J. Chil. Chem. Soc., 53, Nº 3 (2008)

INTERNAL TEST SET (ITS) METHOD: A NEW CROSS-VALIDATION TECHNIQUE TO ASSESS THE PREDICTIVE CAPABILITY OF QSAR MODELS. APPLICATION TO A BENCHMARK SET OF STEROIDS

EMILI BESALÚ[*,1] *AND LEONEL VERA*[2]

[1]Institute of Computational Chemistry. Universitat de Girona. Campus de Montilivi. 17071 Girona, Spain. e-mail: emili.besalu@udg.edu
[2]Dept. of Chemistry. Universidad Católica del Norte. Avda. Angamos 0610. Casilla 1280. Antofagasta, Chile.
(Received: September 3, 2007 - Accepted: May 31, 2008)

ABSTRACT

A new internal cross-validation method is presented for assessing the true predictive capability of QSAR models. The test is general and can be applied in many QSAR/QSPR approaches. In this work, the method is tested on a well-known benchmark set of steroids. In order to make the calculations, Topological Quantum Similarity Indices and Multiple Linear Regression models were considered.

Keywords: benchmark steroids; cross-validation; predictions; internal test set method; statistical validation; topological quantum similarity indexes.

INTRODUCTION

One of the crucial steps involved in a QSPR/QSAR study consists of the statistical validation of results. Some of the most popular validation techniques are the well-known Leave-one-out (LOO) and Leave-many-out (LMO) cross-validation (CV) procedures[1-4]. A good cross-validation method must be an honest and transparent procedure. Wold[3,4] warned that these simple aspects of paramount importance for method reliability must be well taken care of. In this work an internal validation method inspired by the concept of the external validation test set is described. The method is general and applicable to many QSAR approaches. It can be easily generalized to any LMO cross-validation protocol but, for the sake of simplicity, the algorithm presented here lies within the context of a simpler approach: the LOO procedure.

In its basic formulation, the LOO-CV procedure can be applied when a fixed set of descriptors is considered for a set of n molecules. For each molecule, a model is built considering the experimental property value of the n-1 remaining compounds and the respective descriptors. Then, this model is used to make a prediction for the molecule left out. The above primary algorithm is named here the *kernel* of the standard LOO-CV method. This kernel algorithm can be applied using linear or non-linear methods such as Neural Networks[5,6], among others more sophisticated such as molecular field analysis[7,8]. When the ordinary MLR method[9] is considered, it is not necessary to explicitly construct all the linear and independent models to obtain the n predictions: data can be obtained from a single MLR calculation since specific theorems are described in the context of LOO and LMO protocols[10,11] allowing to readily obtain all the cross-validated property values.

Nevertheless, a crucial point is that the protocol usually followed in a MLR-QSAR study consists in repeating the LOO kernel calculations considering a usually large number of descriptor subsets taken form a large pool of parameters. In this way, the final cross-validated predictions arise from the best descriptor set which gives the most satisfactory result. Generally, the reliability of the study is reflected by a statistical parameter such as $q^{(2)}$, the PRESS statistic or the squared correlation coefficient for the cross-validated properties, r^2_{cv}.[12-18] In the evaluation of these statistical parameters the experimental molecular properties are considered in an indirect fashion. For this reason, the standard LOO/LMO procedures can be understood as a variable selection method with indirect property supervision. This characteristic confers this kind of CV protocols undesirable features, as some of those pointed out by Wold[3,4]. For instance, this author does not recommend selecting variables unless this selection is not associated with the molecular properties. Of course, this situation has some advantages as it helps to choose the number and kind of parameters to be considered in a QSAR study.[6,19,20]

The situation described above is also related to a common problem arising in a QSAR study: overparametrisation. This is so because the final accepted and cross-validated model arises from a large number of tests done in order to provide an optimal statistical indicator. Then, the question is: Is the best result a consequence of chance correlations?[21] It is well known that there are several useful post-treatment techniques to study this problem. The best known is the randomisation test[4,22]. This control procedure can be mainly implemented in two different and extreme ways:[4] the randomisation test is applied only to the final model (this can be understood as a mere model validation) or, more desirable, all the calculations are repeated with the free generation of new models involving descriptors from the entire pool of parameters (full method validation).

The literature has plenty of QSAR results obtained via LOO or LMO protocols. The main idea involved is awareness of the whole structure of the study, especially if the number of descriptors is considerable and if, at some level, variable selection methods are applied. Additionally, if present, the randomization test design has to be checked. Finally, the reported statistical parameters ($q^{(2)}$, r^2_{cv},...) do not directly reflect the eventual predictive performance of the QSAR model against an external molecular test set.

METHOD

The necessity of obtaining reliable QSAR models led the authors to reconsider the LOO/LMO protocol design. The internal test sets (ITS) method proposed here takes the idea from the concept of the external validation test sets. At the end of the numerical process which will be described below, every molecule will bear a single predicted property value (or many of them in the case of a LMO-like protocol) which can be compared to the experimental one. The performance of the methodology can be measured by means of the correlation coefficient of both series of values. We will denote this coefficient r_{ITS}. The aim of this internal validation test is to provide a new statistical parameter. This is so because the r_{ITS} statistic is expected to measure the true predictive power of a given QSAR approach more realistically.

The ITS algorithm is simple and general and has been described elsewhere.[23-26] As shown below, it is very easy to extend this approach to the LMO formalism but, for the sake of simplicity, it will be described within the paradigmatic framework of a LOO protocol. The algorithm, which must be applied on a molecular set, proceeds as follows:

1. Consider n molecules, a pool of indexes describing them and the respective property values.
2. Loop: for each molecule (one at a time) do
2.1. Define an internal test set containing this molecule and construct an internal training set with the n-1 remaining ones.
2.2. Obtain a QSAR CV model from the data available in the internal training set.
2.3. Make a prediction for the internal test molecule using the previous model.
3. Collect all the internal predictions and obtain the correlation coefficient, r_{ITS}, against the experimental values.

For each molecule, the prediction made in step 2.3 is kept as a *definitive* result. That is the crucial ITS method concept and the main difference with respect to the standard LOO/LMO procedures: the prediction for the internal test molecule is definitive because this compound acts as a validation test set temporarily, behaving as a true external object for which a real prediction has

 e-mail: lvera@ucn.cl

to be made. From now on, we will call an *ITS step* the processes involved in the loop 2 of the ITS algorithm described above. Consequently, an ITS step consists of the test molecule selection, the corresponding model construction (involving only the remaining ones), and the respective prediction.

If the algorithm above is followed, in every ITS step the concept of model construction at stage 2.2 can be freely interpreted by using the data related to the *n*-1 internal training molecular set. The model construction can be conceived in a general way. It can involve embedded procedures of variable selection, dimensionality reduction, randomisation tests, other CV techniques, linear or non-linear methods, and so on. In this way, the internal training molecular data can be freely manipulated. On the other side, the key idea is that, in every ITS step, the internal test molecule must act as a true validation object and its information must be temporally hidden to the calculation system, especially in step 2.2.

The important concept is that the ITS procedure embeds an external loop which generates the test molecule (or test molecules in a LMO protocol) and an internal procedure which determines the model to be applied over this test set. If needed, this internal loop (step 2.2 above) may expand the generation of descriptor combinations entering the models, as shown below. This design differs from the classical (and most commonly used) CV procedure, which is equivalent to having an external loop generating combinations of descriptors and an internal loop generating the predictions. This important conceptual difference is explained in references 23 and 26.

In the calculations presented below, a standard MLR-LOO method is considered in step 2.2 to select a model. That is, for every internal test molecule the remaining *n*-1 ones are left apart one at a time and, for each one, models involving information of *n*-2 compounds are generated. This internal standard LOO protocol is followed using, as customary, a quite large set of descriptors. Then, at each stage, the best set of descriptors is selected according to the r_{cv} value (also denoted r_{LOO} or r_{LMO} in the literature). Finally, a linear model involving the *n*-1 molecules is obtained and the prediction for the internal test molecule is made in step 2.3. This process is repeated *n* times, generating at each ITS step a different linear model attached to every internal test molecule.

The process of model and prediction generation can be automatically performed or it can be user-supervised. The last option is preferred because the data generated can help to detect the system instabilities, outliers, representative parameters, optimal number of descriptors to consider or, in general, to grasp useful system information. As the ITS method requires constructing QSAR models when training set molecules are temporally excluded, the procedure is somehow related to the Jackknife technique.[27,28] That is, the process of molecule hiding in step 2.1 above can serve to analyse the procedure behaviour or stability against compound inclusion/exclusion. Another interesting feature of the ITS method is related to the commonly encountered problems of overfitting or the generation of chance correlations. Intrinsically, the researcher who wants to perform an ITS procedure is really encouraged by the method to use good descriptors, to apply an efficient methodology, and to generate reliable QSAR models in step 2.2. If these requirements are not met, the numerical predictions in step 2.3 will be very poor and the final value of r_{ITS} obtained in the last step 3 will be almost zero... or even negative. By using the ITS algorithm the user is automatically penalised if its methodology allows overparametrisation, chance correlations or if the so-called outliers are present in the data. On the other hand, if the protocol is followed and an acceptable result (high value of r_{ITS}) is obtained in step 3, the absence of overfitting should be accepted in the study. Thus, this kind of results will show that the selected QSAR parameters are relevant, that they contain significant information, and that the obtained models have a real predictive power.

The ITS method is a demanding and time-consuming internal validation test because it requires generating true predictions for every test molecule (or for every cross-validated test set in a LMO approach). In fact, despite the ITS method is an internal validation test, in the context of the LOO methodology it can be understood as the generator of a set of *n* artificial external validation processes. The consequence is that the procedure is about *n* times slower than a typical LOO calculation. That is the consequence of computing the r_{ITS} parameter. Nevertheless, this parameter can help us to achieve a very desirable goal: to really measure the performance or efficiency of a QSAR methodology when it is focused on making *real* predictions for an eventual external molecular test set. This is so because the excluded molecules, during the internal test set generation, act as true external compounds for which a real prediction has to be made.

Another consideration must be stated here. Usually, the obtained values of r_{ITS} are small, as shown below. In fact, a bad result can lead to a negative value of this parameter, especially if randomisation tests are performed to validate the whole ITS methodology. For this reason, r_{ITS} values are reported instead of their squares.

RESULTS AND DISCUSSION

The ITS approach has been applied to study the affinity of a series of 31 steroids interacting with the corticosteroid binding globuline (CBG). This steroid series has been widely used as a benchmark set for testing various theoretical QSAR methodologies.[7,8,18,20,29-47] That is the reason why this molecular set was selected for application. For the same reason, the molecular structures are not drawn here. Direct information can be retrieved from the cited literature. For this set, a standard LOO study can be found in reference (47) where a value of r^2_{cv}=0.846 (r_{cv}=0.920) was obtained for a model involving 4 linear descriptors. In both, the present work and the reference, Topological Quantum Similarity Indexes (TQSI) are used as molecular descriptors.[45,47,48-50] Nevertheless, attention is not focused on the source of molecular descriptors but on the concept of the ITS method.

In order to have available data related to predictions over an external test set, the well-known problem consisting in modelling the first 21 steroids (training set) are dealt with and predictions for the remaining 10 ones (true external test set) are made. Concerning this approach, many studies can be found in the literature. Some of the predictions and related data are shown in Table I. Here, the r^2_{cv} and $q^{(2)}$ values are the cross-validated correlation coefficient or the value of $q^{(2)}$ (also denoted q^2 or Q^2 in the literature) obtained when the model is build from the 21 training molecules. For instance, Hahn and Rogers[34] report a value of $q^{(2)}$ or r^2_{cv} equal to 0.628, which was obtained using a receptor surface model (RSM). Klebe et al.[30] obtained a value of 0.665 using CoMSIA. Cramer and co-workers[7], using CoMFA, obtained a value of 0.662 when predictions are made from the mean values coming from different atom probes. Chen et al.[44] reported a value of 0.806, using PARM. Bravi et al.[20] obtained the values 0.605 and 0.729, respectively, by means of MS-WHIM and CoMFA. Jain et al.[33] reported a value of 0.89, using Compass. The available data of Robinson and co-workers[8] using SOMFA method or Similarity Matrix Analysis (SMA) are also tabulated. Robert et al., using Tuned Molecular Quantum Similarity Matrices (TQSAR), report a value of 0.832 for a linear six-parameter model.[46] Finally, there is a column of predictions obtained by Liu and co-workers[19] using MEDV-13, a method based on electrotopological state indexes.

In Table I, RMSE is the root of the mean squared error for the 10 predictions and r_{test} is the correlation coefficient attached to the fitting of these values against the corresponding experimental ones. The series of r_{test} values are duplicated: the first entry considers the 10 predictions and the second one shows how the results are systematically improved if the prediction for the molecule number 31 (2a-methyl-9a-fluorocortisol) is not taken into account in any of the final tests. In the literature this molecule has been described as an outlier or an anomalous compound mainly due to the presence of a fluorine atom.[7,18,29,35,45-47]

Table I: Bibliographic and present results of the property value prediction for the 10 last steroids (external test set). The authors' references are cited in the text. RMSE is the root of the mean squared error. r_{test} is the correlation coefficient between experimental and predicted values. r_{ITS} stands for the correlation coefficient between the 21 predictions arising from ITS algorithm and the experimental values.

Steroid	Exp. activity	*Other references*											*This work*		
		Hahn Rogers	*Klebe et al.*	*Cramer et al.*	*Chen et al.*	*Bravi et al. MS-WHIM*	*Bravi et al. CoMFA*	*Jain et al.*	*Robinson et al. SOMFA*	*Robinson et al. SMA*	*Robert et al.*	*Liu et al.*	*1 desc.*	*2 desc.*	*3 desc.*
22	7.512	7.505	7.40	6.984	7.449	7.300	7.883	7.062	7.279	7.453	7.237	8.166	7.635	7.290	7.775
23	7.553	4.083	7.42	7.764	8.037	8.332	7.430	7.729	7.034	7.022	7.879	7.553	8.836	6.285	7.522
24	6.779	6.575	7.04	6.723	6.601	6.821	6.642	6.462	6.925	6.939	6.648	7.652	6.778	5.492	5.934
25	7.2	6.975	6.65	7.460	6.015	7.445	7.705	7.466	7.232	7.146	7.809	7.765	7.15	6.146	6.853
26	6.144	6.060	6.44	6.156	6.246	6.121	6.495	5.994	5.744	5.908	6.832	5.659	5.007	5.388	5.048
27	6.247	6.720	5.32	7.145	5.742	6.901	6.962	6.383	6.800	7.046	7.318	6.666	7.263	6.014	6.395
28	7.12	6.520	6.89	5.453	6.925	6.532	6.848	6.625	6.603	6.569	7.363	7.340	6.778	6.374	6.581
29	6.817	6.461	6.74	6.958	6.100	6.838	6.816	7.403	6.692	6.850	7.540	5.740	5.836	5.917	5.794
30	7.688	7.070	7.72	7.461	6.108	7.860	7.767	7.741	7.345	7.539	7.628	7.642	8.092	8.478	8.871
31	5.797	6.049	3.78	7.331	5.991	7.491	7.793	7.779	7.283	7.457	7.537	6.845	8.577	7.603	8.429
RMSE		1.153 1.213	0.740 0.395	0.800 0.671	0.709 0.744	0.662 0.411	0.716 0.356	0.705 0.339	0.585 0.367	0.640 0.385	0.762 0.555	0.650 0.590	1.138 0.762	1.015 0.884	1.086 0.735
r^2_{test}		0.003 –	0.757 0.707	0.045 0.172	0.335 0.303	0.277 0.628	0.154 0.657	0.154 0.687	0.192 0.61	0.118 0.504	0.157 0.359	0.441 0.561	0.100 0.575	0.106 0.563	0.141 0.755
r_{test}		0.055 -0.029	0.870 0.841	0.212 0.415	0.579 0.550	0.526 0.792	0.393 0.811	0.392 0.829	0.438 0.781	0.343 0.71	0.396 0.599	0.664 0.749	0.317 0.758	0.326 0.750	0.375 0.869
r^2_{cv} or $q^{(2)}$		0.628	0.665	0.662	0.806	0.605	0.729	0.89			0.832		0.466	0.626	0.692
r_{cv}		0.792	0.815	0.814	0.898	0.778	0.854	0.94			0.912		0.683	0.791	0.832
r^2_{ITS}													0.320 0.456	0.166 0.219	0.039 0.388
r_{ITS}													0.565 0.675	0.407 0.468	0.197 0.623

In Table I, r_{ITS} is the correlation coefficient between the experimental values and the 21 internal predictions arising from the ITS algorithm when applied over the training set. The r_{ITS} values increase if molecule number 1 (aldosterone) is not taken into account (see the second entry for each item). This gives the researcher a clue with respect to the uniformity of the data along the training set, and to the possible classification of molecule 1 as an outlier. The cross-validated predictive coefficient value for the 21 molecules, $q^{(2)}$, is substantially higher than the r_{ITS} one. Nevertheless, the r_{ITS} coefficient will be generally more similar to the r_{test} parameter unless some specific test molecules are removed or any other particular manipulations are performed over the training set. Among others, Shao[51] warned that the classical LOO procedure (and hence, the $q^{(2)}$ coefficient) tends to overestimate the predictive capability of a method. In fact, the direct relationship between the standard correlation coefficient and the one arising from a common LOO calculation was recently demonstrated.[52] The ITS methodology does not overestimate the model predictive ability, as the optimisation of the r_{ITS} coefficient is a task harder than maximizing the standard r_{cv} or $q^{(2)}$ values. This is so because in the ITS methodology the relationship between the predicted and experimental values is very weak. A specific r_{ITS} optimization procedure was not taken into account in this work, that is, all the ITS results presented here were obtained "on the fly" from a single prediction for each molecule and no post-treatment has been done.

The main conclusion is that the new numerical values of r^2_{cv} or $q^{(2)}$ in Table I do not directly reflect the true prediction ability of each methodology. In fact, Table I shows that these parameters are much greater than the corresponding test ones, and a similar behaviour can be found in the literature. For instance, So and co-workers[53] reported several values of the statistic r^2_{test} using different methodologies in the study of a set of 56 steroids (43 training and 13 for test). The values of r^2_{test} ranged from 0.108 up to 0.526 and the range rose up to 0.127–0.610 when some outliers were removed. In the respective trainings the values of r^2_{cv} or $q^{(2)}$ were 0.590–0.880, substantially higher quantities and closer to the training performances.

The most valuable parameter appearing in Table I is the correlation coefficient, r_{test}, because it arises from the predictions over the legitimate external molecular test set. In general, this value is substantially lower than the corresponding r_{cv} one, as shown in Table I. Thus, the following chain of inequalities must be expected when a molecular set is treated using the same protocol and molecular descriptors:

$$r_{fitting} > r_{cv} > r_{ITS} \approx r_{test} \qquad (1)$$

In equation (1), r_{cv} can also be interpreted as the square root of $q^{(2)}$. In fact, the proposed correlation coefficient, r_{ITS}, belongs to the same category of cross-validation parameters (because it is obtained from a cross-validation procedure among the training molecules) but, by construction, this quantity is expected to be lower than the other commonly used cross-validation performance indicators above. Consequently, the r_{ITS} value should be closer to the r_{test} one. For instance, Table I shows the data related to the three models presented in this study. The respective inequalities in (1) for the model involving only one descriptor (fitting and making a prediction for all the 10 test molecules) are 0.740 > 0.683 > 0.565 > 0.317. Another (quite opposite) case is the one reported by Hahn and Rogers: despite the value of $q^{(2)}$ is high (0.628), the value for r_{test} is negative. The corresponding r_{ITS} value (not computed by the author) is expected to be very small and, in this case, it could *a priori* reveal the potential bad methodology performance against a real external test set.

The values of r_{test} in Table I can also be compared to other common values of r^2_{cv} or $q^{(2)}$ given by several authors for the training set of 21 steroids: Oprea et al.[32] reported a value of 0.70 (MTD method); Silverman and Platt[31], using CoMMA and 3 components, obtained 0.828 and 0.674 with 2 components; Kellogg and co-workers[36] obtained 0.803 (3 components using the electrotopological state approach); Turner et al.[40] obtained 0.83 (2 components and EVA method); Tominaga and Fujiwara[43], with 3 descriptors, reached the value of 0.807. Finally, in Amat et al. work[18], values of r^2_{cv} of the order of 0.842 are given. Generally speaking, these researchers focused their attention on maximising the r^2_{cv} or $q^{(2)}$ values, but this does not warrant the maximisation of the r_{test} value, the ultimate relevant parameter. Although the same is valid for the r_{ITS} statistic (its maximisation is not equivalent to optimising the r_{test} parameter) the maximisation of the r_{ITS} parameter is expected to be more correlated to the optimisation of the r_{test} one.

As stated above, three multilinear models are considered in this work. Each model involves, respectively, 1, 2 or 3 independent TQSI descriptors

selected from a collection of 119. The corresponding values of r_{ITS} shown in Table I are 0.565, 0.407 and 0.197, respectively (the model with 4 descriptors produced an even worse result than 0.197). In each case, the molecule number 1 (aldosterone) has an attached bad prediction. This is shown in Figure 1 where, for the set of 21 training molecules, the predicted activities by the ITS models involving one descriptor are represented against the experimental ones. The figure reveals several instabilities, but it is necessary to recall that every point in the graph represents a *true* external test prediction. In the three linear models presented here, if aldosterose is not considered, the values of r_{ITS} systematically increase to 0.675, 0.468 and 0.623, respectively (see the second entry in Table I). As stated above, the correlation coefficient r_{ITS} is statistically expected to be smaller than the corresponding r_{cv} parameter values (r_{cv} is 0.683, 0.791 and 0.832, for each respective model), but the r_{ITS} value shows the predictive ability more realistically. For instance, in constructing linear models involving TQSI, the value of r_{cv} increases up to 0.932 if 4 descriptors are considered, or even up to 0.940 for a model involving 5 descriptors. Whereas, the ITS methodology reveals that, if no other information is available, the most predictive option is to consider a smaller number of terms in the linear model. This does not seem to be a particularity of this study. The related results concerning other molecular families will be published elsewhere.

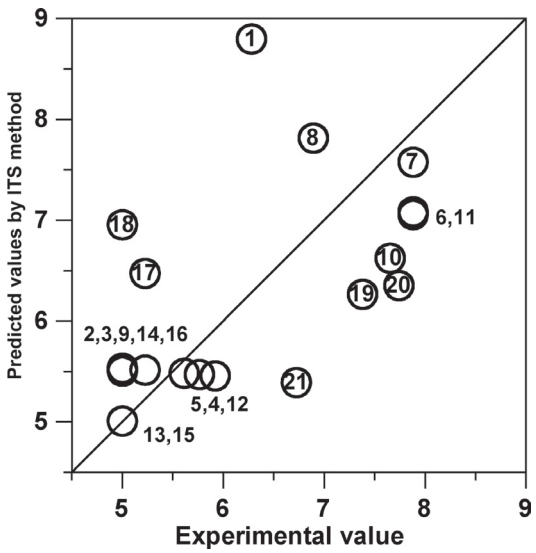

Figure 1 Internally predicted activities by ITS algorithm against the experimental ones for the set of 21 training molecules. The result was obtained using TQSI as descriptors and one-descriptor MLR models.

In the ITS algorithm context, the linear models generated to predict the property for every test molecule are different. This confers the method another characteristic to be checked and evaluated. Regularity of appearance is desired among the indexes selected to construct the models, whereas the presence of fluctuations is more suspicious. For instance, a desirable regularity among the predictive models is found in this study because, in the 21 internal predictive models, the same index (except in two cases) appears in the model of 1 descriptor: the topological connectivity index[54] of order 0, denoted here as ${}^{1}\chi_0(T)$. This index is used to build the final QSAR training model. For this particular case and due to the index simplicity, the model fitting and predictions can be easily reproduced.[55] Other kinds of TQSI connectivity indexes[45,47-50] play an important role for the models of 2 and 3 descriptors and are regularly selected. The final training models are the following:

$pK = 0.525644\ {}^{1}\chi_0(T) - 2.30274$

n=21, r^2_{fit}=0.548, r^2_{cv}=0.466, F=16.58, p=0.0007, $q^{(2)}$= 0.462, PRESS=14.878, r_{ITS}=0.565

$pK = 20.4893\ {}^{8}\chi_p(C) - 23.6851\ {}^{9}\chi_p(C) - 1.31793$

n=21, r^2_{fit}=0.682, r^2_{cv}=0.626, F=31.82, p<0.0001, $q^{(2)}$=0.624, PRESS=10.388, r_{ITS}=0.407

$pK = 0.451322\ {}^{1}\chi_p(C) + 14.4145\ {}^{8}\chi_p(C) - 17.2569\ {}^{9}\chi_1(C) - 3.04739$

n=21, r^2_{fit}=0.769, r^2_{cv}=0.692, F=42.69, p<0.0001, $q^{(2)}$=0.69, PRESS=8.583, r_{ITS}=0.197

Finally, randomisation tests are considered as a tool to *a priori* select the best among the three previous models. Figure 2 presents the obtained values of r_{ITS} when 120 random calculations (artificially exchanging molecular property values) are performed in constructing models involving 1 (a), 2 (b) or 3 (c) descriptors. For each drawn point, an automated full ITS study was made and the system had full freedom to re-select descriptors from the entire data set from the beginning. Hence, this is not a final model test (as mentioned above) but a methodology test, a much more desirable procedure for checking purposes. In each graph, the random points fit a gaussian distribution also graphically represented. The horizontal segments appearing in the gaussian curve signal the position of the points lying to 1, 2 or 3 standard deviation values from the centred mean one (almost zero in all the cases). The upper horizontal line signals the correct value of r_{ITS}. The standardisation of this value according to the respective gaussian distribution gives a measure of the randomness of the deterministic result. The percentages attached to the upper tails (the error levels) are 0.1%, 0.3% and 8.2% for the 1, 2 and 3 descriptor models, respectively. The first case gives the most reliable result. Thus, statistically speaking, the predictions on the 10 test molecules using the single descriptor model is the most robust result according to the ITS protocol. This agrees with the fact that the single parameter model gives the highest r_{ITS} value.

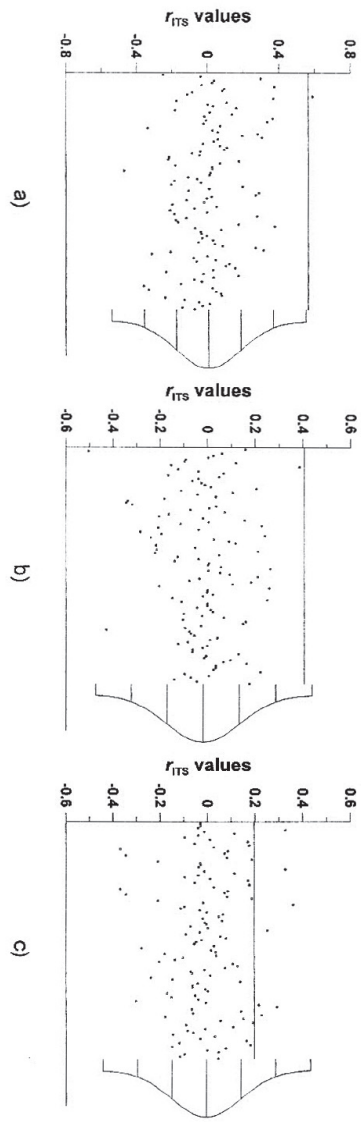

Figure 2 Randomisation tests considering models of 1 (a), 2 (b) and 3(c) descriptors by model. Each point is attached to a full ITS procedure calculation. The random points follow a gaussian distribution (depicted). The horizontal line depicts the actual result which can be contrasted against the normal distribution (see text).

CONCLUSIONS

The ITS methodology is presented and its advantages revealed when comparing predictions and CV methodologies in the literature. This study was undertaken considering a well-known benchmark molecular set, the so called Cramer steroids, a set of 31 steroids interacting with the corticosteroid binding globuline. The relevant conclusion is that the ITS method serves to direct the model construction to a minimum overparametrisation degree, providing the r_{ITS} parameter, which measures the predictive ability of the methodology with respect to a potential external or validation molecular set. This predictive capability quantification appears to be more realistic than the ones obtained from other cross-validation parameters, as r_{CV} or $q^{(2)}$ terms.

ACKNOWLEDGEMENTS

E. Besalú thanks the Spanish Research Projects SAF2000-0223-01, CTQ2006-04410/BQU and the Fundación Maria Francisca de Roviralta for their financial support. L. Vera thanks the Secretaría de Estado de Educación y Universidades (Spain) for the financial help which allowed visiting the Institute of Computational Chemistry at the University of Girona where this work started. He also thanks the Dirección General de Investigaciones y Cooperación Técnica of Universidad Católica del Norte (Chile) for supporting this study.

REFERENCES

1. M. Stone, *J. of the Roy. Stat. Soc*. **B 36**, 111, (1974).
2. S. Wold, *Technometrics* **20**, 397, (1978).
3. S. Wold, *Struct.-Act. Relat*. **10**, 191, (1991).
4. S. Wold, and L. Eriksson, "Statistical validation of QSAR results. Validation tools". In *"Chemometric Methods in Molecular Design"* (H. van de Waterbeemd, Eds.). VCH, Weinheim, 1995.
5. Ajay, J. Med. Chem. 36, 3565, (1993).
6. Ajay, *Chemom. Intell. Lab. Syst*. **24**, 19, (1994).
7. R. D. Cramer, D. E. Patterson, and J. D. Bunce, *J. Am. Chem. Soc.* **110**, 5959, (1988).
8. D. D. Robinson, P. J. Winn, P. D. Lyne, and W. G. Richards, *J. Med. Chem*. **42**, 573, (1999).
9. N. R. Draper, and H. Smith, "*Applied Regression Analysis*". Wiley, New York, 1966.
10. D. C. Montgomery, and E. A. Peck, "*Introduction to linear regression analysis*". Wiley, New York, 1992.
11. E. Besalú, *J. Math. Chem*. **29**, 191, (2001).
12. H. Kubinyi, and U. Abraham, "Practical problems in PLS Analyses". In "*3D QSAR in Drug Design*" (H. Kubinyi, Eds.). ESCOM, Leiden, 1993.
13. M. Baroni, G. Constantino, G. Cruciani, D. Riganelli, R. Valigi, and S. Clementi, *Quant. Struct.-Act. Relat*. **12**, 9, (1993).
14. S. Clementi, G. Cruciani, D. Riganelli, and R. Valigi, "GOLPE: Merits and Drawbacks in 3D-QSAR". In "*QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*" (F. Sanz, J. Giraldo and F. Manaut, Eds.). Prous Pub., Barcelona, 1995.
15. H. Van De Waterbeemd, "Chemometric Methods Used in Drug Discovery". In "*Structure-Property Correlations in Drug Research*" (H. van de Waterbeemd, Eds.). Acad. Press, San Diego, CA, 1996.
16. R. D. Cramer, S. A. Depriest, D. E. Patterson, and P. Hecht, "The Developing practice of Comparative Molecular Field Analysis". In "*3D QSAR in Drug Design*" (H. Kubinyi, Eds.). ESCOM, Leiden, 1993.
17. D. M. Allen, *Technometrics* **16**, 125, (1974).
18. L. Amat, E. Besalú, R. Carbó-Dorca, and R. Ponec, *J. of Chem. Inf. and Comput. Sci*. **41**, 978, (2001).
19. S. S. Liu, C. S. Yin, Z. L. Li, and S. X. Cai, *J. Chem. Inf. Comput. Sci*. **41**, 321, (2001).
20. G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, and A. Zaliani, *J. Comput.-Aided Mol. Des*. **11**, 79, (1997).
21. J. G. Topliss, and R. P. Edwards, *J. Med. Chem*. **22**, 1238, (1979).
22. C. L. Waller, and M. P. Bradley, *J. Chem. Inf. Comput. Sci*. **39**, 345, (1999).
23. E. Besalú, R. Ponec and J. V. de Julián-Ortiz, *Mol. Divers.* **6**(2), 107, (2003).
24. J. V. de Julián-Ortiz, E. Besalú and R. García-Domenech, *Indian J. of Chem.* **42**(6)A, 1392, (2003).
25. R. García-Domenech, J. V. de Julián-Ortiz, E. Besalú, *Mol. Divers*. **10,** 159, (2006).
26. J. V. de Julián-Ortiz and E. Besalú, *Int. J. Mol. Sci.* **7**(10) 456, (2006).
27. S. W. Dietrich, N. D. Dreyer, and C. Hansch, *J. Med. Chem.* **23**, 1201, (1980).
28. J. Huuskonen, *J. Chem. Inf. Comput. Sci.* **41**, 425, (2001).
29. A. C. Good, S. S. So, and W. G. Richards, *J. Med. Chem*. **36**, 433, (1993).
30. G. Klebe, U. Abraham, and T. Mietzner, *J. Med. Chem*. **37**, 4130, (1994).
31. B. D. Silverman, and D. E. Platt, *J. Med. Chem*. **39**, 2129, (1996).
32. T. I. Oprea, D. Ciubotariu, T. L. Sulea, and Z. Simon, *Quant. Struct.-Act. Relat*. **12**, 21, (1993).
33. A. N. Jain, K. Koile, and D. Chapman, *J. Med. Chem*. **37**, 2315, (1994).
34. M. Hahn, and D. Rogers, *J. Med. Chem*. **38**, 2091, (1995).
35. M. Wagener, J. Sadowski, and J. Gasteiger, *J. Am. Chem. Soc*. **117**, 7769, (1995).
36. G. E. Kellogg, L. B. Kier, P. Gaillard, and L. H.Hall, *J. Comput.-Aided Mol. Des*. **10**, 513, (1996).
37. S. Anzali, G. Barnickel, M. Krug, J. Sadowski, M. Wagener, J. Gasteiger, and J. Polanski, *J. Comput.-Aided Mol. Des.* **10,** 521, (1996).
38. U. Norinder, *J. Chemom.* **10**, 533, (1996).
39. J. Schnitker, R. Gopalaswamy, and G. M. Crippen, *J. Comput.-Aided Mol. Des*. **11**, 93, (1997).
40. D. B. Turner, P. Willett, A. M. Ferguson, and T. Heritage, *J. Comput.-Aided Mol. Des.* **11**, 409, (1997).
41. M. F. Parretti, R. T. Kroemer, J. H. Rothman, and W. G. Richards, *J. Comput. Chem*. **18**, 1344 (1997).
42. S. S. So, and M. Karplus, *J. Med. Chem*. **40**, 4347, (1997).
43. Y. Tominaga, and I. Fujiwara, *J. Chem. Inf. Comput. Sci.* **37**, 1152, (1997).
44. H.Chen, J. Zhou, and G. Xie, *J. Chem. Inf. Comput. Sci.* **38**, 243, (1998).
45. M. Lobato, L. Amat, E. Besalú, and R. Carbó-Dorca, *Quant. Struct.-Act. Relat.* **16**, 465, (1997).
46. D. Robert, L. Amat, and R. Carbó-Dorca, *J. Chem. Inf. Comput. Sci.* **39**, 333, (1999).
47. R. Carbó-Dorca, L. Amat, E. Besalú, X. Gironés, and D. Robert, "Quantum Molecular Similarity: Theory and Applications to the Evaluation of Molecular Properties, Biological Activities and Toxicity". In "*Fundamentals of Molecular Similarity*" (R. Carbó-Dorca, X. Gironés and P. G. Mezey, Eds.). Kluwer Acad./ Plenum Pub., New York, 2001.
48. E. Besalú, and R. Carbó, Scientia Gerundensis **21**, 145, (1995).
49. R. Carbó-Dorca, L. Amat, E. Besalú, and M. Lobato, "Quantum Similarity". In "*Advances in Molecular Similarity"* (R. Carbó-Dorca and P. G. Mezey, Eds.). Jai Press Inc., Greenwich (Conn.), 1998.
50. E. Besalú, A. Gallegos and R. Carbó-Dorca, *MATCH Commun. Math. Comput. Chem.* **44**, 41, (2001).
51. J. Shao, *J. Am. Stat. Assoc*. **88**, 486 (1993).
52. E. Besalú, J. V. de Julián-Ortiz, L. Pogliani , *J. Chem. Inf. Model.* **47**(3) 751, (2007).
53. S. S. So, S. P. Van Helden, V. J. Van Geerestein, and M. Karplus, *J. Chem. Inf. Comput. Sci*. **40**, 762, (2000).
54. L. B. Kier, and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research". Academic Press, New York, 1976.
55. Structures can be also found at http://iqc.udg.es/cat/similarity/QSAR/steroids/ (accessed on May 29th 2008)