molecular informatics

DOI: 10.1002/minf.201000181

Digital Filters for Molecular Interaction Field Descriptors

Euzébio Guimarães Barbosa[a] and Márcia Miguel Castro Ferreira*[a]

Abstract: Descriptor properties are often neglected when building 3D-QSAR models. The relevance of correlation and distribution profiles is tested in terms of the models' prediction power. A different approach to filter descriptors prior to variable selection is proposed. Additionally, a protocol for molecular interaction field descriptors selection and model validation is presented. The algorithms and protocols presented are quite simple and enable a different and powerful way to create parsimonious interaction field-based models.

Keywords: Correlation coefficient · Distribution profiles · MIF descriptors

1 Introduction

The aim of quantitative structure-activity and structure-property relationships (QSAR/QSPR) is to create models that can predict activity or property and explain the relationships between molecular features and the biological activity or a given property. Once such a model is found, a set of compounds of reasonable size, including those not yet synthesized and placed on the model's response surface, can be screened automatically to design structures with the desired properties. It is then possible to select the most promising compounds for synthesis and bioassays in the laboratory. Thus, QSAR/QSPR studies can reduce costs and accelerate the process of designing new substances to be used as drugs, new materials, additives, or even for other purposes.[1]

Spatial molecular properties are used to generate descriptors to build QSAR models in the so-called 3D-QSAR. The best known 3D-QSAR methodology is the comparative molecular field analysis (CoMFA).[2] In a standard CoMFA procedure selected conformers of the studied molecules (one per molecule) are superimposed in a manner defined by a supposed mode of interaction with a target macromolecule. Then, the steric and electrostatic molecular interaction fields (MIF) of these compounds are calculated with a probe such as a sp^3 carbon atom with $+1$ charge, at regularly spaced points of a three-dimensional grid. The calculated energy values are then related to the biological activity by partial least-squares (PLS) regression. The final CoMFA model is derived using the optimum number of latent variables (LV) determined by cross-validation and the results are usually displayed as contour maps.[3] A good CoMFA model must have satisfactory statistical significance, explanatory capability for the variance of the activity of the compounds in the training set, and sufficient predictive power for new compounds.

MIF are computed using intrinsically continuous functions, described by analytic expressions at the 3D grid points defined by regular intervals over the space surrounding the molecules. This method has the disadvantage of producing a huge amount of computed data (potential descriptors). Not all of these variables describing the space around the molecules are equally relevant. Usually, useful descriptors are concentrated in specific regions characterized by important molecule-probe interactions within the 3D grid. If the MIF is used to describe a binding pocket within a biological receptor, then the corresponding regions represent promising locations for molecular modification. On the other hand, when the MIF is used to characterize intrinsic ligand properties, the corresponding regions represent groups of the receptor binding site with which the ligand molecule could establish favorable binding interactions. In either case, these regions mean favored locations holding highly relevant information for describing the ability of the two molecules to interact.[4]

It is obvious that variable selection is a very important issue when building 3D-QSAR models. By selecting appropriate descriptors, it is possible to construct interpretable, robust and predictive models. Numerous methods have been developed for this purpose and are widely applied in 3D-QSAR.[5] The MIF descriptors are most commonly submitted to variable selection using the GOLPE approach (generating optimal linear PLS estimates),[6] the Smart Region Definition (SRD)[7] and the Modified Iterative/Uniformative Variable Elimination-PLS (IVE/UVE-PLS) Method.[8] These procedures were optimized to produce interpretable contour maps.

[a] *E. G. Barbosa, M. M. C. Ferreira*
University of Campinas – UNICAMP
Campinas, Brazil
zip box: 6154, zip code: 13083-970
**e-mail: marcia@iqm.unicamp.br*

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201100181.

A new MIF methodology has recently been proposed[9] that uses certain features of the CoMFA and 4D-QSAR paradigms proposed by Hopfinger et al.[10] This methodology was named LQTA-QSAR after our research group LQTA (Laboratório de Quimiometria Teórica e Aplicada - LQTA-QSAR) and makes use of the free GROMACS package[11] to execute molecular dynamics simulations and provide conformational ensemble profiles (CEPs) generated for each compound in a data set. Instead of using a genetic algorithm for variable selection as in 4D-QSAR,[12] the MIF descriptors in LQTA-QSAR are selected by means of Ordered Prediction Selection (OPS).[13]

During the development of the LQTA-QSAR methodology, special attention was devoted to eliminate certain descriptors (filtering) prior to using the specialized variable selection algorithm OPS. Two MIF descriptor features were considered. One of these features is the correlation of a given descriptor studied and the biological activity data (*y*). The other was the spread of each descriptor when compared to *y*, i.e., pronounced dispersion of the data points around the regression line in the corresponding bivariate plot (Figure 1).

The MIF-based QSAR model must make use of descriptors that are well correlated to *y* in terms of normal or quasi-normal bivariate distributions and their distributions should also be normal or quasi-normal. This type of descriptor responds to variation in *y* unlike binary ones. Common normal distribution tests among others can be used to detect such features as the Kolmogorov-Smirnov,[14] D'Agostino-Pearson,[15] Shapiro-Wilk,[16] Lilliefors[17] and Jarque-Bera tests,[18]. Another "simpler" way to assess the normality or quasi-normality is by inspection of scatterplots of *y* versus a given descriptor, as shown in Figure 1. Descriptor 1 shows good bivariate distribution profile and descriptor 2 a bimodal distribution. MIF descriptors are derived from continuous functions and, thus, well scattered normal/quasi-normal descriptors are more suitable for 3D-QSAR models. Descriptors, such as those in Figure 1 (descriptor 2), have to be avoided. As mentioned before, one may eliminate these poorly distributed descriptors by using normality tests, but such tests can eliminate useful descriptors, since minor deviations from normality or quasi-normality can be detected.

In 1973 Anscombe[19] had already pointed out the usefulness of scatterplots in graphical analyses to distinguish how independent variables are distributed in relation to *y*. Common numerical statistical parameters like correlation coefficients alone or linear regression parameters are not able to successfully give the same information. It means that by analyzing bivariate plots of *y* and a descriptor it is possible to know if a descriptor follows a normal or quasi-normal distribution such as that indicated in Figure 1 (Descriptor 1). Unfortunately, the use of human intervention to inspect each descriptor generated from any grid-based methodology is unfeasible. In this work, a new automated methodology, named the Comparative Distribution Detection Algorithm (CDDA) is proposed, which is able to classify descriptors by means of distribution profile. CDDA compares individual distributions of a descriptor and *y* and calculates dissimilarity statistics, which enables quick numerical inspection of bivariate scatterplots and aids in selection (filtering) of descriptors suitable to build a multivariate model.

In order to demonstrate the usefulness of descriptor filtering prior to variable selection, three data sets were selected from the literature, for which CoMFA models were re-built. Poorly correlated descriptors and those not well distributed in relation to *y* were eliminated and PLS models were constructed with the remainder of descriptors. The resulting models were compared with analogues built without using the CDDA filter. Comparisons with results from the original articles were also made. LQTA-QSAR software

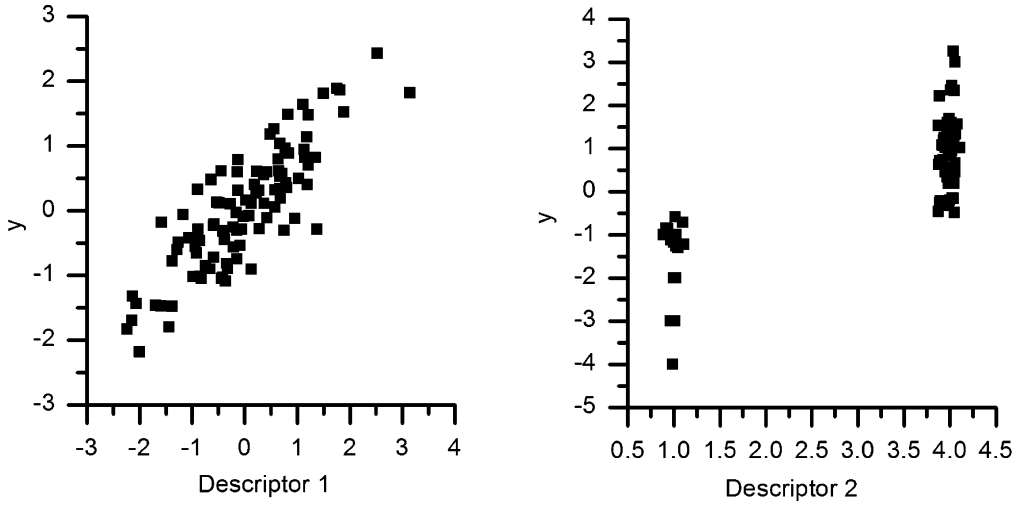

Figure 1. Scatterplots of two hypothetical biological activities *y* and two descriptors with different scattering profiles. Both descriptors have similar Pearson correlations with *y* (0.8), but the distribution profiles are quite different.

was adapted to reproduce the original data matrices, rebuilt in human readable format. Alongside with the CDDA application in descriptor filtering, a new protocol to build 3D-QSAR models is discussed in detail. The algorithms were written in MATLAB open code and are available at lqta.iqm.unicamp.br.

2 Computational Methods

Notation: Scalars are defined as italic lower case characters, vectors are typed in bold lower case characters and matrices as bold upper case characters. A descriptor retrieved from the *j*-th column of the **X**(*I,J*) matrix is denoted as x_j.

Prior to any filtering and variable selection, descriptors related to probe positions far from molecular conformations were eliminated by a variance cut-off in which descriptors with variance below of 0.01 were excluded.[20]

The Lennard–Jones (LJ) descriptors were submitted to a special treatment due to the presence of positive values of large orders of magnitude. When the probe is in a position of the grid very close to a particular atom of the target molecule, the potential energy values are of high orders of magnitude due to extremely high repulsion energies derived from the LJ equation. Since using both, large and small numerical values can be harmful for PLS modeling, they were submitted to a pretreatment satisfying the rules given in Equation 1. This transformation ensures that the information regarding grid points close to ligand atoms were not completely lost. If a given LJ descriptor computed at any *x,y,z* position had its value of energy equal to or higher than 30 kcal/mol, the log10 of the remainder energy value was added to 30 kcal/mol. If positions in the grid have had all the energy values transformed, then the particular grid points were eliminated. Such transformations were applied to LJ but not to Coulomb (QQ) descriptors.

$$\begin{cases} LJ_{x,y,z} < 30 \text{ kcal/mol} \Rightarrow LJ'_{x,y,z} = LJ_{x,y,z} \\ LJ_{x,y,z} \geq 30 \text{ kcal/mol} \Rightarrow \\ LJ'_{x,y,z} = 30 + \log\left(\frac{LJ_{x,y,z}}{\text{kcal}\cdot\text{mol}^{-1}} - 29\right) \end{cases} \quad (1)$$

The resulting descriptors were then filtered by the use of a correlation coefficient cut-off. In this simple procedure, descriptors having absolute Pearson correlation coefficients ($|r|$) with the **y** vector at the same level of random noise are eliminated from the pool. This level is obtained by calculating *r* of **y** and a large number of random vectors ($\mathbf{r}_{rand}$). The histogram of $\mathbf{r}_{rand}$ is to follow a normal distribution around the mean ($\mu = 0$) and the upper confidence limit at 99% of confidence for this distribution can be used as the $|r|_{cutoff}$ level. It was observed that the confidence limit varies, depending on how **y** is distributed and the degrees of freedom involved. Therefore, each data set that was chosen to test the filtering procedure had its own $|r|_{cut\text{-}off}$ level calculated by Equation 2.

$$|r|_{cut\text{-}off} = Z_{0.99}\sigma_\mu \quad (2)$$

where $Z_{0.99}$ is the number of standard deviations extending from the mean of a normal distribution ($\mu = 0$) required to contain 99% of the area and σ_μ is the standard error of the mean. When the calculated $|r|_{cut\text{-}off}$ was lower than 0.3 this threshold was used.[21]

The next task is to eliminate descriptors with poor distribution profiles with respect to **y** using CDDA filtering. CDDA provides a way to quantify how similarly distributed is **y** and a given descriptor, aiding the removal of those similar to descriptor 2 in Figure 1. CDDA is rather simple, which can be seen from the step-by-step complete algorithm's description that follows.

Step 1: Each descriptor vector $\mathbf{x}_j$ ($j = 1, 2, \ldots, J$) is range-scaled so its minimum value is zero and the maximum value is equal to one (Equation 3). A similar treatment is applied to the dependent variable **y** to obtain **y**′

$$\mathbf{x}'_j = \frac{\mathbf{x}_j - \mathbf{1}x_{j(\min)}}{x_{j(\max)} - x_{j(\min)}} \quad (3)$$

where **1** is a ($I \times 1$) vector of ones, and *I* is the number of samples within the data set,

$$x_{j(\min)} = \min_{1 \leq i \leq I} |x_{ij}|$$

and

$$x_{j(\max)} = \max_{1 \leq i \leq I} |x_{ij}|.$$

Step 2: Similar to a distribution histogram, for each descriptor *j*, the interval [0,1] is divided into *K* subdivisions, $k = 2^n$ ($n = 1, 2, 3, \ldots$), say for example, [0,1/2] and [1/2,1] for $k = 2$, and how many values are in each subdivision is checked. This is carried out by the routine below.

for $k = 1, \ldots, K$

for $i = 1, \ldots, I$

if $\{ x_i \geq 2^{-n} (k-1)$ and $x_i < 2^{-n} k \}$

or $\{ x_i = 2^{-n} k$ and $x_i = 1 \}$;

$f_{ik} = 1$

else

$f_{ik} = 0$; end; end; end

The results are recorded in a logical matrix **F**($I \times K$) and then transformed into the row vector $\mathbf{f}_{k(j)}$ by summing over the rows (accordingly to Equation 4). Each element of $\mathbf{f}_{k(j)}$ vector contains the number of samples in each subdivision *k* for each variable *j*.

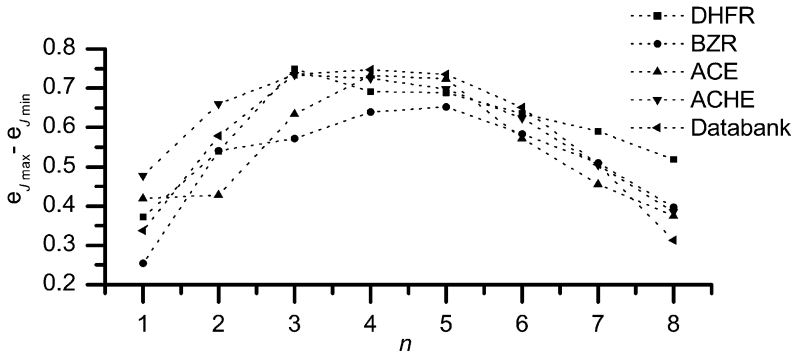

Figure 2. Results from investigations to define the default value for *n* for some data sets retrieved from the literature and our databank. DHFR stands for dihydrofolate reductase,[23] BZR for benzodiazepine receptor,[24] ACE for angiotensin converting enzyme[25] and ACHEfor acetyl-cholinesterase.[26]

$$\mathbf{f}_{k(j)} = \sum_{i=1}^{I} f_{i,k(j)}, \qquad k = 1, \cdots, K \tag{4}$$

A similar procedure is done for $\mathbf{y}'$ resulting in the reference frequency vector $\mathbf{f}_{k(y')}$.

Step 3: The difference between each $\mathbf{f}_{k(j)}$ and the $\mathbf{f}_{k(y')}$ vectors is computed as the residual vector $\boldsymbol{v}_{k(j)}$ (Equation 5).

$$\mathbf{v}_{k(j)} = \mathbf{f}_{k(j)} - \mathbf{f}_{k(y')} \tag{5}$$

Step 4: The total error (residual) for the *j*-th descriptor for a given *n*, $\varepsilon_j(n)$, is calculated by taking the **1**-norm of vector $\boldsymbol{v}_{k(j)}$ (Equation 6).

$$\epsilon_j(n) = \sum_{k=1}^{K} |V_{k(j)}| \tag{6}$$

If a descriptor has exactly the same distribution as **y**, then any $\varepsilon_j(n)$ value is zero. But when a descriptor $\boldsymbol{x}_j$ and **y** have different distributions, $\boldsymbol{v}_{k(j)}$ is no longer a null vector, indicating that there are dislocated values. This means that there are over-populated subdivisions and, at the same time, under-populated ones, increasing the value of $\varepsilon_j(n)$. Furthermore, the two extreme values (0 and 1) are fixed and do not contribute to the $\boldsymbol{v}_{k(j)}$ vector. Thus, the maximum value ε_j for a descriptor can be given as the number of samples (*I*) minus 2, and each value of $\varepsilon_j(n)$ can be normalized to give $e_j(n)$ (Equation 7). The values of $e_j(n)$ provide a parameter for comparison of **y** and a descriptor by means of their distribution profiles. The closer $e_j(n)$ is to 1.00, the more similar are the distributions of **y** and a descriptor *j*.

$$e_j(n) = 1 - \frac{\varepsilon_j(n)}{2I - 2} \tag{7}$$

The $e_j(n)$ values are functions of *n* and, thus, of the number of subdivisions. To define a default value for *n*, several data sets were used for computing the difference between maximum and minimum values of $e_j(n)$ for the entire descriptor pools, in order to investigate how $e_j(n)$ varies as *n* changes. This was achieved by varying *n* from 1 to 8 (2 to 256 subdivisions), and calculating $e_j(n)$ for matrices reproduced by the Dragon[22] and LQTA-QSAR programs for data sets retrieved from literature and our own databank. The values of *n*, for which the largest differences between maximum and minimum e_j were obtained, were 3, 4 and 5 (Figure 2).

By analysing Figure 2, the default chosen for *n* was 4, meaning the value at which the CDDA was the most sensitive. The parameter $e_j(4)$ is simply denoted as e_j in the remainder of the text.

Based on tests carried out on the data sets a default value for e_j cutoff of 0.5 was suggested. This parameter is large enough to remove only very poorly distributed descriptors. This parameter is a user-definable input in the Matlab command line.

The χ^2 goodness-of-fit[27] is a well-known method for determining whether observed frequencies deviate significantly from their expected frequency. The observed frequencies (O) can be associated to $\mathbf{f}_{k(j)}$ and the expected frequency (*E*) to $\mathbf{f}_{k(j)}$. However, it was observed that the way CDDA is presented provides a stricter way to differentiate poorly and badly distributed descriptors since it doubles the penalty for each error when comparing *O* and *E*. From equations in (8) the difference between the two approaches is visible.

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{I} (O_i - E_i)^2 / E_i \\ e_j &= \sum_{i=1}^{I} 1 - \frac{(O_i - E_i)}{2I - 2} \end{aligned} \tag{8}$$

3 Method Validation

In order to test the applicability of CDDA for MIF descriptors and to verify whether it is able to improve 3D-QSAR

models in terms of prediction and interpretation, three data sets from literature in which the CoMFA method was applied were selected. The chosen data sets contained 3D molecular conformations in mol2 format with atom type specifications and attributed atomic charges. Data set (1) consisted of 114 angiotensin converting enzyme (ACE) inhibitors (84 compounds in the training set and 30 for external validation) retrieved from the work of Depriest et al.[25] The activities, expressed as pIC_{50}, are spread over a wide range of values (from 2.1 to 9.9). Data set (2) consisted of 40 phase-transfer asymmetric catalysts[28] (30 samples for calibration and 10 for external validation), based on quaternary ammonium salts, with experimentally observed selectivity expressed in percentages ranging from −30% to 91%. Data set (3) had 362 dihydrofolate reductase (DHFR) inhibitors (280 samples for calibration and 84 samples for external validation) extracted from the work of J. J. Sutherland and D. F. Weaver,[23] with pIC_{50} values ranging from 3.34 to 9.8.

At first, a model with all samples was built and completely validated, according to the suggested statistical tests recommended in the literature.[21] Then, the external data samples (test set) were defined based on dendrograms built with the selected descriptors using Hierarchical Cluster Analysis (HCA).[29] In HCA, the Euclidian distances and the '*complete*' linkage method to cluster similar samples, were used. HCA provided a visual aid to define the prediction set and to avoid removing samples in isolated clusters. The lowest and highest values for *y* were never selected. Once the samples from the test set were defined, the complete data set was split into training and test sets. New models were built using the training set, never including samples from the test set during the processes of variable filtering and selection. *If and only if* the models built with the training set were considered well validated, had good figures of merit and presented no discrepancies between the sign of correlation and regression coefficient,[29] they were used to make predictions on the test set.

Since it is not possible to freely access the CoMFA data matrices, one can re-generate similar descriptors using the LQTAgrid module of the LQTA-QSAR package.[9] LQTAgrid was designed to deal with an ensemble of molecular conformations when calculating the MIF descriptors, however, a single conformation can be handled as well as in CoMFA. Using a similar probe (carbon sp^3 bearing single +1 point charge) and the published mol2 charges, the Coulomb interaction energy descriptors were calculated. The Lennard-Jones descriptors were created using interactions of GAFF[30] atom types instead of those in Sybyl. It is reasonable to expect that the regenerated descriptors, Coulomb and Lennard–Jones, are related to those from CoMFA and so the results from this work can be somewhat comparable to those from the literature.

The literature mol2 files were converted into the appropriate LQTAgrid usage format. The grid dimensions used were large enough to outgap the coordinates by 5 Å, with 0.5 Å increments, producing a total of 158400 descriptors, for dataset (1). For dataset (2) 162922 descriptors were computed, and for dataset (3) 112800 descriptors. The descriptor fields were separated into Coulomb interaction descriptors (QQ field) and Lennard-Jones interaction descriptors (LJ field).

After the classical variance cut-off, the first step in data filtering was the elimination of descriptors for which the correlation coefficient $|r|$ with the dependent variable *y* was lower than the noise level. This procedure guarantees the elimination of the regions in the grid space that cannot be identified as related to biological activity in terms of statistical significance. The resulting QQ and LJ fields were submitted to the CDDA filter. This resulting matrix was then ready for variable selection and regression model building using autoscaled data (each descriptor was mean-centered and scaled to unit variance). Variable selection was carried out by the OPS algorithm.[13]

The final models were validated by using leave-*N*-out (LNO) crossvalidation and **y**-randomization tests which are highly recommended to check model robustness and the presence of chance correlations, respectively. In the first test, the training set of *I* samples is divided into consecutive blocks of *N* samples. Each block is excluded once, a new model is built without it, and the values of the dependent variable are predicted for the block in question. LNO is performed for $N = 2$, 3, etc., and the leave-*N*-out cross-validated correlation coefficients (Q^2 LNO, Table 1) are calculated. The model is considered robust if the deviation of Q^2 leave-one-out (LOO) and other Q^2LNO does not exceed 0.05 for all values of *N* up to 20–30% of samples from the training set.[21,31]

The **y**-randomization test consists of several runs for which the original descriptor matrix **X** is kept fixed, and only the **y** vector is randomized (scrambled). The models obtained under such conditions should be of poor quality and without real meaning. Two **y**-randomization plots $|r|$ x Q^2_{yrand} and $|r|$ x R^2_{yrand} are drawn for randomized and real models, and the linear regression lines are obtained. ($Q^2_{yrand} = a_Q + b_Q|r|$, $R^2_{yrand} = a_R + b_R|r|$). The real model is characterized as free of chance correlation when the intercepts are $a_Q < 0.05$ and $a_R < 0.3$.[32]

The most successful models were presented and compared to those originally from the literature. The descriptors illustrated in 3D space were depicted as spheres using Chimera software[33] to aid visualization. The statistical quality of the final models was expressed as Q^2 LOO and by the correlation coefficient of external validation (Q^2_{pred}, Table 1) obtained for the external validation set. The same procedure: correlation cut-off, variable selection and validation were applied to QQ and LJ fields without using the filter CDDA for the sake of comparison.

Table 1. Basic statistical parameters for regression models. *I* is the number of samples (training set or external validation set), *i* is the summation index and also the index of the *i*-th sample ($i = = 1, 2, \ldots, I$); *y*: experimental values of **y**; y_c: calculated values of **y** for the training set; y_v: calculated values of **y** from internal validation (LOO, LNO or **y**-randomization); y_p: predicted values of **y** for the external validation set. $<\mathbf{y}>$ is the average for experimental values of **y** calculated for the training set and not for the external validation set.[21]

Parameter	Definition
Leave-one-out (LOO) and leave-*N*-out (LNO) cross-validation correlation coefficients	$Q^2 = 1 - \frac{\sum_{i=1}^{I}(y_i - y_{vi})^2}{\sum_{i=1}^{I}(y_i - \langle \mathbf{y} \rangle)^2}$
Correlation coefficient of multiple determination	$R^2 = 1 - \frac{\sum_{i}^{I}(y_i - y_{ci})^2}{\sum_{i}^{I}(y_i - \langle \mathbf{y} \rangle)^2}$
External validation correlation coefficient	$Q^2_{\text{pred}} = 1 - \frac{\sum_{i}\left(y_i - y_{pi}\right)^2}{\sum_{i}(y_i - \langle \mathbf{y} \rangle)^2}$

4 Results and Discussion

The variable count reduction after the application of the cut-offs mentioned are summarized in Table 2. It can be noted that the filters enabled substantial decrease from the initial pool and clearly resulted in a more tractable amount of descriptors for specialized variable selection. Highly intercorrelated descriptors (data not shown) were also removed.

The figures of merit for all models after the procedures described are summarized in Table 3. Comparisons with the original articles are reported accordingly.

For the sake of comparison with the CoMFA procedure 15 bootstrapping run were performed and the resulting values of $Q_{\text{pred}}{}^2$ for randomized external data sets were: (1) 0.74 ± 0.12, (2) 0.84 ± 0.15 and (3) 0.60 ± 0.15 for data sets (1), (2) and (3), respectively. These results are in agreement with the selected external data set demonstrating the model robustness of the model to predict.

Dataset (1). The resulting matrix after the correlation and CDDA applied in the LJ block for data set (1) is shown in 3D space in Figure 3. Can be noted a clear definition of regions accompanying the molecular shape. After applying

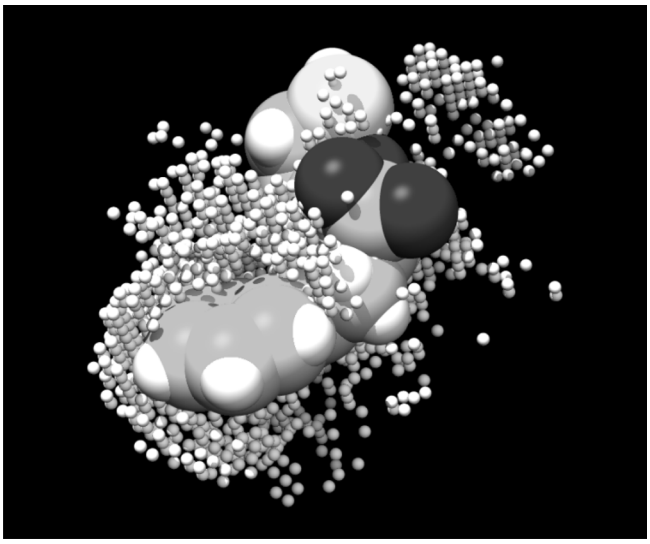

Figure 3. Grid points left after using correlation and CDDA cutoffs on the LJ field. One conformation is shown by its van der Walls volume.

Table 2. Descriptor count reduction along the filtering procedure and the final model descriptor count. FM stands for Final Model.

Data set	Initial count	Variance cut-off (0.01)	$\lvert r \rvert_{\text{cut-off}}$ level [a]	$\lvert r \rvert$ cut-off	*CDDA* (0.5)	*FM*
(1)	158400	60512	0.26 (0.3)	29047	2156	6
(2)	162922	69518	0.43	14841	2825	5
(3)	112800	38516	0.14 (0.3)	2297	277	9

[a] The actual values used are in parenthesis.

Table 3. Figures of merit of the final models obtained for each data set.

Data Set	Q^2 LOO	R^2	$Q_{\text{pred}}{}^2$	*SEP*	*NV*[a]	LV[b]
(1)	0.76	0.79	0.64	1.11	6	2
Depriest et al.[25]	0.66	0.77	–	1.31	–	–
(2)	0.96	0.97	0.73	5.49	5	4
Melville et al.[28]	0.78	0.94	0.64	–	–	4
(3)	0.59	0.62	0.60	0.88	9	9
Sutherland[23]	0.65	0.76	0.52	–	–	7

[a] Number of variables included in the final model. [b] Number of optimum latent variables. In bold characters are important comparison figures. Not all values are expressed for comparison because not all of them were presented in the original articles.

correlation and CDDA cutoffs on the QQ field, the number of descriptors was reduced to 1 720. Figure 4 illustrates the remaining descriptors.

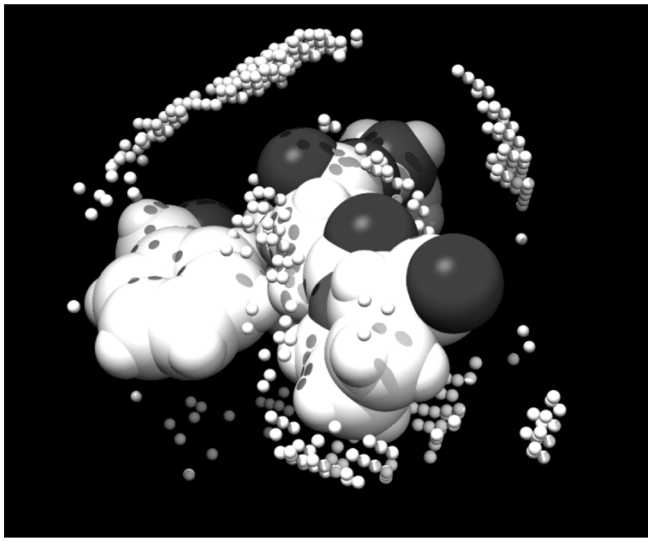

Figure 4. Grid points left after using correlation and CDDA cutoffs on the QQ field. One conformation is shown by its van der Waals volume.

The matrix formed by the joined QQ and LJ fields resulted in 2156 descriptors that were ready to be submitted to variable selection with common computational tools such as genetic algorithms,[34] feature selection[35] and so forth. LQTA-QSAR uses OPS[13] as the default variable selection tool, hence the final models were achieved by means of this algorithm. The final PLS models have 6 descriptors (Figure 5) and 2 latent variables. The obtained Q^2 LOO value is 0.77, higher than the literature value (0.68).[25] The obtained Q^2_{pred} value was 0.68, which can be considered a fairly reasonable predictive power. The leave-*N*-out results in Figure 6 show clearly the robustness of the model. The LNO validation shows that deviation of Q^2 LNO from Q^2 LOO stays lower than 0.05 after taking 33 samples out (39%). The **y**-randomizations tests (Figure 7) indicate that this model it was not obtained by chance.

When the 3D-QSAR study by Depriest et al.[25] was published, no crystallographic data regarding any compound within the enzyme binding pocket were known. Such information could be used to generate a better model for dataset (1). Only in 2003 was a structure of the human angiotensin-converting enzyme lisinopril complex published.[36] The conformations used and the alignment are likely to be biased by the neglected data. Unfortunately, the literature lacks a paper presenting such an approach[37] and applying the new crystallographic data is beyond the scope of this article.

The usefulness of CDDA was verified by producing a models without its use. The descriptor matrix submitted only to correlation cut-off was much larger (29 047 descriptors). After OPS variable selection, interestingly, the final model also yielded good figures of merit. The model had 8 descriptors and 3 latent variables; the Q^2 LOO value obtained was 0.80, a littler higher than the value obtained using the filter (0.77). The Q^2_{pred} dropped from 0.68 (CDDA) to 0.63. The model performed well in the **y**-randomization test, but the LNO test proved to be much worse (Figure 8). Even though this model was still capable of doing prediction, it appears that when poorly distributed descriptors are used, there is a higher possibility of removing samples that entirely change the descriptor distribution, leading to

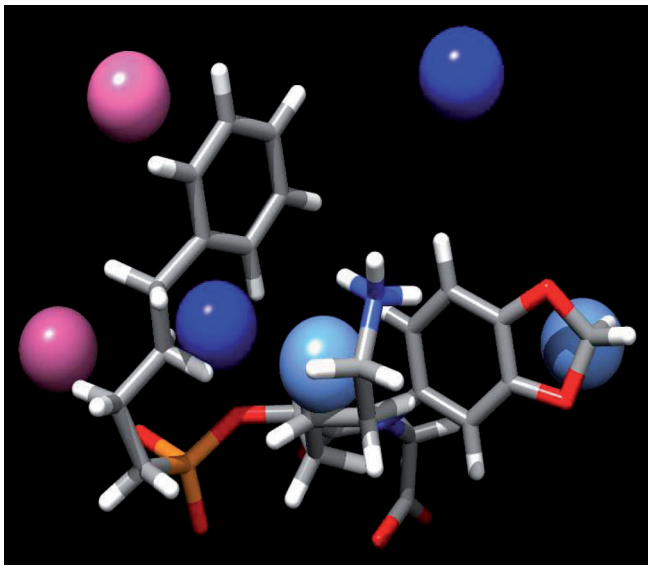

Figure 5. The best PLS model for data set (1). Negatively correlated descriptors are depicted in pink and red, and blue color is for the positively correlated descriptors.

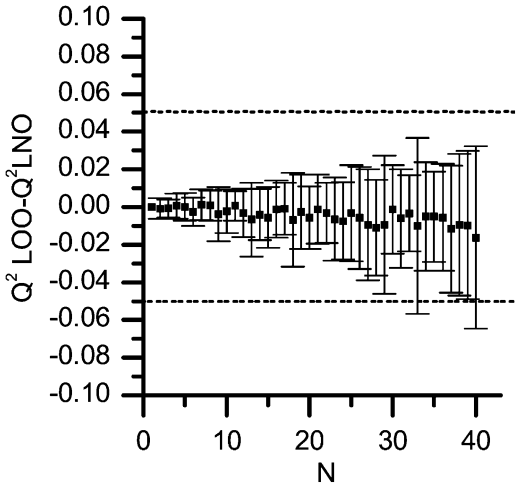

Figure 6. Results of leave-*N*-out test for data set (1). The values are the average of 20 rearrangements of the data. The error bars are two times the standard deviations.

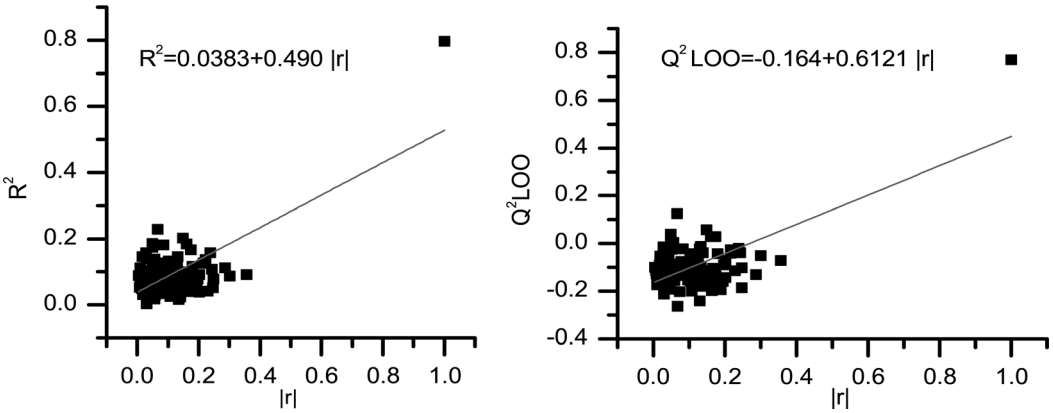

Figure 7. Results from **y**-randomization for data set (1). The intercept values are adequate for the validation procedure, showing that the model does not suffer from chance correlation.

unpredicted behavior when the model is used for internal validation. The higher obtained Q^2 LOO seems to be prone to overfiting during variable selection with OPS. Any other variable selection procedure would be affected by the same pitfalls.[38]

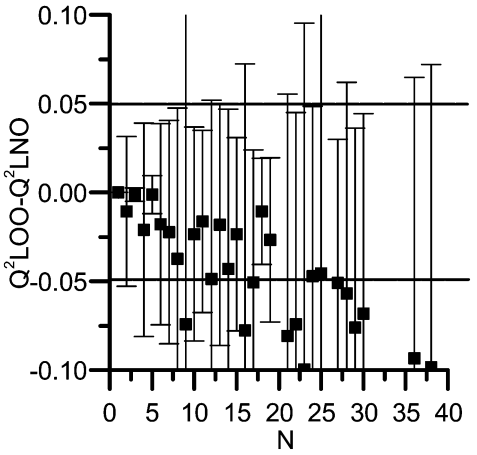

Figure 8. Results of the leave-*N*-out test for dataset (1) when the model is built without CDDA filtering. The values are the average of 20 rearrangements of the data. The error bars are two times the standard deviations.

Dataset (2). No good models were obtained with this data set using the proposed method. This problem was overcome when the atomic charges were changed to AM1-BCC charges.[39] The work of Mittal et al.[40] demonstrates improvements when the charges are changed for the same data set.

After the correlation and CDDA filtering procedure, the QQ field was reduced from 34 759 to 2 348 descriptors. The LJ field presented only 477 descriptors after filtering. These two fields were submitted to the already described variable selection and validation tests. The final model is surprisingly simple, having only 5 descriptors (Table 2) and 4 latent variables. The Q^2 LOO obtained is 0.955 and R^2 0.973, and the prediction is also satisfactory ($Q^2_{pred} = 0.73$). The model performed well in the validation tests. Detailed information can be retrieved in the Supporting Information.

The original paper where the CoMFA model is presented,[28] in addition to the single conformation to obtain MIF, also employed molecular dynamics simulation to explore the conformational space to produce 4D-QSSR (Quantitative Structure-Selectivity Relationship) and used Boltzmann-weighting the contribution of selected minimized conformations to produce "3.5D"-QSSR models. In order to avoid mixing up the concepts of 4D-LQTA-QSAR and approaches of the original paper only monoconfigurational mol2 files were retrieved. Concerning model interpretation, little was discussed about the descriptors obtained, since the mechanism of selectivity of the catalysts was not explored.[41]

Dataset (3). This data set possesses a much smaller number of descriptors well correlated to **y**. The final model has still satisfactory statistics (Q^2 LOO of 0.61 and R^2 of 0.65). Although the Q^2 LOO value from the original article[21] is somewhat higher (0.65), the model presented in this work had better performance when used to make prediction (Q^2_{pred} of 0.60 versus 0.52). As mentioned before, when using current information on targets in the process of QSAR model building, recent crystallographic information could be used to produce better models for this data set as well.

Detailed interpretation of the models obtained is beyond the scope of this article, but judging by its prediction power expressed as Q^2_{pred}, the concordance of correlation and regression vector signs, and their simplicity, it can be stated that these models are more reliable than the literature analogues. Some of the results could be improved if more recent knowledge about mechanism of action and the receptor structures were used.

When CDDA was not applied to dataset (2) the external prediction for the blind set was largely compromised (Q^2 LOO $= -0.16$). Interestingly, for dataset (3) not using CDDA filter did not implicate on detrimental results for the final

model's prediction power or Leave-*N*-out test (Supporting Information). After variable selection by OPS, basically the same descriptors were selected.

5 Conclusions

The descriptors resulting from variance cut-off enabled the removal of descriptors too far from the molecular surface. A new way of transforming LJ descriptors was presented. The proposed transformation allows for maintenance of descriptor variability, unlike the usual cut-off procedure at 30 kcal/mol.

The elimination of non-informative grid points was derived by the determination of a critical correlation cut-off value. This value aids when deciding whether a descriptor has a 99% of chance to be correlated to a random vector. This approach proves to be very useful to set a specific level of cut-off based on the biological data available.

Regard the CDDA simplicity, when it is applied to restrain variable selection for well distributed MIF descriptors, valuable results are obtained, mainly regarding the stability of the models. Correlation analysis and scatterplots to analyze relations between the descriptors and the dependent variable are features commonly neglected in the QSAR literature. Often it is possible to find very poor quality QSAR and 3D-QSAR models being published, which is the reason for a certain degree of skepticism about the methodology and its applicability.

Correlation and CDDA cut-offs do not work as variable selection procedures. They must be understood as a filtering step prior to variable selection. The models obtained with the filtered pool had good prediction power and the descriptors were the most informative regions in space. The 3D-QSAR models obtained are as simple as classical QSAR models.

There are certain types of descriptors derived from discrete properties of compounds, such as the number of oxygen and carbon atoms and scaffold positions for a substituent, among others, that should not be submitted to CDDA since they can be important and should not be eliminated. So it is recommended to use CDDA to filter only intrinsically continuous descriptors.

It can happen that **y** is not well distributed, not having normal or quasi-normal distribution, hence CDDA is likely to keep descriptors with the same distribution profile. A model with this type of *y* will produce gaps within the response model surface and, consequently, predictions falling inside these regions might not be reliable. One solution for this problem is finding data to fill these gaps to improve the distribution of **y**.

Finally, the MIF descriptor filtering performed by the correlation cut-off and by CDDA proved to be useful tools to aid in formal variable selection.

Acknowledgements

The authors acknowledge the Brazilian Governmental Agencies *CAPES* and *FAPESP* for financial support and Professor Carol H. Collins for English revision. Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the *University of California, San Francisco* (supported by *NIH* P41 RR001081).

References

[1] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev.* **1996**, *96*, 1027 – 1044.
[2] R. D. Cramer, D. E. Patterson, J. D. J. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959 – 5967.
[3] G. Cruciani, in *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*, Wiley-VCH, Weinheim, **2006**. p. 5.
[4] A. N. Durán, G. C. Martínez, M. Pastor, *J. Chem. Inf. Model.* **2008**, *48*, 1813 – 1823.
[5] M. Arakawa, K. Hasegawa, K. Funatsu, *Curr. Comput-Aid. Drug Des.* **2007**, *3*, 254 – 262.
[6] M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, *Quant. Struct.-Act. Relat.* **1993**, *12*, 9 – 20.
[7] M. Pastor, G. Cruciani, S. Clementi, *J. Med. Chem.* **1997**, *40*, 1455 – 1464.
[8] R. Grohmann, T. Schindler, *J. Comput. Chem.* **2008**, *29*, 847 – 860.
[9] J. P. A. Martins, E. G. Barbosa, K. F. M. Pasqualoto, M. M. C. Ferreira, *J. Chem. Inf. Model.* **2009**, *49*, 1428 – 1436.
[10] A. J. Hopfinger, S. Wang, J. S. Tokarski, B. Jin, M. Albuquerque, P. J. Madhav, C. Duraiswami, *J. Am. Chem. Soc.* **1997**, *119*, 10509 – 10524.
[11] S. Van Der David, L. Erik, H. Berk, G. Gerrit, E. M. Alan, J. C. B. Herman, *J. Comput. Chem.* **2005**, *26*, 1701 – 1718.
[12] D. Rogers, *WOLF 6.2 GFA Program,* Molecular Simulations, San Diego, **1994**.
[13] R. F. Teófilo, J. P. A. Martins, M. M. C. Ferreira, *J. Chemometr.* **2009**, *23*, 32 – 48.
[14] F. J. Jr. Massey, *J. Am. Stat. Assoc.* **1951**, *46*, 68 – 78.
[15] E. S. Pearson, R. B. D'Agostino, K. O. Bowman, *Biometrika* **1977**, *64*, 231 – 246.
[16] S. S. Shapiro, M. B. Wilk, *Biometrika* **1965**, *52*, 591 – 611.
[17] H. W. Lilliefors, *J. Am. Stat. Assoc.* **1967**, *62*, 399 – 402.
[18] C. M. Jarque, A. K. Bera, *Int. Stat. Rev.* **1987**, *55*, 163 – 172.
[19] F. Anscombe, *J. Am. Stat.* **1973**, *27*, 17 – 21.
[20] H. Kubinyi, *Drug Discov. Today* **1997**, *2*, 457 – 467.
[21] R. Kiralj, M. M. C. Ferreira, *J. Braz. Chem. Soc.* **2009**, *20*, 770 – 787.
[22] R. Todeschini, V. Consonni, M. Pavan, *DRAGON for Windows*, Milano Chemometrics and QSAR Research Group, Milano, Italy, **2002**.
[23] J. J. Sutherland, D. F. Weaver, *J. Comput. Aid. Mol. Des,* **2004**, *18*, 309 – 331.
[24] D. J. Maddalena, G. A. R. Johnston, *J. Med. Chem.* **1995**, *38*, 715 – 724.
[25] S. A. DePriest, D. Mayer, C. B. Naylor, G. R. Marshall, *J. Am. Chem. Soc.* **1993**, *115*, 5372 – 5384.
[26] A. Golbraikh, P. Bernard, J. R. Chrétien, *Eur. J. Med. Chem.* **2000**, *35*, 123 – 136.

[27] G. W. Snedecor, W. G. Cochran, *Statistical Methods*, 8th ed, Iowa State University Press, Ames, Iowa, **1989**.
[28] J. L. Melville, K. R. J. Lovelock, C. Wilson, B. Allbutt, E. K. Burke, B. Lygo, J. D. Hirst, *J. Chem. Inf. Model.* **2005**, *45*, 971 – 981.
[29] R. Kiralj, M. M. C. Ferreira, *J. Chemometr.* **2010**, *24*, 681 – 693.
[30] J. Wang, W. Wang, P. A. Kollman, D. A. Case, *J. Mol. Graph. Model.* **2006**, *25*, 247 – 260.
[31] P. Gramatica, *QSAR Comb. Sci.* **2007**, *26*, 694 – 701.
[32] L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, *Environ. Health Perspect.* **2003**, *111*, 1361.
[33] E.F Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.* **2004**, *25*, 1605 – 1612.
[34] M. M. C. Ferreira, C. A. Montanari, A. C. Gaudio, *Quim. Nova* **2002**, 25, 439 – 448.
[35] A. H. Mark, *Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, in Proc. XVII Int. Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA, **2000**.
[36] R. Natesh, S. L. U. Schwager, E. D. Sturrock, K. R. Acharya, *Nature*, **2003**, *421*, 551 – 554.
[37] J. San, A. Amor, C. H. O. Seung Joo, *Kor. Chem. Soc.* **2005**, *26*, 952 – 958.
[38] T. Scior, J. L. Medina-Franco, Q. T. Do, K. Martinez-Mayorga, J. A. Yunes Rojas, P. Bernard, *Curr. Med. Chem.* **2009**, *16*, 4297 – 4313.
[39] A. Jakalian, D. B. Jack, C. I. Bayly, *J. Comput. Chem.* **2002**, *23*, 1623 – 1641.
[40] R. R. Mittal, L. Harris, R. A. McKinnon. M. J. Sorich, *J. Chem. Inf. Model.* **2009**, *49*, 704 – 709.
[41] B. Lygo, B. Allbutt, S. R. James, *Tetrahedron Lett.* **2003**, *44*, 5629 – 5632.

Received: December 6, 2010
Accepted: September 28, 2011
Published online: January 11, 2012

