QSAR modeling: a new open source computational package to generate and validate QSAR models

*João Paulo A. Martins and Márcia M. C. Ferreira**

Laboratory for Theoretical and Applied Chemometrics, Institute of Chemistry, University of Campinas - UNICAMP, Campinas, SP 13084-971, P.O.B. 6154, Brazil

* To whom the correspondence should be addressed. E-mail: marcia@iqm.unicamp.br

Abstract. *QSAR modeling* is a novel computer program developed to generate and validate QSAR or QSPR (quantitative structure- activity or property relationships) models. With *QSAR modeling*, users can build partial least squares (PLS) regression models, perform variable selection with the ordered predictors selection (OPS) algorithm, and validate models by using **y**-randomization and leave-*N*-out cross validation. An additional new feature is outlier detection carried out by simultaneous comparison of sample leverage with the respective studentized residuals. The program was developed using Java version 6, and runs on any operating system that supports the Java Runtime Environment version 6. The use of the program is illustrated. This program is available for download at lqta.iqm.unicamp.br.

Keywords: QSAR models, PLS regression, OPS variable selection, **y**-randomization, leave-*N*-out cross validation, outliers detection.

Introduction

The study of quantitative relationships between chemical structure and biological activity or a physicochemical property (QSAR/QSPR) is an important research field nowadays. For example, QSPR studies are of great help in predicting physicochemical properties that are difficult to be obtained experimentally. Regarding theoretical medicinal chemistry, the prediction of biological activities of new compounds using mathematical relationships based on structural, physicochemical and conformational properties of previously tested potential agents is an area of intense research. QSAR relationships are helpful to understand and explain the mechanism of drug action at molecular level and allow the design and development of new compounds with desirable biological properties [1].

A QSAR (or QSPR)[1] relationship is expressed through an equation that relates the properties of the investigated compounds to their biological activity, with sufficient statistical significance. This equation must have not only a good predictive power, but it should be validated to show its robustness and that was not obtained by chance [2-7].

There are several programs available in the literature to generate and validate QSAR models. Among them, the most well-known are: MobyDigs[8] BuildQSAR[9], VCCLAB[10,11], QSAR+[12], BILIN[13], MOLGEN QSPR [14], CORAL[15], CODESSA PRO[16] and WOLF [17]. Table 1 shows a comparison between *QSAR modeling* and these programs. Among the free programs, *QSAR Modeling* is the only one that incorporates all the validation tests suggested in literature[3] to generate models that are robust, that do not suffer form chance correlation and in which all the compounds are inside the applicability domain (no outliers or atypical compounds in the model).

[1] From here on, we will refer only to QSAR, but the same procedures are applied to QSPR studies

Tabela 1. Comparison between the main features of *QSAR modeling* and other programs available in literature.

Program	Robustness[a]	Test for chance correlation[b]	Outlier detection	Free program
MobyDigs	No	Yes	No	No
BuildQSAR	No	No	Yes	Yes
VCCLAB	No	No	No	Yes
QSAR+	No	Yes	Yes	No
BILIN	No	No	No	Yes
MOLGEN QSPR	No	Yes	No	No
CORAL	No	No	No	Yes
CODESSA PRO	No	Yes	Yes	No
WOLF	No	No	Yes	No
QSAR Modeling	**Yes**	**Yes**	**Yes**	**Yes**

[a]Leave *N* out Cross validation [b]**y** randomization

The purpose of this work is to introduce the new open source computer program, named *QSAR modeling*, that is able to build and validate QSAR models according to basic chemometrical principles. This is the first QSAR program that implements the newly developed variable selection ordered predictors selection (OPS) algorithm [18], the validation procedures of leave-*N*-out cross validation and **y**-randomization, besides outlier detection. Outlier detection features, frequently neglected in QSAR programs, is implement here by combining the samples leverage and their studentized residuals, which is a common procedure in chemometrics and not implemented in any of the free programs from Table 1. Outlier detection is implemented in BuildQSAR by using the residual standard deviation.

The process of model building using *QSAR modeling* is described using a data set formed by 37 polycyclic aromatic hydrocarbons (PAH), having their log P (octanol-water partition coefficient) as dependent variable[19]. Besides the descriptors available in reference 19, topological descriptors were calculated by using the program DRAGON 6.[20]

Methodology

QSAR modeling was coded in Java 6 language [21] and has an object oriented structure. It was designed to work under any operating system having Java Virtual Machine (JVM) available (Windows XP, Windows Vista, Windows 7, Linux, Mac OS X, Solaris, among others). It is necessary to have the Java Runtime Environment (JRE) version 6 installed in the operating system to run *QSAR modeling*.

Results and Discussion

QSAR modeling program requires as input, two text files containing a matrix with the numerical values of the descriptors (**X** matrix formed by *I* rows and *J* columns) and a vector of biological activities (**y** vector consisting of *I* elements) for the *I* compounds under investigation. In the file containing the descriptors the user can, optionally, add the name of each descriptor in the first row. The main screen of the program showing a data set is available in the supplementary material (Figure 1S).

QSAR modeling program integrates the following tools:

1. Data pre-processing
2. Variable selection – OPS algorithm
3. Regression modeling - PLS method
4. Outlier detection – Leverage and studantized residuals
5. Model validation - leave-*N*-out cross validation and **y**-randomization tests

Data pre-processing

Pre-processing the data is a common procedure when building QSAR models. If variables are of various nature having distinct units or if they present different orders of magnitude and variances, as frequently happens in QSAR studies, it is recommended to preprocess the data. The

standard procedure consists of auto-scaling the matrix **X** and the **y** vector. This corresponds to assigning each variable with the same weight, minimizing dominant variables' influence in further calculations. The mean and standard deviation (SD) of each column of **X** and of vector **y** are obtained. The mean is subtracted from each column values and thereafter divided by the corresponding SD (expression 1).

$$x_{ij(as)} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (1)$$

where x_{ij} and $x_{ij(as)}$ are the values of j^{th} variable, from i^{th} compound before and after auto-scaling, respectively; $\bar{x}_j$ is the mean value of j^{th} variable and s_j its standard deviation. This pre-treatment is applied to the **X** matrix of descriptors and the vector **y** containing the biological activities. Auto-scaling is less sensitive to outliers and autoscaled variables have a mean of zero and unity variance (standard deviation).

In a few cases, mean-centering is preferred to auto-scaling, which is accomplished by subtracting the mean of a given variable from all of its elements. QSAR modeling has implemented both preprocessing types.

Model building with PLS regression method

The mathematical models used in QSAR are often obtained through a linear regression [1,2,22,23] between the descriptors matrix and the biological activities vector. Usually, this regression is performed in three different ways: *i*) multiple linear regression (MLR); *ii*) principal components regression (PCR); and *iii*) partial least squares regression (PLS).

Historically, multivariate linear regression was first performed using MLR, which has always worked well because the number of descriptors was smaller than the number of samples. Nowadays, when using MLR in QSAR studies, it is common to use one descriptor for at least 5 or 6 molecules, and it is assumed that the correlation between descriptors is not high (> 0.7). However, modern modeling programs used in QSAR studies generate thousands of descriptors which are highly intercorrelated, especially in 3D- and 4D-QSAR analysis [24-27]. Thus, MLR cannot be

used in such cases, unless a careful variable selection is carried out. To avoid these problems, good alternatives are to use projection methods such as Principal Components Regression (PCR) or Partial Least Squares (PLS) [22,28,29]. When applying these methods, the number of descriptors and the correlations among them are no longer a problem. Between PLS and PCR, the former became more popular in QSAR studies and it is the regression method implemented in *QSAR modeling*. Although PLS gives similar results to those from PCR method, it usually yields more parsimonious regression models, i.e., models with fewer factors while still retaining a good fit.

The optimum number of latent variables (LV) in the PLS model is determined by internal cross validation. This methodology is applied because projection methods are biased and it is desirable to avoid overfitted models. In cross validation, the data set is split into a certain amount of groups (of size *N*) and several models are built always leaving one of these groups out of the model building. The regression model obtained is used to predict the dependent variable (biological activity or physicochemical property) of samples left out from the analysis. This process is repeated until all samples have been excluded once. This strategy, called leave-*N*-out (leave-many-out) cross validation, is very important to have some insight about the predictability and robustness of the model and is always used in QSAR studies. It is common to use *N* equal to 1, leading to leave-one-out cross validation, when defining the number of LV in the PLS model..

QSAR modeling provides, as a result from cross validation, tables containing the values of the statistical parameters listed in Table 2, the regression coefficients of the PLS model (*b(j)* for *j* = 1, 2, ..., *J*), the predicted values for the dependent variable in cross validation ($\hat{y}_{cv}(i)$ for *i* = 1, 2, ..., *I*) and the predicted values of the dependent variable in the model ($\hat{y}_{cal}(i)$ for *i* = 1, 2, ..., *m*).

Table 2. Statistical Parameters calculated by the program *QSAR modeling*

Parameter	Symbol	Equation[a]
Prediction error sum of squares of cross validation	*PRESScv*	$\sum_{i=1}^{I}[y(i) - \hat{y}_{cv}(i)]^2$

Prediction error sum of squares of calibration	*PRESScal*	$\sum_{i=1}^{I}[y(i)-\hat{y}_{cal}(i)]^2$
Pearson correlation coefficient of cross validation	r_{cv}	$\frac{\sum_{i=1}^{I}[y(i)-\bar{y}]\times[\hat{y}_{cv}(i)-\hat{\bar{y}}_{cv}]}{\sqrt{\sum_{i=1}^{I}[y(i)-\bar{y}]^2}\sqrt{\sum_{i=1}^{I}[y_{cv}(i)-\bar{y}_{cv}]^2}}$
Pearson correlation coefficient of calibration	r_{cal}	$\frac{\sum_{i=1}^{I}[y(i)-\bar{y}]\times[\hat{y}_{cal}(i)-\hat{\bar{y}}_{cal}]}{\sqrt{\sum_{i=1}^{I}[y(i)-\bar{y}]^2}\sqrt{\sum_{i=1}^{I}[y_{cal}(i)-\bar{y}_{cal}]^2}}$
Coefficient of determination of cross validation	Q^2	$1-\frac{PRESScv}{\sum_{i=1}^{I}[y(i)-\bar{y}]^2}$
Coefficient of multiple determination of calibration	R^2	$1-\frac{PRESScal}{\sum_{i=1}^{I}[y(i)-\bar{y}]^2}$
Root mean square error of cross validation	*RMSECV*	$\sqrt{\frac{PRESScv}{I}}$
Root mean square error of calibration	*RMSEC*	$\sqrt{\frac{PRESScal}{I}}$

[a] I is the number of samples in the training set. $\hat{y}_{cv}(i)$ and $\hat{y}_{cal}(i)$ are the predicted values for $y(i)$ in cross validation and in the final model, respectively. $\bar{y}$, $\hat{\bar{y}}_{cv}$ e $\hat{\bar{y}}_{cal}$ are mean values of $y(i)$, $\hat{y}_{cv}(i)$ and $\hat{y}_{cal}(i)$, respectively.

The cross validation procedure, as implemented in *QSAR modeling* program, allows the user to choose the maximum number of latent variables (LV) and the number of samples to be removed when performing cross validation (Figure 2S from supplementary information). Figure 1 shows the results of cross validation provided by *QSAR modeling* for the data set used in this study after variable selection.

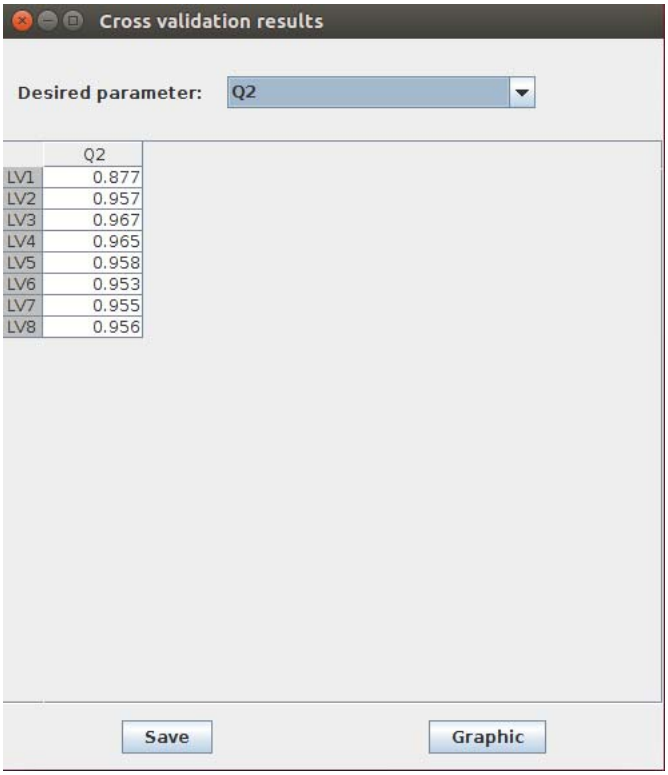

Figure 1. *QSAR modeling* window where the cross validation results are shown. All parameters from Table 2, regression coefficients, predicted values for the dependent variable in cross validation and predicted values of the dependent variable in the model can be seen in this window

Table 3 shows the results from leave-one-out cross validation applied to the data set used in this study after variable selection using *QSAR modeling*. The final PLS model was obtained with eight descriptors and three latent variables (LV).

Table 3. Statistical parameters generated by *QSAR modeling* for the PLS model with 3LV and after variable selection.

Parameter	*PRESScv*	*PRESScal*	r_{cv}	r_{cal}	Q^2	R^2	*RMSECV*	*RMSEC*
Value	1.23	0.74	0.98	0.99	0.97	0.98	0.18	0.14

OPS variable selection

The ordered predictors selection (OPS) algorithm was recently developed to perform variable selection [18] and it has already been successfully used in QSAR/QSPR studies [24,30-34]. The main idea of this algorithm is to attribute an importance to each descriptor based on informative vectors. The columns of the data matrix are reordered in such a way that the most important descriptors are presented in the first columns. Then, successive PLS regressions are built with increasing number of descriptors in order to find the best PLS model, which can be selected according to some of the parameters shown in Table 2.

QSAR modeling implements the OPS algorithm with the following informative vectors: *i*) correlation vector; *ii*) PLS regression vector and *iii*) element-wise product of the two previous vectors. Figure 3S (supplementary material) shows the window of *QSAR modeling* where the user chooses the appropriate options to run the OPS algorithm.

As can be seen in Figure 3S, the program presents the following options to run OPS:

- **Number of latent variables for OPS algorithm** - indicates the number of latent variables in a model when the regression vector is used as informative vector to sort the variables. Different number of latent variables in a regression model lead to distinct regression vectors (informative vectors) and, consequently, the sorting of variables might be affected.
- **Number of latent variables in the model** - indicates the maximum number of latent variables in the models built during the OPS algorithm execution (see reference 18 for more details).
- **Number of samples to be removed during cross validation** - indicates the *N* value in the leave-*N*-out procedure.
- **Window** - indicates the initial number of sorted descriptors in the matrix analyzed by the OPS algorithm.
- **Increment** - indicates the number of sorted descriptors added to the matrix analyzed in each

step by the OPS algorithm.

- **Percentage of variables** - indicates the fraction of the descriptors to be analyzed by the OPS algorithm.
- **Vector** - indicates the informative vector to be used to sort the descriptors.
- **Criterion to classify the model** - indicates the statistical parameter to be used to evaluate the quality of the model.

As output from variable selection with OPS, *QSAR modeling* program provides a table listing the best models obtained with the algorithm. It is possible to select one of the listed models and perform all the validation tests available in the program. Besides, it is possible to save the table with the descriptors selected for further analysis. Figure 2 shows the *QSAR modeling* window that presents the OPS results.

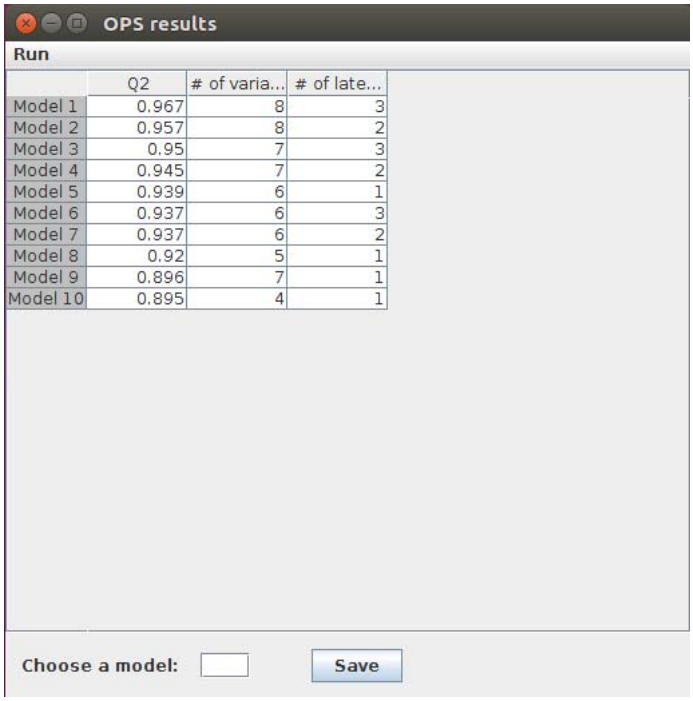

Figure 2. Output from OPS algorithm. In the columns are the values of the parameter chosen to evaluate the models, the number of variables selected and the number of latent variables for the best ten models.

To illustrate the use of the OPS algorithm with *QSAR modeling*, a data set consisting of 37 compounds and 407 descriptors was chosen. Electronic, steric, topological and electrotopological descriptors were used in this example. Such descriptors are of different nature and so, autoscaling (eq. 1) was the preprocessing applied prior to data analysis (descriptors and biological activities were both autoscaled). A correlation cutoff, also available in *QSAR modeling* program, was applied before the first run of OPS algorithm. Descriptors presenting Pearson correlation coefficient with the biological activity lower than 0.3, were eliminated from the pool and the number of descriptors decreased from 407 to 305. This new matrix (37x305) was submitted to the OPS algorithm and the best model was obtained with 15 descriptors, 3 latent variables and a Q^2 value of 0.959. In order to have a less complex model, a new variable selection was carried out by applying the OPS algorithm to this matrix containing the 15 descriptors selected above. The results obtained for the final model in this second run are shown in Table 3 (8 descriptors, 3 latent variables and Q^2 = 0,967).

Outliers detection

In order to verify the quality of the training set being used to build the QSAR model, the homogeneity of the set should be assured by a standard procedure for outliers detection. Compounds structurally different from their counterparts in the training set or with an atypical value of the measured biological activity do not belong to the applicability domain and should be removed from the training set before building the final model. A common procedure in Chemometrics to detect outliers in the training set is to use the leverage and the studentized residuals.[2, 22,35] The leverage measures the influence of a sample in the regression model while the studentized residual (a standardized residual) represents the difference between the experimental value of the biological activity and the value predicted by the model, divided by the residual standard deviation. The advantage of using this residual definition is that it has zero mean and unity standard deviation.

The outliers detection performed by *QSAR modeling* allows the user to choose the number of latent variables to be used in the PLS model and the results are given in a table containing the values of the leverage and studentized residuals for the compounds in the training set (supplementary information, Figure 4S). Samples with leverage higher than $3k/I$, where k is the number of LV in the model and I is the number of samples, can be considered suspicious and should be analyzed carefully [2,35]. Regarding the studentized residuals, they should be randomly scattered around the origin indicating that they follow a normal distribution. Assuming that they are normally distributed at 95% confidence level (α = 0.05), the critical value of a bilateral t test is 1.96, when the residuals are limited to the ± 1.96 interval (in general the interval ± 2.0 is used). The studentized residuals are measured in standard deviation units and values higher than 2.0 or lower than -2.0 can be already considered as statistically significant.

Samples presenting, simultaneously, leverage and studantized residuals above the limits indicated above are atypical and should be excluded from the data set.

QSAR modeling was used to check for outliers in the model, after variable selection performed by OPS algorithm. The resulting leverage and studantized residuals are shown in Figure 3. It can be observed that there are no compounds presenting simultaneously leverage and studantized residuals outside the limits recommended in the literature. However, compound 10 presents high leverage value compared to other compounds and can be characterized as an outlier. Besides, the residuals from compounds 2 and 23 are slightly below the lower limit. These two compounds can be temporarily excluded from the data, a new model is built and the improvements are evaluated. If being significant they are eliminated from the data otherwise, they stay in the model. Sample 2 has low leverage and so won't cause significant changes in the regression vector. On the other hand, the leverage from compound 23 is significant. With the exclusion of the three compounds, Q^2 increase from 0.97 to 0.98. The high residuals observed for compounds 2 and 23 can be an indication of uncertainty in the experimental measurements. It should be stressed that the elimination of samples must be done with great care and justified from the chemical or biological

point of view.

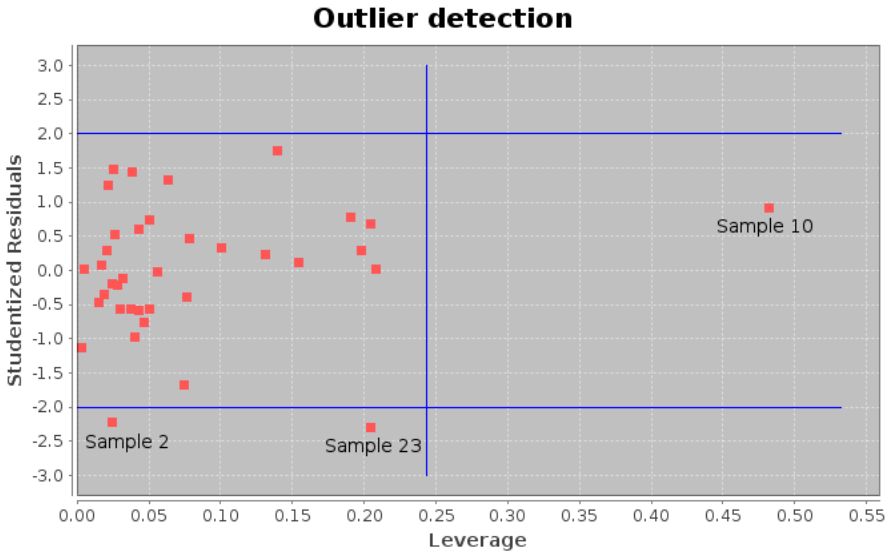

Figure 3. Plot of leverage *versus* studantized residuals, used for outlier detection. Blue lines indicate the limits suggested in the literature.

Leave-N-out cross validation

If the leave-*N*-out cross validation process is repeated several times for different integers *N*, different values of coefficient of multiple determination of cross validation (Q^2) will be obtained. In addition, for the same value of *N* (assuming it is not equal to one), different runs of leave-*N*-out procedure will also lead to distinct Q^2 values, because the order of the samples is randomized prior to splitting the data in the cross validation procedure.

However, these Q^2 values should not be significantly different from each other. Since the QSAR model is built to predict the activities of new compounds, it should not be sensitive to samples removed during the cross validation procedure. Thus, to evaluate the robustness of the model, it is highly recommended to perform repeated leave-*N*-out cross validation runs for different values of *N* (varying from 2 to 20% - 30% of the number of compounds).[7]

The robustness of the model can be accessed by leave-*N*-out process using *QSAR modeling*. In this step, it is possible to choose the maximum number of samples to be removed for cross

validation, the number of latent variables in the model, which is kept constant during the validation, as well as the number of repetitions in each validation for each number of samples removed (Figure 5S from supplementary information). The output of this test is a table containing the values of *RMSECV* or Q^2 depending on the users' choice. Figure 4 shows the results of Q^2 from this validation test for a model with 3 LV, 3 repetitions and a maximum of 10 samples being removed.

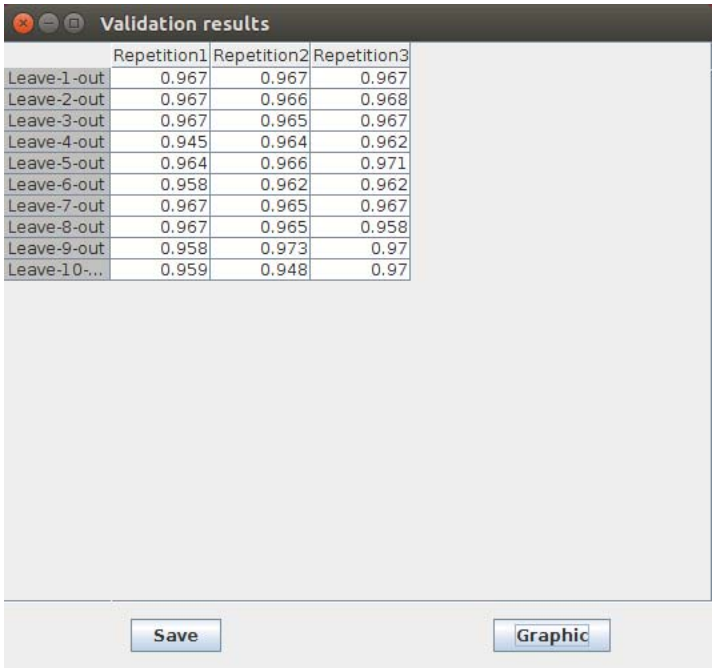

Figure 4. Results obtained from leave-*N*-out cross validation by *QSAR modeling*

The regression model obtained, after variable selection by OPS algorithm, was submitted to leave-*N*-out validation procedures and the graphical results are presented in Figure 5. As can be seen, the model can be considered robust, since small fluctuations in Q^2 values are observed up to 10 samples removed. For each value of *N* the procedure was repeated three times (triplicate).

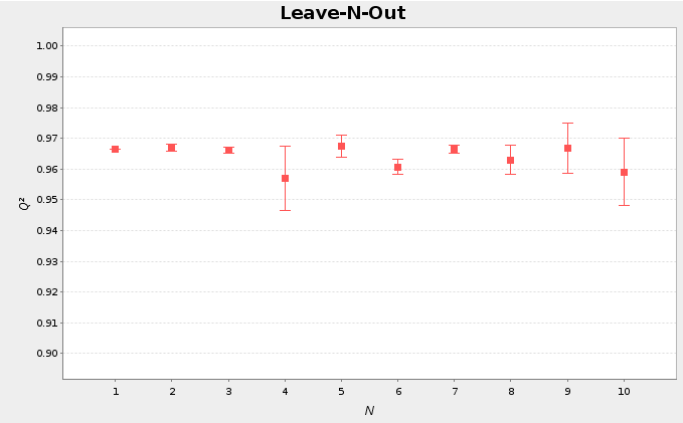

Figure 5. Mean and standard deviation values of Q^2 obtained from Leave-*N*-out validation test applied to the PLS model built after variable selection by OPS.

__y__-randomization

The purpose of the **y**-randomization test is to detect and quantify chance correlations between the dependent variable and descriptors [2,5-7]. To obtain an estimate of the significance of a Q^2 value obtained for a given model, parallel models should be built with the biological activity values (**y** vector) permuted among the molecules while the original descriptors (from **X** matrix) are kept fixed. Thus, the Q^2 value of the real model must be much greater than the values obtained for the parallel models to assure that the real model was not obtained by chance.

Performing **y**-randomization test with the *QSAR modeling* program, it is possible to choose the number of randomizations to be performed in this validation step (Figure 6S from supplementary information). The program provides as result a table containing the R^2 and Q^2 values calculated for the models obtained with the biological activity shuffled and the Pearson correlation coefficient ($r(\mathbf{y}_{al},\mathbf{y})$) between the real and the scrambled **y** for each model (Figure 6). The last row in this table contains the R^2 and Q^2 values for the real model, so it can be compared to those from parallel models.

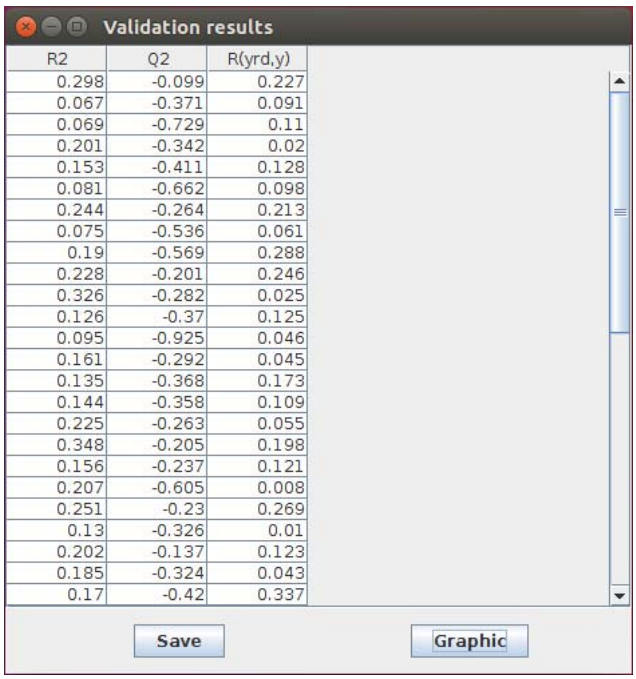

Figure 6. Results from **y**-randomization test provided by *QSAR modeling*

The model obtained after variable selection by OPS method was submitted to **y**-randomization test considering 50 randomizations for **y** and excluding one sample at time (leave-one-out). The results are presented in Figure 7. As can be seen, all R^2 and Q^2 values of the models obtained with vector **y** scrambled, $\mathbf{y}_{al}$, are lower than 0.4 and 0.0, respectively,[2] confirming that the real model was not obtained by chance.[7]

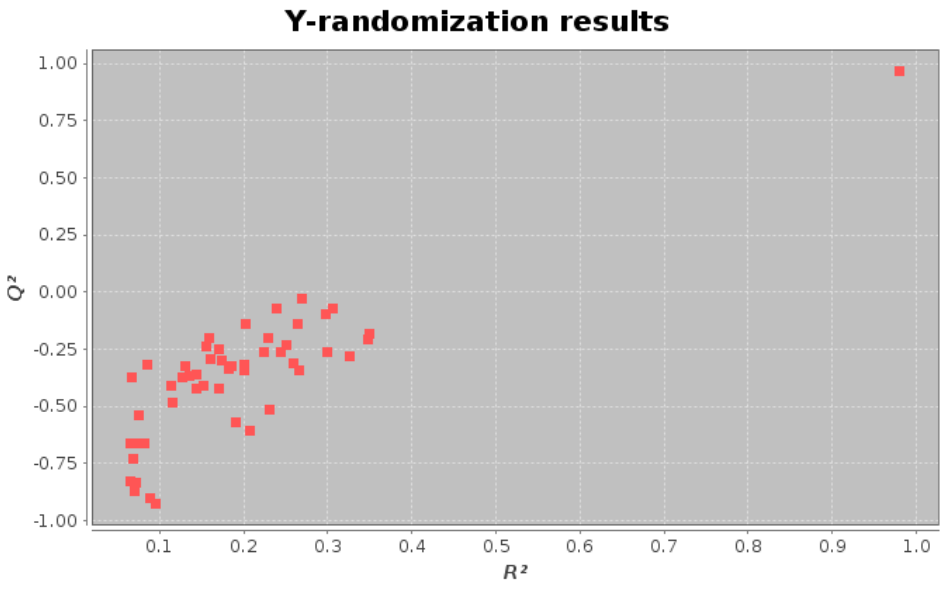

Figure 7: Graphical representation of R^2 and Q^2 from models obtained with **y** scrambled. The distant point stands for R^2 and Q^2 values of the real model.

Software comparison

With the aim to evaluate the performance of *QSAR modeling* in building and validating models, the same data set was used to build QSAR models by using some of the programs cited in Table 1.

The VCCLAB,[11], a free program that can be downloaded from the WEB, was used to build PLS regression model. In this program, variable selection is performed in two steps: i) descriptors with small variance are eliminated; ii) the remaining descriptors are selected by using genetic algorithm based on Q^2 values. The final model containing 190 descriptors and 2 LV, presented Q^2 = 0. 963. Even though Q^2 is similar to that from *QSAR modeling* (0,967), it is impossible to give any physical interpretation to this model due to the excessively high number of descriptors. Besides, this PLS model is underfitted; the projection of 190 descriptors e a 2D space spanned by 2LV could lead to loss of relevant information. Unfortunately, the number of LV is selected automatically and there is no option to change it.

BuildQSAR[9] is another free program and was used to build a MLR regression model. The program allows the use of principal component regression (PCR), but when making variable selection using systematic search or genetic algorithm, the only option in the program is to build MLR models. The final model with 7 descriptors and no outlier detected presented Q^2 = 0. 963. The data matrix with these 7 descriptors selected used *QSAR modeling* to validate their model (PLS model with 7LV) but the proposed model was not robust and suffers from chance correlation (did not pass in both tests).

The program Wolf[17] was used to build a MLR model and using genetic algorithm to carry out variable selection. Sample 23 was identified as an outlier and the final model with 5 descriptors and Q^2 = 0. 961 (also inferior to that from *QSAR modeling*) was tested in *QSAR modeling* for model validation. This models was robust but did not pass in the **y**-randomization test.

Conclusions

The *QSAR modeling* program allows building QSAR or QSPR models in a simple and fast way. Besides, it joins in a single program a newly developed variable selection algorithm to build PLS models, a procedure to detect outliers and the main validation procedures of QSAR models demanded by the scientific community.

A data set was used to illustrate the use of all tools provided by *QSAR modeling* and the results were superior to those obtained by the other programs used for comparison. Besides, several of the functionalities incorporated in *QSAR modeling*, are not available in other programs.

Being a free open source program, *QSAR modeling* is a new QSAR tool available to everyone, and as such it can be enhanced for the needs in a variety of research fields.

Acknowledgments

The authors thank FAPESP and CNPq for financial support and, Dr. Kerly F. M. Pasqualoto from Laboratório de Bioquímica e Biofísica - Instituto Butantan for the help in building the QSAR model using the program Wolf and Dr. Eduardo Borges de Melo for his help in using the BuildQSAR program.

References

1. Ferreira, M. M. C., Multivariate QSAR. *J. Braz. Chem. Soc.* **2002**, *13*(6), 742-753.
2. Ferreira, M.M.C.; Kiralj, R. Em *Química Medicinal, Métodos e Fundamentos em Planejamento de Fármacos*; Montanari, C., ed.; EDUSP, 2011.3, chapter. 12.
3. Guidance Document on the Validation of (Quantitative) OECD Environment Health and Safety Publications Series on Testing and Assessment No. 69. OECD: Paris, 2007. http://www.oecd.org/dataoecd/55/35/38130292.pdf (accessed Jul 07, 2009).
4. Gramatica, P., Principles of QSAR models validation: internal and external. *QSAR & Comb.*

Sci. **2007**, *26*, 694-701.

5. Tropsha, A.; Gramatica, P.; Gombar, V. K., The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Comb. Sci.* **2003**, *22*, 69-77.
6. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P., Methods for Reliability and Uncertainty Assesment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environmental Health Perspectives.* **2003**, *111*, 1361-1375.
7. Kiralj, R.; Ferreira, M. M. C., Basic validation procedures for regression models in QSAR and QSPR studies: theory and applications. *J. Braz. Chem. Soc.* **2009**, *20*, 770-787.
8. MobyDigs, Version 1-2004, Talete srl, Milano, Italy.
9. de Oliveira, D. B.; Gaudio, A. C., BuildQSAR: A New Computer Program for QSAR Analysis. *Quant. Struct.-Act. Relat.* **2000**, *19*, 599-601.
10. 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.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description, *J. Comput. Aid. Mol. Des.*, **2005**, 19, 453-63.
11. VCCLAB, Virtual Computational Chemistry Laboratory, 2005. http://www.vcclab.org, (accessed Oct 26, 2011).
12. Cerius² QSAR+, 2000. http://www.scripps.edu/rc/softwaredocs/msi/cerius45/qsar/QSAR40TOC.html, (accessed Oct 26, 2010).
13. Bilinear Model, BILIN, 1976. http://www.kubinyi.de/bilin-program.html, (accessed Oct 26, 2010).
14. Molecular Structure Generation MOLGEN QSPR, 2003. http://www.molgen.de/?src=documents/molgenqspr.html (accessed Oct 26, 2011).
15. Correlation and Logic, CORAL, 2010. http://www.insilico.eu/coral/ (accessed Oct 26, 2011).
16. Comprehensive Descriptors for Structural and Statistical Analysis, CODESSA PRO, 2001. http://www.codessa-pro.com/index.htm (accessed Oct 26, 2011).
17. D. Rogers, WOLF Reference Manual Version 5.5, The Chem21 Group Inc., Chicago, IL 1994.

18. Teófilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C., Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J. Chemometr.* **2009**, *23*, 32-48.
19. http://www.moleculardescriptors.eu/dataset/dataset.htm (accessed May 4, 2012).
20. DRAGON, Version 6-2010, Talete srl, Milano, Italy.
21. *Java*, version 6 update 10; java development kit; Sun microsystems, Inc: Santa Clara, CA 95054 USA, 2008.
22. Martens, H.; Naes, T., *Multivariate Calibration*. Ed. Chichester, Wiley, 1989.
23. Beebe, K. R.; Pell, R. J.; Seasholtz, M. B., *Chemometrics: A Practical Guide*. Wiley, 1989.
24. Martins, J. P. A.; Barbosa, E. G.; Pasqualoto, K. F. M.; Ferreira, M. M. C., LQTA-QSAR: a new 4D-QSAR methodology. *J. Chem. Inf. Model.*, **2009**, *49*(6), 1428–1436.
25. Nilsson, J.; de Jong, S.; Smilde, A. K., Multiway calibration in 3D QSAR. *J. Chemometr.* **1997**, *11*(6), 511-524.
26. Cramer, R. D.; Patterson, D. E.; Bunce, J. D., Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*(18), 5959-5967.
27. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C., Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*(43), 10509-10524.
28. Agnar, H., PLS regression methods. *J. Chemometr.* **1988**, *2*(3), 211-228.
29. Geladi, P.; Kowalski, B. R.; Partial least-squares regression - A tutorial. *Anal. Chim. Acta.* **1986**, *185*, 1-17.
30. de Melo, E. B.; Ferreira, M. M. C., Multivariate QSAR study of 4,5-dihydroxypyrimidine carboxamides as HIV-1 integrase inhibitors. *Eur. J. Med. Chem.* **2009**, *44*(9), 3577-3583.
31. Teófilo, R. F.; Kiralj, R.; Ceragioli, H. J.; Peterlevitz, A. C.; Baranauskas, V.; Kubota, L. T.; Ferreira, M. M. C., QSPR Study of Passivation by Phenolic Compounds at Platinum and

Boron-Doped Diamond Electrodes. *J. Eletrochem. Soc.* **2008**, *155*(10), D640-D650.

32. Hernández, N.; Kiralj, R.; Ferreira, M. M. C.; Talavera, I., Critical comparative analysis, validation and interpretation of SVM and PLS regression models in a QSAR study on HIV-1 protease inhibitors. *Chemom. Intell. Lab. Syst.* **2009**, *98*(1), 65-77 .
33. de Melo, E. B.; Ferreira, M. M. C., *J. .Chem. Inf. Model*. **1012**, *52*, 1722.
34. Barbosa, E. G.; Pasqualoto, K. F. M.; Ferreira, M. M. C., J. Comput.-Aided Mol. Des. (no prelo).
35. Ferreira, M. M. C.; Antunes, A. M.; Melgo, M. S.; Volpe, P. L. O., Quimiometria I: Calibração multivariada, um tutorial. *Quim. Nova.* **1999**, *22*(5), 724-731.