier detection and applicability domain

Other two important aspects that should be checked in the generation of QSAR models are the outlier detection and the analysis of the applicability domain. These two properties are robustness measures of QSAR models that will be used for predicting compounds with unknown activity. In this study, for outlier detection, it was applied a method proposed by Filzmoser *et al.*[48] that combines the ordered squared robust Mahalanobis distances (MD) of the observations and the distribution of chi-squared. Initially, the MD values for each observation are calculated. Afterwards, to perform the search for outliers, observations that exceed a certain value of the chi-squared distribution are marked. More details about this method can be found in Filzmoser *et al*.[48]

The applicability domain is widely used to express the scope and limitations of a QSAR model, i.e., the range of chemical structures for which the model is considered to be applicable.[49] In our study, we used the leverage value and Studentized residuals to determine the applicability domain of the compounds. The leverage method provides a distance measure of the compounds from the centroid of the data set (i.e., vector mean of the dataset). Compounds near to centroid are less influential in QSAR model than that in extreme points. More details about these techniques can be found in references.[50,51]

Construction of QSAR models

The generation of QSAR models was performed using Partial Least Squares (PLS) method, implemented in Pirouette3.11 software.[52] The PLS method can handle data with numerous independent variables by constructing principal components (PCs) from a non-linear combination of all X variables used to construct the QSAR model. A short description of PLS technique involves the following idea: the X matrix of independent variables (containing the descriptors) is correlated with the Y vector (representing the biological data, in this case) in such a way that the projected coordinates (T) are good predictors of Y.[53] An important feature of PLS is the fact of the biological data is included in the decomposition procedure. Besides, the loading matrix (W) is defined in such a way that the product (variance in X) times (the correlation XW to Y) is maximized.[53] A detailed description of PLS can be found in other references.[54,55] The quality model was evaluated according to its internal consistency (q^2, values of leave-one-out and leave-N-out methods), external predictive ability (r^2 of the test set and residual values), sensitivity of randomization (Y-scrambling) and external predictive ability potential (r^2_m).

Results and Discussion

In order to define the best model, there was specified a flow chart as shown in Figure 2. Initially, from 1719 calculated descriptors, we applied an intermediary filter using WPCA, OPS and two forms of Fisher's weight. After the application of these techniques, the SFS algorithm was used aiming to achieve models with 8 variables, according to the rule of 1 descriptor for each 5 compounds, since the training set of our study contains 46 compounds. To carry out the selection of variables with WPCA, it was used the software MATLAB.[56] As parameters to WPCA, we applied the error range equals a 0.01 ($\beta = 0.01$). The number of variables obtained with this method was equal to 42. To perform the variable selection with OPS, it was used the package OPS developed by Teofilo *et al.*,[42] also implemented in MATLAB software. As parameters to the variable selection, it was employed the minimal value of root mean squared error, obtained after the application of PLS technique. From this procedure, 256 variables were selected.

The initial version of FW was applied separating the molecules in two classes of biological activity: (*i*) a class with the biological range between 4.95 and 7.32; (*ii*) a class with the range between 7.33 and 7.92. The choice for the splitting of the dataset using a non-uniform distribution of biological ranges is due to this threshold (pIC_{50} ca. 7.33) separates the compounds in two balanced subsets, biologically and structurally, with about 23 compounds each. Finally, the weight higher than 5.00 were selected, resulting in 357 variables selected with this methodology.

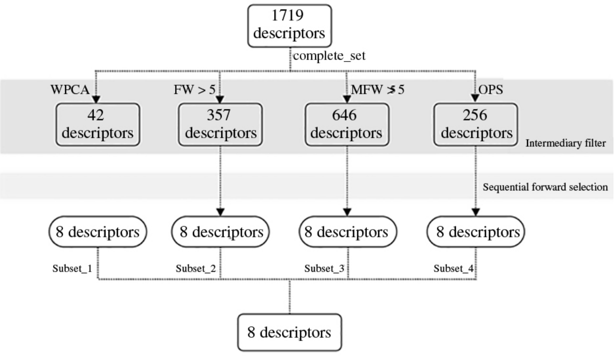


Figure 2. Scheme used to select chemical descriptors.

To the application of the second version of Fisher's weight (MFW), initially, the dataset was divided in six classes, according to the following ranges of biological activity: (*i*) class 1 (4.95-5.49); (*ii*) class 2 (5.50-5.99); (*iii*) class 3 (6.00-6.49); class 4 (6.50-6.99); class 5 (7.00-7.49) and class 6 (7.50-7.99). The main idea of MFW is to select the descriptors that are important to discriminate between the most active compounds and the least ones. In other words, MFW is designed to discriminate the most active compounds (class 6) in each other class individually.

After the definition of the classes, various comparisons between the most activity class (class 6) and the five remaining classes were carried out using the FW and the weights for each comparison were determined. Finally, for each variable, we calculated the sum of the weights found in each comparison from equation 3.

$$MFW = 0.35FW_{6-1} + 0.30FW_{6-2} + 0.20FW_{6-3} + 0.10FW_{6-4} + 0.05FW_{6-5} \quad (3)$$

In the last step employed in the application of MFW, we selected the variables with weight higher than 5.0, as done in the selection with FW. The major difference between FW and MFW is that the initial method can provide the variables related to the split of the data set in two classes (the most and the least potent compounds) and the second one provides the X variables that discriminate gradually the most active compounds (class 6) from the least active class.

Table 1. Results of PLS regression combined with SFS technique

Model	Feature selection method	q^{2a}	SEV[b]	r^{2c}	SEC[d]	PCs[e]
1	WPCA[f]	0.73	0.44	0.81	0.40	6
2	FW[g]	0.80	0.38	0.85	0.35	5
3	MFW[h]	0.80	0.38	0.86	0.34	6
4	OPS[i]	0.73	0.45	0.84	0.37	6

[a]q^2: validation coefficient; [b]SEV: standard error of validation; [c]r^2: calibration coefficient; [d]SEC: standard error of calibration; [e]PCs: number of principal components; [f]weighted principal components analysis (WPCA); [g]Fisher's weight (FW); [h]second version of Fisher's weight (MFW); [i]ordered predictor selection (OPS).

The application of MFW returned 646 variables. After the initial step of the variable selection, applying WPCA, FW, MFW and OPS, it was used the SFS technique to select eight variables from each subset of the variables cited previously, as shown in Figure 2. The SFS technique was combined with the PLS method and eight variables were selected, which resulted in a best value of q^2. The main results are summarized in Table 1.

From Table 1, we selected the variables indicated by the MFW method, since these variables returned the best values of q^2 and the lowest value of standard error of estimation. The difference between the models generated with MFW and FW methods is not significant, then we employed the MFW model to perform a physicochemical interpretation of the selected variables but we also analyzed the other models.

After choosing the best set of variables using MFW, we performed several analyses of outliers and also different splitting of training and test sets. For the analysis of outliers, the technique described by Filzmoser *et al.*[48] was applied, making the search for outliers in a chi-squared (Figure 3) distribution with limit value equals to 0.95. Moreover, the values of leverage obtained after the variable selection were calculated. Among all compounds, it was observed that the compound **3** was identified as an outlier by the Filzmoser's technique, as well as the coefficient of leverage. Thus, this compound was removed from the data set in the further analysis.

The splitting of training and test sets was performed in two steps: (*i*) the data was divided in two subsets according to the levels of biological activity: 4.95-7.32 and 7:33-7.92; (*ii*) after this initial splitting, the Kennard-Stone method was applied in each subset, separating 80% for the training set and 20% for the test set. As a final result, 46 molecules were selected for the training and 12 for the test set (Supplementary Material, Table S2).

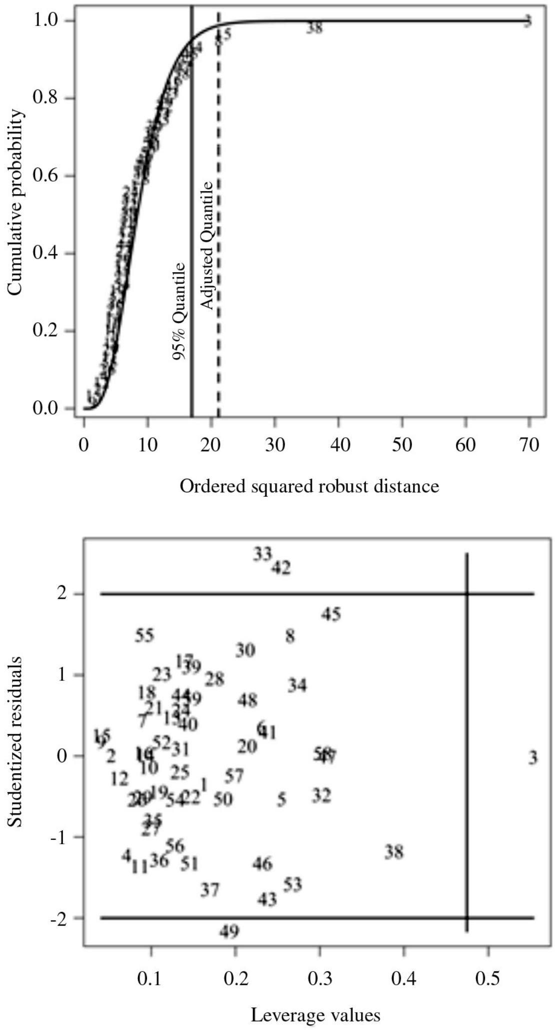


Figure 3. (a) Analysis of outliers and (b) plot of leverage *versus* Studentized residuals.

Statistical analysis of model 3

In comparison with the other models, the model 3 displays satisfactory internal and external correlation coefficients (q^2_{LOO} and q^2_{LNO} = 0.74; r^2 = 0.83 and $r^2_{test\ set}$ = 0.87) and the Y-scrambling results (the average values of q^2 and r^2 for the scrambled models) indicate that the model was not obtained by chance (Table 2). Finally, the best quality of the model 3 can be observed by comparison of r^2_m of all models. Only the MFW and OPS models (models 3 and 4, respectively) showed acceptable external predictive ability, but clearly the external predictive ability of the model 3 was strongly superior to the model 4. However, the model obtained with the combination of MFW and SFS presented the lowest SEV and SEC values.

Table 2. Others statistical parameters for all obtained models

	WPCA[a]	FW[b]	MFW[c]	OPS[d]
q^2_{LOO}[e]	0.66	0.74	0.74	0.74
SEV[f]	0.46	0.40	0.40	0.40
q^2_{LNO}[e]	0.65	0.74	0.74	0.73
r^2[g]	0.77	0.83	0.83	0.84
SEC[h]	0.40	0.35	0.35	0.34
$r^2_{test\ set}$[g]	0.67	0.92	0.87	0.70
r^2_m[g]	0.42	0.47	0.57	0.54
PCs[i]	6	5	6	6
$q^2_{Y\text{-}scrambling}$[e]	0.17	0.18	0.18	0.18
$r^2_{Y\text{-}scrambling}$[g]	−0.34	−0.33	−0.31	−0.33

[a]Weighted principal components analysis (WPCA); [b]Fisher's weight (FW); [c]ordered predictor selection (OPS); [d]second version of Fisher's weight (MFW); [e]q^2: validation coefficient; [f]SEV: standard error of validation; [g]r^2: calibration coefficient; [h]SEC: standard error of calibration; [i]PCs: number of principal components.

To evaluate the robustness and the stability of the selected model, leave-N-out and y-scrambling tests were carried out (Table 2). In fact, a good QSAR model must have an average value of q^2 close to the q^2 obtained with the leave-one-out procedure, while the standard deviation for each N should not exceed 0.1.[51] The model obtained with the variable selection using MFW and SFS was stable with deviations from q^2 for each N being lower than 0.020. These findings confirm the stability and robustness of the model 3 (Figure 4).

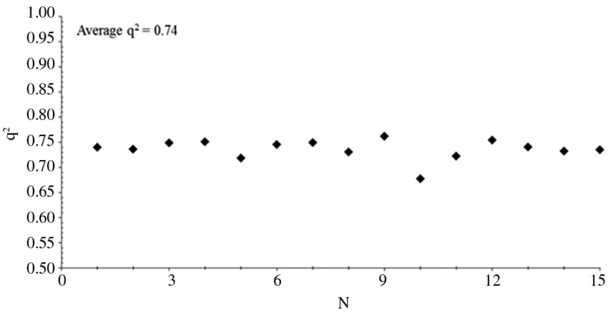


Figure 4. Plot of the results obtained for the leave-N-out validation.

The predictive power of the model 3 was also evaluated by predicting the biological activity of the compounds from the test set (external validation). Experimental and predicted pIC_{50} values are listed in Table 3. The obtained results indicate that the obtained model is very predictive since the residual values of external predictions were lower than 0.80 log unities.

Table 3. Experimental and predicted pIC_{50} values for the test set compounds

Compound	pIC_{50} experimental	pIC_{50} predicted	Residual
1	6.97	7.04	−0.07
4	6.72	7.14	−0.42
7	7.55	7.41	0.14
14	6.97	6.94	0.03
21	7.92	7.72	0.20
23	7.72	7.36	0.36
31	7.31	7.30	0.01
38	7.03	7.56	−0.53
43	5.40	6.19	−0.79
44	5.93	5.76	0.17
46	5.00	5.63	−0.63
47	5.75	5.79	−0.04

A plot of the experimental *versus* predicted pIC_{50} for the compounds in training and test sets is shown in Figure 5. The good agreement between the experimental and calculated values indicates that a predictive MFW model was obtained and can be used to accurately predict the biological activity of other compounds within this structural class.

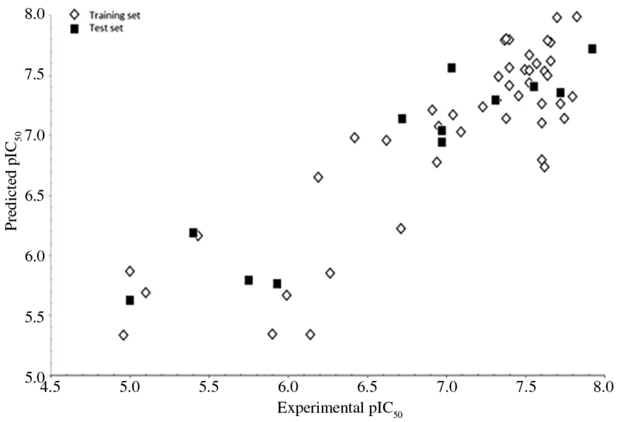


Figure 5. Experimental *versus* predicted pIC_{50} of the training and test set compounds.

The y-scrambling validation was also employed to verify the possibility of chance correlations between the dependent variable and the selected descriptors. In this study, the pIC_{50} values were scrambled and the r^2 and q^2 values were calculated (Figure 6). In the 100 y-scrambling experiments performed in our data, only low values of r^2

and q^2 were obtained, with average of −0.31 and 0.18, respectively. If low values were found for both parameters, then one can be sure that a true correlation between the selected descriptors and the response variable exists in our data set.

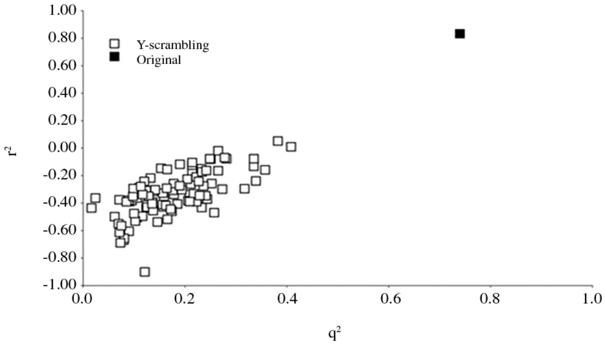


Figure 6. Plot of the results obtained in the y-scrambling tests.

In summary, all internal and external validations indicate that the model 3 is suitable for the prediction of the biological activity of new ALK-5 inhibitors and, consequently, this model contain statistically relevant information in the relationships between the calculated descriptors and the biological activity.

Physicochemical interpretation of the best model

For the model obtained using the MFW and SFS algorithms (variable selection), 8 descriptors were selected: MATS4v, EEig04x, ESpm12r, BELp5, SPH, Mor26e, R8m+ and R5e+. Table 4 displays the description of each variable employed in the construction of the model 3.

Table 4. Symbols, types and definitions of the selected descriptors

Descriptor	Type	Definition
SPH	geometrical	spherosity[57]
EEig04x	edge adjacency indices	eigenvalue 04 from the edge-adjacency matrix weighted by edge degrees[58]
ESpm12r		spectral moment 12 from edge adjacent matrix weighted by resonance integrals[59]
MATS4v	2D autocorrelation	Moran autocorrelation, lag 4 weighted by atomic van der Waals volumes[60]
Mor26e	3D-MoRSE descriptor	3D-MoRSE, signal 26 weighted by atomic Sanderson electronegativity[61]
R8m+	GETAWAY descriptor	R maximal autocorrelation of lag 8 weighted by atomic masses[62]
R5e+		R maximal autocorrelation of lag 5 weighted by atomic Sanderson electronegativities[63]
BELp5	Burden eigenvalue descriptor	lowest eigenvalue n = 5 of Burden matrix / weighted by atomic polarizabilities[64]

The calculated values for the 8 selected descriptors are shown in Supplementary Information (Table S2) and the contributions of each descriptor to the regression vector, in the model 3, are displayed in Figure 7.

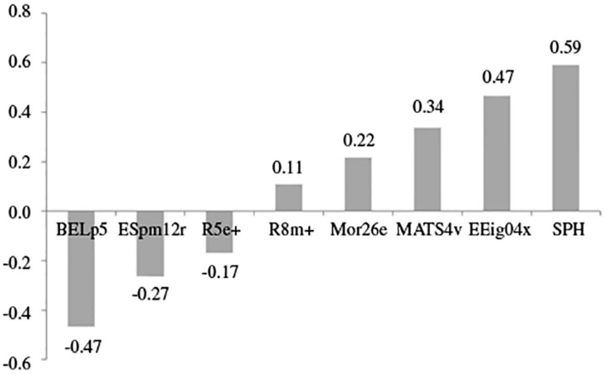


Figure 7. Contribution of all selected descriptors.

Regarding the selected descriptors used to build the model presented in this study, some considerations can be pointed out (Figure 7):

(*i*) SPH is a geometrical descriptor and refers to the spherical format of the molecule. This variable suggests that the spherical shape of the compounds is an important parameter in the ALK-5 inhibition since this descriptor showed the highest contribution to PC. Compounds with values of SPH nearest to 1 indicate higher spherical shape while values nearest to 0 indicate compounds not spherical.[57] In this study, the SPH descriptor presented important contribution (Figure 7) indicating a better complementarity between spherical compounds and the active site. Indeed, the three more potent compounds (**21**, **19** and **39**) have values of SPH equal to 0.937, 0.949 and 0.853, respectively, while the three least potent ones (**50**, **46** and **49**) have SPH values equal to 0.732, 0.783 and 0.802, respectively. These results indicate that the most potent compounds have higher values of SPH and, consequently, they are more spherical and can be performed more interactions in the active site of the biological target.

(*ii*) EEig04x is the second descriptor with high positive contribution and represents the eigenvalue 04 from the edge-adjacency matrix weighted by edge degrees, which belongs to edge-adjacency indices. The adjacency matrix also provides some generalized descriptors of network connectivity like the average vertex degree and connectivity.[58,59]

(*iii*) MATS4v represents the distribution mode of the atomic van der Waals volumes along the topological structure of the compounds.[60,65] Therefore, the positive contribution of this descriptor indicates the relationship between the topological structure weighted by van der Waals volume and the biological activity.

(*iv*) MorSE descriptors have structural information by means of 3D atomic coordinates. In this case, the Mor26e descriptor represents a 3D-MorSE descriptor weighted by atomic electronegativity and this descriptor has the fourth positive contribution. Thus, the atomic electronegativity of the compounds showed high statistical importance for the protein-ligand interaction.[35,61]

(*v*) R8m+ and R5e+ are GETAWAY descriptors that mean geometry, topology and atom-weights assembly descriptors derived from the leverage matrix, which is deduced by the centering of all atomic coordinates.[62,63] Thereby, R8m+ is weighted by atomic masses with a positive contribution for the dataset and R5e+ is weighted by atomic electronegativities with negative contribution (see Figure 7). Therefore, these descriptors can contribute for the size (R8m+) and the shape (R5e+) of the ALK-5 inhibitor weighted by the properties of the data set from pIC_{50} values.

(*vi*) ESpm12r represents the resonance effects or resonance integrals between atoms twelve bonds apart.[59,66] As its contribution to the model was negative, the resonance effects could inversely be related to the biological activity.

(*vii*) BELp5 is a 2D Burden eigenvalue descriptor that has the lowest contribution (Figure 7). This descriptor is weighted by the atomic polarizabilities, encoding molecular branching, position and length. This topological descriptor is designed to encode atomic properties that drive intramolecular interactions.[64,67]

Based on the results obtained in this study and in the face of the continuous search for new anti-cancer compounds, statistical models can play an important role in the discovery and optimization of new drug candidates. In this work, WPCA, FW, MFW and OPS-PLS models were developed to provide insights on relevant molecular features for the ALK-5 inhibition. A set of 8 descriptors selected by MFW and SFS techniques has demonstrated to be suitable for the construction of reliable models. The good statistical parameters, stability and robustness of the models obtained here, as assured by the validation tests applied over our data, indicate that these models can be used to design other inhibitors with improved anti-cancer activity, i.e., using this model as virtual screening filter. Therefore, the selected descriptors could be employed to construct focused chemical libraries to find out new ALK-5 inhibitors.

Conclusions

In this study, four models were investigated with the aim to describe the relationships between the chemical structure of a series containing bioactive ligands and the ALK-5 receptor. WPCA, FW, MFW and OPS-PLS algorithms were employed to select the most relevant descriptors. MFW was the best algorithm for the variable selection, because it resulted in significant correlation coefficients ($q^2 = 0.83$, $r^2 = 0.74$ and $r^2_{\text{Test}} = 0.87$). The strategy employed in this work has provided a reliable model for the ALK-5 inhibition regarding the class of the studied ligands. Our findings suggest the importance of topological, geometrical, edge adjacency indices, 2D autocorrelation and 3D features for the anti-cancer activity presented by the studied compounds. The descriptors selected using the MFW method describe molecular features as the geometry (SPH) and connectivity (EEig04x), which are defined as dragon descriptors. Additionally, the influence of the distribution mode of atomic van der Waals volume (MATS4v) is indicated by 2D autocorrelations descriptor, as well as Mor26e and atomic electronegativity. Therefore, these results can be used to design other ALK5 inhibitors with anti-cancer activity.

Supplementary Information

Supplementary information (structure and pIC_{50} values of the studied compounds, as well as the values of the selected descriptors for the training and test sets) are available free of charge at http://jbcs.sbq.org.br as PDF file.

Acknowledgements

The authors would like to thank L'Oreal, ABC (Brazilian Science Academy), UNESCO (United Nations Educational, Scientific and Cultural Organization), FAPESP, CNPq, CAPES and UFABC for funding.

References

1. Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA); http://www.inca.gov.br/estimativa/2014/estimativa-24042014.pdf accessed in June 2015.
2. Siegel, R. L.; Miller, K. D.; Jemal, A.; *CA-Cancer J. Clin.* **2015**, *65*, 5.

3. Torre, L. A.; Bray, F.; Siegel, R. L.; Ferlay, J.; Lortet-Tieulent, J.; Jemal, A.; *CA-Cancer J. Clin.* **2015**, *65*, 87.
4. Siegel, R.; Ma, J.; Zou, Z.; Jemal, A.; *CA-Cancer J. Clin.* **2014**, *64*, 9.
5. Akhurst, R. J.; Hata, A.; *Nat. Rev. Drug Discovery* **2012**, *11*, 790.
6. Connolly, E. C.; Freimuth, J.; Akhurst, R. J.; *Int. J. Biol. Sci.* **2012**, *8*, 964.
7. Wang, H.; Sessions, R. B.; Prime, S. S.; Shoemark, D. K.; Allen, S. J.; Hong, W.; Narayanan, S.; Paterson, I. C.; *J. Comput. Aided Mol. Des.* **2013**, *27*, 365.
8. Arjaans, M.; Munnink, T. H. O.; Timmer-Bosscha, H.; Reiss, M.; Walenkamp, A. M. E.; Lub-de Hooge, M. N.; de Vries, E. G. E.; Schroder, C. P.; *Pharmacol. Ther.* **2012**, *135*, 123.
9. Bierie, B.; Moses, H. L.; *Cytokine Growth F. R.* **2006**, *17*, 29.
10. Goldberg, F. W.; Ward, R. A.; Powell, S. J.; Debreczeni, J. E.; Norman, R. A.; Roberts, N. J.; Dishington, A. P.; Gingell, H. J.; Wickson, K. F.; Roberts, A. L.; *J. Med. Chem.* **2009**, *52*, 7901.
11. Ciayadi, R.; Potdar, M.; Walton, K. L.; Harrison, C. A.; Kelso, G. F.; Harris, S. J.; Hearn, M. T. W.; *Bioorg. Med. Chem. Lett.* **2011**, *21*, 5642.
12. Ebrahimi, M.; Khayamian, T.; Gharaghani, S.; *J. Braz. Chem. Soc.* **2012**, *23*, 2043.
13. Scott Sawyer, J.; Beight, D. W.; Britt, K. S.; Anderson, B. D.; Campbell, R. M.; Goodson Jr, T.; Herron, D. K.; Li, H.-Y.; McMillen, W. T.; Mort, N.; Parsons, S.; Smith, E. C. R.; Wagner, J. R.; Yan, L.; Zhang, F.; Yingling, J. M.; *Bioorg. Med. Chem. Lett.* **2004**, *14*, 3581.
14. Zhou, L.; McMahon, C.; Bhagat, T.; Alencar, C.; Yu, Y.; Fazzari, M.; Sohal, D.; Heuck, C.; Gundabolu, K.; Ng, C.; Mo, Y.; Shen, W.; Wickrema, A.; Kong, G.; Friedman, E.; Sokol, L.; Mantzaris, G.; Pellagatti, A.; Boultwood, J.; Platanias, L. C.; Steidl, U.; Yan, L.; Yingling, J. M.; Lahn, M. M.; List, A.; Bitzer, M.; Verma, A.; *Cancer Res.* **2011**, *71*, 955.
15. Owens, J.; *Drug Discov. Today* **2001**, *6*, 1145.
16. Zhang, Y.; Feng, X.-H.; Wu, R.-Y.; *Lett. Nat.* **1996**, *383*, 168.
17. E. Ling, L.; Lee, W.-C.; *Curr. Pharm. Biotechnol.* **2011**, *12*, 2190.
18. Gaudio, A. C.; Zandonade, E.; *Quim. Nova* **2001**, *24*, 658.
19. Castro, F. M. M.; Alberto, M. C.; Coser, G. A.; *Quim. Nova* **2002**, *25*, 439.
20. Martins, J. P. A.; Ferreira, M. M.; *Quim. Nova* **2013**, *36*, 554.
21. Yang, M.; Zhou, L.; Zuo, Z. L.; Mancera, R.; Song, H.; Tang, X. Y.; Ma, X.; *QSAR Comb. Sci.* **2009**, *28*, 1300.
22. Geldenhuys, W. J.; Nakamura, H.; *Bioorg. Med. Chem. Lett.* **2010**, *20*, 1918.
23. Araujo, S. C.; Maltarollo, V. G.; Honorio, K. M.; *Eur. J. Pharm. Sci.* **2013**, *49*, 542.
24. Gellibert, F.; Woolven, J.; Fouchet, M.-H.; Mathews, N.; Goodland, H.; Lovegrove, V.; Laroze, A.; Nguyen, V.-L.; Sautet, S.; Wang, R.; *J. Med. Chem.* **2004**, *47*, 4494.
25. Gellibert, F.; de Gouville, A.-C.; Woolven, J.; Mathews, N.; Nguyen, V.-L.; Bertho-Ruault, C.; Patikis, A.; Grygielko, E. T.; Laping, N. J.; Huet, S.; *J. Med. Chem.* **2006**, *49*, 2210.
26. Gellibert, F.; Fouchet, M. H.; Nguyen, V. L.; Wang, R.; Krysa, G.; de Gouville, A. C.; Huet, S.; Dodic, N.; *Bioorg. Med. Chem. Lett.* **2009**, *19*, 2277.
27. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R.; *J. Mol. Biol.* **1997**, *267*, 727.
28. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W.; *J. Med. Chem.* **2007**, *50*, 726.
29. McLean, A.; Chandler, G.; *J. Chem. Phys.* **1980**, *72*, 5639.
30. Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A.; *J. Chem. Phys.* **1980**, *72*, 650.
31. *Spartan*, version 08; Wavefunction, Inc.: Irvine, CA, USA, 2008.
32. *HyperChem,* version 8.0; Hypercube, Inc.: Gainesville, FL, USA, 2007.
33. *Sybyl*, version 8.1; Tripos, Inc.: St. Louis, MO, USA, 2008.
34. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; *J. Comp. Aided Mol. Des.* **2005**, *19*, 453.
35. Todeschini, R.; Consonni, V.; *Handbook of Molecular Descriptor;* Wiley-VCH Verlag GmbH: Weinheim, Germany, 2000.
36. Arroio, A.; Honorio, K. M.; da Silva, A. B. F.; *Quim. Nova* **2010**, *33*, 694.
37. Hawkins, D. M.; *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1.
38. Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M.; Haussler, D.; *Bioinformatics* **2000**, *16*, 906.
39. Duda, R. O.; Hart, P. E.; Stork, D. G.; *Pattern Classification*; Wiley: Menlo Park, CA, 1997.
40. Guyon, I.; Elisseeff, A.; *J. Mac. Learn. Res.* **2003**, *3*, 1157.
41. de Melo, E. B.; *Eur. J. Med. Chem.* **2010**, *45*, 5817.
42. Teofilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C.; *J. Chemom.* **2009**, *23*, 32.
43. Kim, S. B.; Rattakorn, P.; *Expert Syst. Appl.* **2011**, *38*, 5704.
44. Jolliffe, I.; *Principal Component Analysis*, 2nd ed.; Springer: Berlin, 2002.
45. Aha, D. W.; Bankert, R. L.; *A Comparative Evaluation of Sequential Feature Selection Algorithms, Learning from Data;* Springer: New York, 1996.
46. Kennard, R. W.; Stone, L. A.; *Technometrics* **1969**, *11*, 137.
47. Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A.; *J. Chem. Inf. Model.* **2012**, *52*, 2570.
48. Filzmoser, P.; Garrett, R. G.; Reimann, C.; *Comput. Geosci.* **2005**, *31*, 579.

49. Roberts, D. W.; Patlewicz, G.; Kern, P. S.; Gerberick, F.; Kimber, I.; Dearman, R. J.; Ryan, C. A.; Basketter, D. A.; Aptula, A. O.; *Chem. Res. Toxicol.* **2007**, *20*, 1019.
50. Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. M.; Tong, W. D.; Veith, G.; Yang, C. H.; *Atla-Altern. Lab. Anim.* **2005**, *33*, 155.
51. Ferreira, M. M.; Antunes, A. M.; Melgo, M. S.; Volpe, P. L.; *Quim. Nova* **1999**, *22*, 724.
52. *Pirouette*, version 3.11; Infometrix Inc.: Bothel, USA, 2003.
53. Ferreira, M.; *J. Braz. Chem. Soc.* **2002**, *13*, 742.
54. Martens, H.; Naes, T.; *Multivariate Calibration*; John Wiley & Sons: Chichester, UK, 1992.
55. Geladi, P.; Kowalski, B. R.; *Anal. Chim.* **1986**, *185*, 17.
56. *Matlab*, version 7.8.0 (R2009a); The MathWorks Inc.: Natick, MA, USA, 2009.
57. Robinson, D. D.; Barlow, T. W.; Richards, W. G.; *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 939.
58. Azimi, G.; Afiuni-Zadeh, S.; Karami, A.; *J. Chemom.* **2012**, *26*, 135.
59. Bonchev, D.; Chemiker, B.; Chemist, B.; *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, UK, 1983.
60. Borota, A.; Mracec, M.; Gruia, A.; Rad-Curpăn, R.; Ostopovici-Halip, L.; Mracec, M.; *Eur. J. Med. Chem.* **2011**, *46*, 877.
61. Yan, D.; Jiang, X.; Yu, G.; Zhao, Z.; Bian, Y.; Wang, F.; *Chemosphere* **2006**, *63*, 744.
62. Deshpande, S.; Goodarzi, M.; Katti, S. B.; Prabhakar, Y. S.; *J. Chem.* **2013**, *2013*, 1.
63. Che, Z.; Zhang, S.; Shao, Y.; Fan, L.; Xu, H.; Yu, X.; Zhi, X.; Yao, X.; Zhang, R.; *J. Agric. Food Chem.* **2013**, *61*, 5696.
64. Kim, M.-G.; Shin, H.-S.; Kim, J.-H.; *Mol. Cell. Toxicol.* **2009**, *5*, 14.
65. Moran, P. A.; *Biometrika* **1950**, *37*, 17.
66. Cunha, L. B.; Freitas, H. F.; Castilho, M. S.; *J. Braz. Chem. Soc.* **2013**, *24*, 1623.
67. Kubinyi, H.; *QSAR in Drug Design, Theory Methods and Applications*; Springer: Netherlands, 1993.

Submitted: April 6, 2015
Published online: July 17, 2015

FAPESP has sponsored the publication of this article.