Predictivity of QSAR

Romualdo Benigni* and Cecilia Bossa

Environment and Health Department, Istituto Superiore di Sanita′, Viale Regina Elena 299, 00161 Rome, Italy

Received January 10, 2008

A range of good quality, local QSARs for mutagenicity and carcinogenicity have been assessed and challenged for their predictivity in respect to real external test sets (i.e., chemicals never considered by the authors while developing their models). The QSARs for potency (applicable only to toxic chemicals) generated predictions 30−70% correct, whereas the QSARs for discriminating between active and inactive chemicals were 70−100% correct in their external predictions: thus the latter can be used with good reliability for applicative purposes. On the other hand internal, statistical validation methods, which are often assumed to be good diagnostics for predictivity, did not correlate well with the predictivity of the QSARs when challenged in external prediction tests. Nonlocal models for noncongeneric chemicals were considered as well, pointing to the critical role of an adequate definition of the applicability domain.

INTRODUCTION

The quantitative structure−activity relationships (QSAR) science provides scientific tools for exploring and understanding data and for predicting the activity of new, untested chemicals. A crucial issue in QSAR is the predictive ability of the models. Good predictivity is essential for the practical use of the model (e.g., directing the synthesis of more active or less toxic chemicals, contributing to the toxicological profiling of chemicals for regulatory purposes). In addition the assessment of predictivity provides an essential piece of information on the validity of a QSAR: a QSAR that does not predict beyond the chemicals used for its generation is only a rhetoric description of the data. Thus, assessing the predictivity of a QSAR is an integral part of the process of generating and validating the model.

The QSAR community has been aware of this issue since the beginning of QSAR. A clear definition of the problem is presented by Corwin Hansch in a 1977 paper: ". . .Turning now to the problem of direct predicting of the biological activity of untested molecules from known QSAR, it is instructive to divide the problem into two parts: 1) predictions within spanned substituent space (SSS); 2) predictions outside of SSS. . . . Good predictions within SSS are not a trivial matter since in multiparameter equations, being able to maximize the contributions of all variables within SSS may mean a significant increase in log 1/C. We can say little or nothing with certainty about un-SSS; we know of course that no linear relationship in SAR work will hold forever. . .".[1]

The research on QSAR predictivity, and on the best criteria for its assessment, is still very active, and no definitive solution or recipe exists. In principle, prediction of external validation sets (i.e., chemicals not considered in any way by the modelers in the phase of QSAR generation) is considered to be the gold standard. Among such examples are a number of prospective comparative prediction exercises held under the aegis of the U.S. National Toxicology Program (NTP), that invited modelers to present predictions on the mutagenic or carcinogenic activity of chemicals before the actual experimental results were available.[2−4] Since often such external test sets do not exist or are difficult to retrieve, a number of statistical surrogates that attempt to estimate external predictivity from the characteristics of the training set (e.g., y-scrambling, boot-strapping, cross-validation, artificial test set) have been devised and largely used instead of the check on real external chemicals.[5] These internal validation approaches have strong advocates; however, a number of papers pointed out to limitations and poor reliability of such methods.[6,7]

Recently, the expanding role of QSAR for regulatory purposes has brought out more strongly the need for elucidation of the issue of predictivity. In particular in Europe a new regulation of chemicals, called REACH, has been just enforced. Since REACH envisages a wider role for the theoretical assessment of properties/activities, a range of preparatory initiatives has been undertaken.[8] Among others, the European Chemicals Bureau (ECB) has promoted a survey on the (Q)SARs for mutagens and carcinogens in the public domain. The survey has been performed by our group at the Istituto Superiore di Sanita′ and has produced−among others−evidence on the predictivity of selected QSARs for congeneric sets of chemicals. A congeneric class includes chemicals with a similar basic scaffold (e.g., aromatic amine), acting through the same mechanism, possibly with the same rate-limiting step. The QSARs have been challenged on real external test sets, after checking the congenericity of the test chemicals with the training set chemicals. In this paper, we present the results of the predictions, and we compare them with the indications from internal (statistical) validation approaches. These results are then put in a wider perspective, and previous and new analyses on the predictivity of QSARs for noncongeneric chemicals are considered.

DATA AND METHODS

QSARs for Congeneric Classes of Chemicals. The results of our survey on the (Q)SARs for mutagens and

* Corresponding author e-mail: rbenigni@iss.it.

10.1021/ci8000088 CCC: $40.75
Published on Web 04/22/2008

carcinogens in the public domain are presented in detail in a report[9] freely available on the Internet (http://ecb.jrc.it/DOCUMENTS/QSAR/EUR_22772_EN.pdf).

Briefly, the evaluation of the noncommercial (Q)SARs for mutagenicity and carcinogenicity consisted of a preliminary survey of all the models available in the literature. Based on the information provided by the authors, a number of promising models were short listed and analyzed more in depth in a second phase, where the information provided by the authors was completed and complemented with a series of analyses aimed at generating an overall profile of each of the short listed models. The models can be divided into two families based on their target: (a) congeneric and (b) noncongeneric sets of chemicals. The noncongeneric models included several sets of Structural Alerts (for a review, see ref 10), whose relevance for the assessment of the carcinogenic and mutagenic risk was evaluated.

Another section of the work focused on the QSARs for congeneric classes: this work is summarized in this paper. After examination of the available literature, a number of QSAR models were short listed, based on their quality (e.g., size of training sets, statistical fit, mechanistic understanding). These models were reanalyzed by us, and—when necessary—further results were provided.

For multiple linear regression (MLR) models, the characterization includes the following: r2; adjusted r2; and q2. The cross-validated r2 (q2) is = 1 − (sum of squares of the predictive residuals/sum of squares of the mean-centered response data). The mean leverage (with SD) is used to assess if a model depends in a balanced way on the data points (good models should exhibit low mean leverage).

For discriminant models (linear discriminant analysis or canonical discriminant analysis), the characterization includes the following: accuracy, sensitivity, and specificity, together with the squared canonical correlation. Accuracy is the percentage of all chemicals correctly identified by the model. Sensitivity is the percentage of biologically active (positive) chemicals correctly identified (calculated out of the total number of positives). Specificity is the percentage of biologically inactive (negative) chemicals correctly identified (calculated out of the total number of negatives). The squared canonical correlation is a measure of the correlation between the biological activity variable and the linear combination of descriptor variables that best separates the negatives from the positives.

For all models, cross-validation was performed also by leave-many-out procedures (LMO). Here we report the LMO results obtained leaving out 10% of the data set. Each procedure was applied ten times (by random selection of excluded chemicals). More results are in the report of the study.

For each QSAR model, the external predictivity was assessed by retrieving from the literature experimental data relative to the same chemical classes of the models but not considered by the authors of the models.

For regression based models, the performance in external validation is expressed as a correlation coefficient between experimental and predicted potency. An additional way of measuring the prediction performance is the percentage of test chemicals correctly predicted within one log unit of potency. For discriminant models, the external predictivity is measured as a percentage of test chemicals correctly predicted (accuracy).

The applicability of the models to the retrieved test sets was assessed by checking the concordance between the chemical domains of the training and test sets in terms of (a) types of chemical structures and presence/absence of functional groups; (b) range of the values of the descriptors in the models; and (c) chemical similarity (Tanimoto index).

QSARs for Noncongeneric Sets of Chemicals. An original QSAR model for rodent carcinogenicity was established, based on the use of the Tanimoto similarity metrics as descriptors. The Tanimoto coefficient between two chemicals is the ratio of shared substructures to the number of all substructures that appear in the two chemicals. The Tanimoto coefficient varies between 0 (total lack of similarity) to 1 (the query chemical has an identical constitution to the reference chemical). In this work, the $n \times n$ Tanimoto similarity matrix (between all chemicals) is calculated with the computer software Leadscope Enterprise v. 2.4.11−4 (Leadscope Inc., Columbus OH), based on the range of around 27,000 substructures with which Leadscope characterizes the chemicals. The dimensionality of the $n \times n$ similarity matrix is then reduced with Principal Component Analysis.[11]

The chemical carcinogenicity data were retrieved from the ISSCAN database established at the Istituto Superiore di Sanita′, Rome. ISSCAN can be freely downloaded from the ISS Web site (http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7) or from the DSSTox site (http://www.epa.gov/ncct/dsstox/sdf_isscan_external.html). It contains information on chemical compounds tested with the long-term carcinogenicity bioassay on rodents (rat, mouse). It is characterized by (a) the quality of data [The data were cross-checked on different sources of information available; contradictions were solved going back to the original papers, and results based on insufficient protocols were not included.] and (b) by the fact it is QSAR-ready. The biological data (carcinogenicity and Salmonella mutagenicity) are coded in numerical terms that can be used directly for QSAR analyses.[12]

RESULTS AND DISCUSSION

External Predictivity of QSARs for Congeneric Chemical Classes. The external predictivity of selected QSARs was studied, and the results were compared with the estimations provided by a series of surrogate statistical indices (internal or statistical validation). The QSARs were short listed for their high quality in the course of a survey on the QSARs for mutagens and carcinogens in the public domain.[9] Table 5 displays the QSAR models and their characteristics relevant to this study. The QSARs are summarized in the Appendix, and more details are reported in ref 9.

The QSARs are divided into models for (a) the gradation of potency of active (mutagenic or carcinogenic) chemicals and (b) the discrimination between positive and negative chemicals. In fact, previous work has demonstrated that the two types of models—in the field of mutagenicity and carcinogenicity—are usually different: the factors that modulate the gradation of potency of the positive compounds are usually different from those that make the difference between

Table 1. Regression-Based Models for Potency: Fit and Predictivity Measures[a]

		training set				test set	
QSAR	system	rtra	Q2	q2_10	lever	rte	accte
QSAR1	TA98	0.90	0.78	0.71	0.06	0.41	0.36
QSAR2	TA100	0.88	0.74	0.66	0.06	0.68	0.57
QSAR3	mouse	0.91	0.58	0.00	0.25	0.56	0.58
QSAR4	rat	0.93	0.81	0.79	0.15	0.48	0.71
QSAR9	TA98	0.90	0.89	0.80	0.04	−0.23	0.43
QSAR10	TA100	0.88	0.77	0.73	0.05	0.36	0.32

[a] The fitting, internal validation indices, and external validation outcomes are reported for the regression-based local QSARs presented in Table 5. The training set measures are as follows: rtra, correlation coefficient of the training set; q2, r2, cross-validated (leave-one-out); q2_10, r2 cross-validated (leave-10%-out); lever, mean leverage of the data points. The test set measures are as follows: rte, correlation coefficient between experimental and predicted values; accte, accuracy (percentage of chemicals correctly predicted within 1 log activity unit).

Table 2. Regression-Based Models for Potency: Correlation Coefficients for Indices in Table 1

	training set				test set	
	rtra	q2	q2_10	lever	rte	accte
rtra	1.00	0.09	−0.02	**0.52**	−0.07	**0.60**
q2		1.00	**0.93**	**−0.77**	**−0.69**	−0.25
q2_10			1.00	**−0.84**	−0.39	−0.25
lever				1.00	0.43	**0.62**
rte					1.00	0.41
accte						1.00

positives and negatives (the latter models describing the minimum reactivity requirements).[13,14]

In the selection of the external sets, the constraint for a test set to belong to the applicability domain of the training set was taken into account by considering (a) the types of structures to which the model applies [This was assessed subjectively by us according to our expert knowledge, by checking, among others, the absence of reactive groups different from those that characterize the chemical class under study.], (b) the ranges of descriptors values of the two sets, and (c) a mathematical transform of structural similarity indices. The external test sets used by us were within the applicability domains of the models (i.e., range of structures and features characteristic of the training sets) (for details, see [9]).

Table 1 summarizes the external prediction outcomes for regression based models (i.e., QSAR models for potency), and Table 3 summarizes the outcomes for discriminating models (i.e., QSAR models for activity). The two tables report also parameters for goodness of fit and different internal validations of the training set. Further information is displayed in Tables 2 and 4: Table 2 shows the correlation coefficients among the parameters in Table 1, and Table 4 shows the correlation coefficients among those in Table 3.

The QSARs in Table 1 refer to the gradation of potency of two chemical classes (aromatic amines and nitroarenes) in in vitro (*Salmonella typhimurium* TA98 and TA100 strains) and in vivo (rodents) systems. The biologic al activity is mutagenicity for *S. typhimurium* and carcinogenicity for rodents.

Table 3. Discriminant Models for Activity: Fit and Predictivity Measures[a]

		training set			test set
QSAR	system	sqcc	acctra	acc10	accte
QSAR7	rodent	0.38	0.88	0.75	0.67
QSAR8	rodent	0.50	0.94	0.78	0.70
QSAR5	TA98	0.46	0.89	0.88	0.69
QSAR6	TA100	0.52	0.87	0.87	0.81
QSAR13	TA100	0.61	1.00	0.85	1.00

[a] The fitting, internal validation indices, and external validation outcomes are reported for the discriminant local QSARs presented in Table 5. The training set measures are as follows: sqcc: squared canonical correlation; acctra: accuracy (percentage of chemicals correctly assigned/total number of chemicals); and acc10: cross-validated accuracy (leave-10%-out). The test set measures are as follows: accte: accuracy.

Table 4. Discriminant Models for Activity: Correlation Coefficients for Indices in Table 3

	sqcc	acctra	acc10	accte
sqcc	1.00	**0.75**	0.50	**0.89**
acctra		1.00	−0.05	**0.72**
acc10			1.00	0.39
accte				1.00

Inspection of Table 1 indicates that the goodness of fit in the training set (correlation coefficient, rtra) is always considerably better than the goodness of prediction (rte) for the test set. rte is the correlation coefficient between predicted and experimental potency of the test set. It also appears that the internal validation measures (q2 and q2_10) are lower than the back-fitting of the model (rtra) but still considerably higher than the external prediction rte. q2 and q2_10 are r^2 cross-validated with the leave-one-out and leave-10%-out procedures, respectively. Thus the internal validation measures are quite far from the performance values with external test sets.

An alternative way of measuring the prediction performance for regression based models is to calculate its accuracy as a percentage of test chemicals correctly predicted within one log unit of activity (accte). When expressed as accte, the prediction performance is a more robust estimate than when expressed as rte (see for example QSAR9, which shows a negative correlation between predicted and experimental potency values (rte), whereas the percentage of chemicals correctly predicted within one log unit (accte) is 0.43). This is understandable, since high rte values require exact point estimates, whereas high accte requires correct estimates of intervals; the latter is a less stringent criteria and, in addition, is closer to the regulatory needs. Overall, the QSAR external predictions for the potency of congeneric chemicals are 30−70% correct (accte).

Table 2 (correlation coefficients) indicates that rtra is correlated with accte but not with rte, thus indirectly confirming that accte is a performance index more robust than rte.

Regarding the internal validation indices q2, q2_10, and mean leverage (lever), Table 2 shows that q2 and q2_10 are negatively correlated with both rte and accte. In addition the mean leverage (lever) values (whose high values are supposed to indicate "bad" models, i.e., with uneven influence of individual data points on the models) are positively

Table 5. Local, Noncommercial (Q)SAR Models for Mutagenicity and Carcinogenity That Were Surveyed in the Framework of a Collaboration between the ISS and the ECB[a]

code	chemical class	biological end points	type	*n*	ref
QSAR1	aromatic amines	Salmonella mutagenicity TA98	potency	88	28
QSAR2	aromatic amines	Salmonella mutagenicity TA100	potency	67	28
QSAR3	aromatic amines	mouse carcinogenicity	potency	37	29
QSAR4	aromatic amines	rat carcinogenicity	potency	41	29
QSAR5	aromatic amines	Salmonella mutagenicity TA98	activity	111	30
QSAR6	aromatic amines	Salmonella mutagenicity TA100	activity	111	30
QSAR7	aromatic amines	rodent overall carcinogenicity	activity	66	31
QSAR8	aromatic amines	rodent overall carcinogenicity	activity	64	31
QSAR9	nitroarenes	Salmonella mutagenicity TA98	potency	188	32
QSAR10	nitroarenes	Salmonella mutagenicity TA100	potency	117	32
QSAR13	aldehydes	Salmonella mutagenicity TA100	activity	20	33

[a] A summary of the models, with descriptors definition and statistics, is given in the Appendix. For more details, see the report.9 The codes of the QSARs are the same reported as in the above report. Type: The models are divided by type into the following: potency = QSARs obtained by regression analysis of mutagenic or carcinogenic potency (it applies only to positive chemicals); activity = QSARs obtained by discriminant analysis of dichotomous data (positive/negative); *n*: number of chemicals in the training sets of the models.

correlated with both the external validation indices rte and accte. All these results are contrary to what one would expect (i.e., a positive correlation between internal validation indices and external predictivity, on one hand, and a negative correlation between leverage values and external predictivity) if the internal predictivity indices were good "predictors" of external predictivity.

The results of external validation for the discriminant models (activity) are in Table 3. The QSARs refer to the chemical classes of aromatic amines and $\alpha\beta$-unsaturated aldehydes and to mutagenic activity in *S. typhimurium* or carcinogenic activity in rodents.

Table 3 shows that the overall accuracy in the training set (acctra) is in the range 0.88−1.00 and is systematically higher than that attained in the external test set (accte). The external prediction performance is 0.67−1.00 accurate, considerably higher than that of the regression models for potency (0.32−0.71). This confirms that predicting intervals is more reliable than predicting individual data points.

Table 4 shows that the accuracy in the training set (acctra) has a good correlation with external predictivity (accte). An even better correlation with external predictivity is shown by the squared canonical correlation (sqcc) of the models.

Table 4 also shows that the internal validation index (acc10: cross-validation leave-10%-out) has a mediocre correlation with the external predictivity.

Overall, a coherent picture emerges from the results in Tables 1−4. All the QSARs selected are of good quality, since this was the aim of our survey of (Q)SARs: the QSARs had a good internal statistics (including cross-validation measures) and were directly related to the underlying mechanism of action of the classes/activities under study. However, when challenged for their predictivity of external sets of chemicals retrieved from the literature and belonging to the applicability domain of the models, the performance of the QSARs spread over a wide range of values. Overall, the QSARs for potency (applicable only to positive chemicals) generated predictions 30−70% correct, whereas the QSARs for discriminating between active and inactive chemicals were 70−100% correct in their external predictions. Thus, the QSARs for yes/no activity exhibit a more than satisfactory reliability, remarkably higher than that of the QSARs for potency. This is understandable, since predicting values intervals is a definitely easier task than predicting exact data points. Based on our results, the range 70−100% correct predictions can be considered as a measure of the uncertainty typical of "good" yes/no QSAR models. The present results also indicate that the various internal validation measures are not useful to assess the predictivity of the QSARs, whereas fitting measures, like correlation coefficient and squared canonical correlation, have a better correlation with the output of external predictions.

Nonlocal Models for Sets of Noncongeneric Chemicals. The QSARs evaluated up to here in this paper are all local models for congeneric sets of chemicals. Since each of the models describes a class with one (or a few variations of one) mechanism of action, and its applicability domain can be assessed with relative certainty based on chemical knowledge and consideration of parameters in the models, it can be assumed that these models provide a privileged point of view and are the cases in which best performance should be expected. As shown above, the external predictivity exhibited by the QSARs for yes/no activity is in the range 70−100% accuracy. It is interesting to try to compare these figures with those relative to nonlocal models, which include some of the most popular commercial software implementations for predicting the mutagenicity/carcinogenicity of chemicals. The commercial models were outside the scope of our evaluation work for ECB, so the results shown below are collected from various literature sources. They refer to the software programs Topkat,[15,16] DEREK,[17] and Multicase.[18,19]

It should be noted that these methods are often called global, since they are aimed at (hopefully) predicting the toxicity of any type of chemicals and are based on large training sets of noncongeneric compounds that can be understood as unknown mixtures of various local series for which different QSARs can be potentially developed. However such a series is by no means global, considering the size of the chemical universe as a whole, but is only "nonlocal".

Figures 11−3 display the outcomes of a series of external prediction exercises performed by various investigators with three nonlocal models in the commercial domain: Multicase, Topkat, and Derek. These include the prospective prediction exercises promoted by the NTP as well as several studies performed by industries with their in-house sets of com-

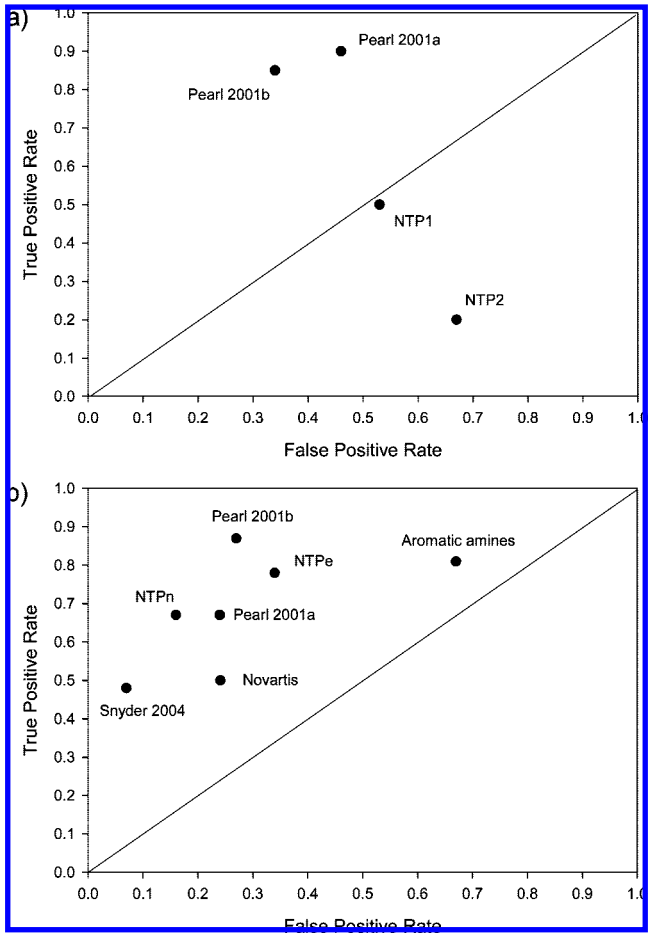

Figure 1. The ROC graph reports the performance of the Multicase software system in a series of external prediction exercises for carcinogenicity (a) and mutagenicity (b) (see more details in the text). The codes make reference to the following original studies: (a) Pearl 2001a and Pearl 2001b,[34] NTP1,[35] and NTP2[4] and (b) Pearl 2001a and Pearl 2001b,[34]Aromatic amines, and Novartis,[36] NTPn, and NTPe[3] Snyder 2004.[37]

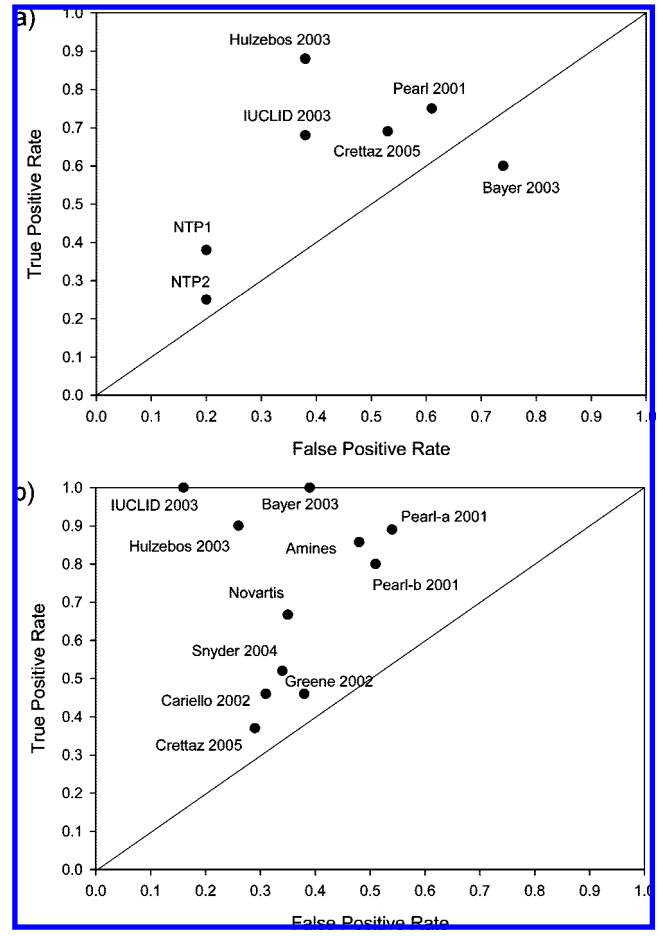

Figure 2. The ROC graph reports the performance of the DEREK software system in a series of external prediction exercises for carcinogenicity (a) and mutagenicity (b) (see more details in the text).The codes make reference to the following original studies: (a): Hulzebos 2003,[38] Bayer 2003,[36] IUCLID 2003,[36] Pearl 2001,[34] NTP1,[35] NTP2,[4] and Crettaz 2005.[39] (b) Hulzebos 2003,[38] Bayer 2003,[36] Amines,[36] IUCLID 2003,[36] Cariello 2002,[40] Greene 2002,[41] Snyder 2004,[37] Pearl-a 2001 and Pearl-b 2001,[34] and Crettaz 2005.[39]

pounds.[14] The common characteristics of these studies is that the external chemicals to be predicted were different from those used as training sets by the authors of the programs, and were performed independently.

Figures 1−3 are receiver operating characteristics (ROC) plots that report the true positive rate (sensitivity) on the Y-axis and false positive rate (1 - specificity) on the X-axis. In a ROC graph, perfect performance is located at the left upper corner, whereas the diagonal line represents random results.[20]

What is striking in the above figures is the extremely large variability of the responses: the predictions of the external chemicals vary very much both in terms of overall accuracy (closeness to the "perfection" point represented by the left top corner in the ROC graph) and in terms of relative proportions of true and false positives. This observation applies in the same way to the three commercial systems under study. This contrasts with the usually good "average" performance reported by the authors, as assessed on large noncongeneric databases: here it appears that the performance with the different segments of the universe of chemicals is largely unpredictable.

Since the training sets for nonlocal models can be considered as unknown mixtures of various local series, an hypothesis about the cause of the above reality is that the systems failed to correctly "warn" the users on which predictions are reliable and which are not (i.e., which predictions are in the applicability domain of the systems, and which do not). The need for further expanding the ability to correctly identify the "allowed" and "nonallowed" portions of the chemical space for each predictive system is widely recognized today, and much work is going on.[21] On the other hand, whereas the generation and validation of local, mechanistically based QSARs can exploit the general knowledge in the fields of QSAR, chemistry and biology and combine this with statistical checks, the nonlocal QSARs tend to be based more and more on the use of computer programs that can generate hundreds of parameters per compound automatically. Clearly, such a situation calls for automatic procedures to arrive to a QSAR model, and this in turn leads to an overestimation of the value of the statistical assessment which turns to be the sole validation tool. As demonstrated by Figures 1−3, this procedure can be misleading. Additional clues are provided by the following exercise performed in our laboratory.

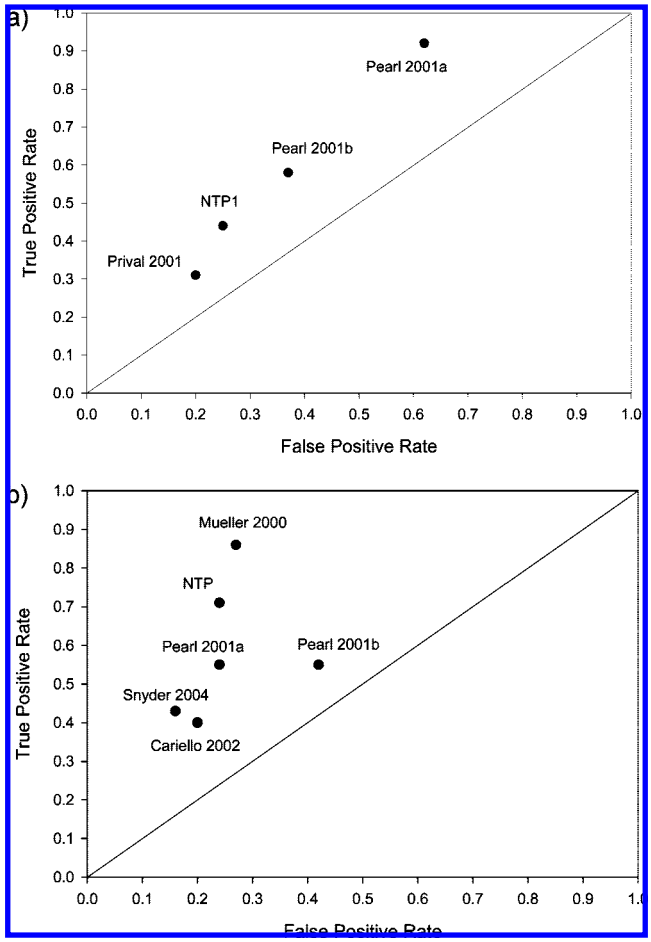

Figure 3. The ROC graph reports the performance of the Topkat software system in a series of external prediction exercises for carcinogenicity (a) and mutagenicity (b) (see more details in the text).The codes make reference to the following original studies: (a) Pearl 2001a and Pearl 2001b,[34] NTP1,[35] Prival 2001[42] and (b) Pearl 2001a and Pearl 2001b,[34] Snyder 2004,[37] Mueller 2000,[43] Cariello 2002,[40] and NTP.[3]

Building and Evaluating a New Nonlocal Model for Carcinogenicity. The QSAR model was based on chemical similarity indices (Tanimoto indices) as descriptors, calculated with the program Leadscope Enterprise version 2.4.11−4 (Leadscope Inc., Columbus, OH). The chemical carcinogenicity data were retrieved from the ISSCAN database (freely downloaded from the ISS Web site http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7). As an external test set, the same chemicals used by the NTP in two prospective prediction exercises on chemical carcinogenicity[2] were considered.

From the ISSCAN database, 873 compounds (bioassayed in rodents) with definite chemical structures were extracted. Out of them, 67 compounds were the target of the NTP prediction exercises; these chemicals were considered as a test set and were not used in the development of the model: the remaining 806 compounds were the training set.

The first step was the calculation of Tanimoto similarity indices of each of the 873 chemicals (whole set) from the 806 training set compounds only: this generated a matrix with 873 rows (chemicals) and 806 variables (similarity). Thus the training set defined the chemical space under study, and the characteristics of the test set did not influence the generation of the model. The 806 variables were treated with

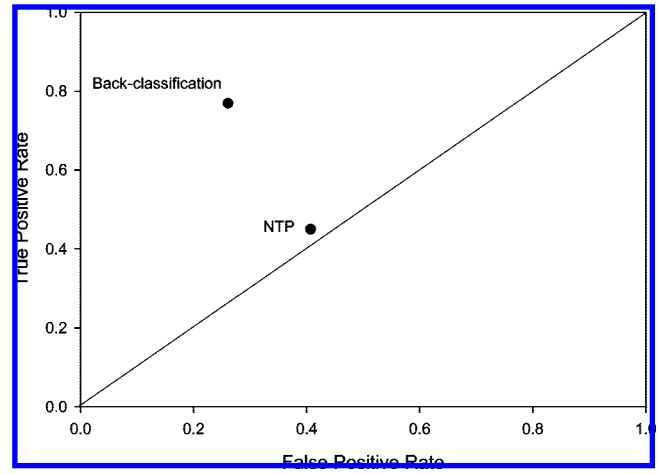

Figure 4. ROC graph of the performance on a Tanimoto-similarity based model on rodent carcinogenicity. The figure displays the performance of the model in terms of back-classification of the training set and of prediction of the external test set (NTP chemicals) (see details in the text).

principal component analysis (PCA), and 160 principal components (PC) (explaining 94.4% of variance) were accepted and submitted to analysis. The number of 160 PCs to be screened for building the model was decided because this is the maximum acceptable for a 5:1 ratio of compounds to variables, in order to minimize the risk of chance correlations.[22]

The second step was the application of stepwise linear discriminant analysis to the 806 compounds of the training set, based on the 160 PCs. As a result a statistically significant model separated carcinogens from non-carcinogens with a squared canonical correlation = 0.29 (47 PCs entered into the model; accuracy = 0.76; sensitivity = 0.77; specificity = 0.74). Thus the mutual chemical similarities of the training set compounds generated a statistically significant model. It should be emphasized that the capability of building a QSAR on a fragment-based similarity matrix is not a novelty, since it has been demonstrated that such a matrix contains information equivalent to that of e.g., a wide range of continuous classical descriptors (see, for example, ref 23).

The 47_PCs model was applied to predict the carcinogenicity of the 67 NTP chemicals left out from the generation phase. Figure 4 displays the performance of the model as back-classification of the training set and as external prediction. The ROC graph shows that the model failed to correctly predict the carcinogenicity status of the NTP chemicals: in fact the performance with the test set is very close to the diagonal line where random results are located.

It should be noted that the NTP chemicals are perfectly inside the applicability domain of the model. In fact, a canonical discriminant analysis based on the 47 model PCs failed to separate the test set from the training set chemicals: Figure 5 shows the overlap between the two groups on the best discriminating linear combination of variables (PCs) (squared canonical correlation = 0.05).

To investigate deeper, various cross-validation approaches were applied to the training set data. These approaches are detailed in the legend to Figure 6 and range from a simple leave-one-out (LOO) procedure where the same PCs of the overall model are used to a most stringent leave-10%-out

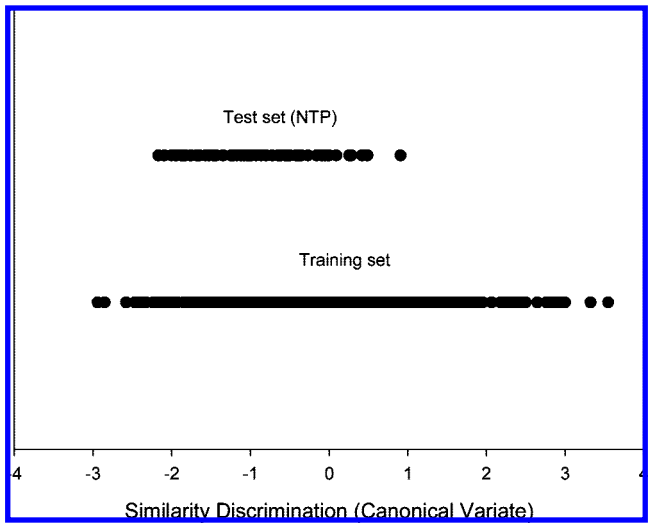

Figure 5. Applicability domain of a Tanimoto similarity-based model on rodent carcinogenicity.The figure shows the overlap between the similarity spaces of the training and the test set (see details in the text).

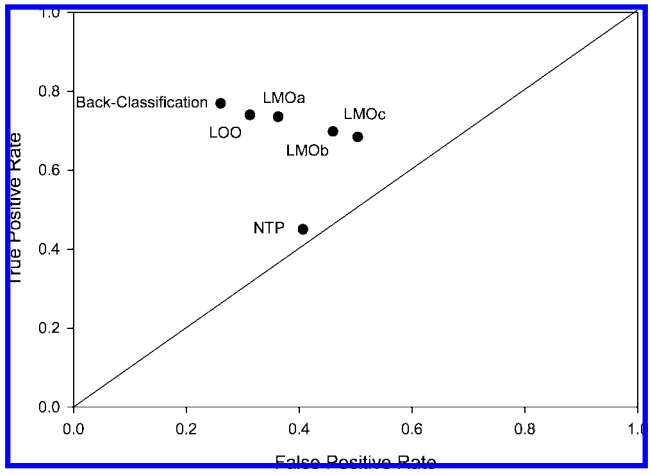

Figure 6. ROC graph of the performance on a Tanimoto-similarity based model on rodent carcinogenicity. Together with the performance of the model in terms of back-classification of the training set and prediction of the external test set (as in Figure 4), the figure reports the results of a series of cross-validation approaches applied to the training set. The cross-validation approaches are the following: LOO: leave-one-out, the same 47 PCs of the model were used. Only the coefficients of the model were recalculated at each iteration; LMOa: leave-10%-out, version_a: the same 47 PCs of the model were used. Only the coefficients of the model were recalculated at each iteration. The results are the average of 10 repetitions (as for the following versions); LMOb: leave-10%-out, version_b: at each iteration, discriminant analysis was applied to select the relevant PCS out of the first 160 PCs originally calculated for establishing the model. Thus, both the subset of PCs and the model coefficients were new at each iteration; LMOc: leave-10%-out, version_c: at each iteration, a new PCA was applied only to the variables relative to the selected training set (90% at each iteration). This step was followed by the application of discriminant analysis. Thus, the data space was dictated only by the different training sets selected at each iteration, and both the subset of PCs and the model coefficients were new at each iteration as well.

(LMO) approach where both the variables (PCs) calculation, the PCs selection, and the model coefficients calculation are repeated at each iteration on the 90% remaining chemicals.

Figure 6 shows that the statistical validation procedures can generate quite different estimates of external predictivity and that the "predicted" predictivity may fade out as the cross-validation criteria become more stringent and approach a real external data testing, where the training set and the test set are completely independent. One lesson of this exercise is that the incomplete exclusion of test set information from the training set can generate an overoptimistic view of prospective performance. Thus, it is necessary to repeat the feature construction/selection process separately for each cross-validation iteration in order to keep the training set and the test set independent. In addition, the test set for cross-validation has to be extracted randomly from the initial (general) training set.

It also appears that, even though the most stringent cross-validation attributes some statistically acceptable validity to the model, the model predicted completely random the 67 NTP external test set chemicals (in spite of the fact that they are in the same range of descriptors values of the training set). This observation parallels the huge variability of external predictivity shown by the commercial nonlocal prediction software systems in Figures 1−3.

CONCLUSIONS

The present work indicates that good local QSARs for congeneric chemicals can attain 70−100% correct external predictions if they are used to discriminate between inactive and active (mutagens, carcinogens) chemicals. This result indicates that these QSARs can be used with good reliability for applicative purposes (e.g., enriching the target for priority setting). It should be emphasized that the above values compare favorably with the known limits in reproducibility of the experimental systems.[24] On the other hand, the local QSARs for the potency of the active compounds exhibit a much lower reliability.

A remarkable result of this study is that a number of internal, statistical validation methods (e.g., cross-validation, etc. . .), which are often assumed to be good diagnostics for predictivity, do not correlate well with the predictivity of the QSARs when challenged in real external prediction tests. In our opinion, these statistical procedures should be better seen as tools to characterize and describe the data sets and to assist in the development of the model. The value of a local QSAR should be better assessed from the combination of the goodness of fitting and of the coherence of the model with the general knowledge in the fields of QSAR, chemistry, and biology.[7,25,26] Mechanistic aspects, expertise on the meaning of chemical descriptors, and comparison with other QSARs are of central importance when assessing predictivity and applicability domain of local QSAR models.

The nonlocal QSARs for noncongeneric chemicals seem to be much more prone to erratic predictions. Unfortunately, the problem of modeling large sets of chemicals acting by different mechanisms makes it almost unavoidable for the use of large numbers of descriptors devoid of any mechanistic meaning (in this case, the PCs), thus the results cannot be evaluated except in statistical terms. Often this procedure is misleading, and, as shown by the analyses presented in Figures 1−3, models that look good at the "average" scale do not work with the same efficiency in the various segments of the chemical space.

In our opinion, local, mechanistically based QSARs should be built and used whenever possible. With the large and

friendly usable databases available today,[27,12] compounds similar to the target to be predicted can often be found, and a local QSAR can be built. On the other hand, the local QSARs suffer a number of practical limitations. For example, the QSARs found to be promising in our study for the ECB cover only a portion of e.g., the EU high production volume chemicals[9] and three chemical classes. Moreover, the mechanistic understanding of many human health effect end points is not possible at this time (e.g., some nongenotoxic carcinogenicity mechanisms) and cannot contribute to the validation of the relative local QSARs. Thus in many instances there is no alternative to the use of nonlocal models. In these cases, stringent checks of the predictivity in the chemical space portion of interest should be preliminarily performed.

APPENDIX: CHARACTERIZATION OF THE LOCAL QSAR MODELS ANALYZED

Mutagenicity of Aromatic Amines in *S. typhimurium* TA98 and TA100 Strains, with S9 Metabolic Activation.[28]

$$\log TA98 = 1.08\ (\pm 0.26)\ \log P + 1.28\ (\pm 0.64)\ HOMO - 0.73\ (\pm 0.41)\ LUMO + 1.46\ (\pm 0.56)\ IL + 7.20\ (\pm 5.4) \quad \textbf{QSAR1}$$

$$n = 88,\ r = 0.898,\ s = 0.860$$

$$\log TA100 = 0.92\ (\pm 0.23)\ \log P + 1.17\ (\pm 0.83)\ HOMO - 1.18\ (\pm 0.44)\ LUMO + 7.35\ (\pm 6.9) \quad \textbf{QSAR2}$$

$$n = 67,\ r = 0.877,\ s = 0.708$$

where logTA98 and logTA100 are the mutagenic potency as log(revertants/nmol), logP is the logarithm of the octanol/water partition coefficient, HOMO is the energy of the highest occupied molecular orbital, LUMO is the energy of the lowest unoccupied molecular orbital, IL is an indicator variable, with value 1 is for compounds with three or more fused rings.

Carcinogenic Potency of Aromatic Amines in Rodents..[29]

$$BRM = 0.88\ (\pm 0.27)\ \log P{*}I(monoNH2) + 0.29\ (\pm 0.20)\ \log P{*}I(diNH2) + 1.38\ (\pm 0.76)\ HOMO - 1.28\ (\pm 0.54)\ LUMO - 1.06\ (\pm 0.34)\ \Sigma MR2,6 - 1.10\ (\pm 0.80)\ MR3 - 0.20\ (\pm 0.16)\ Es(R) + 0.75\ (\pm 0.75)\ I(diNH2) + 11.16\ (\pm 6.68) \quad \textbf{(QSAR3)}$$

$$n = 37,\ r = 0.907,\ r2 = 0.823,\ s = 0.381,\ F = 16.3,\ P < 0.001$$

$$BRR = 0.35\ (\pm 0.18)\ \log P + 1.93\ (\pm 0.48)\ I(Bi) + 1.15\ (\pm 0.60)\ I(F) - 1.06\ (\pm 0.53)\ I(BiBr) + 2.75\ (\pm 0.64)\ I(RNNO) - 0.48\ (\pm 0.30) \quad \textbf{(QSAR4)}$$

$$n = 41,\ r = 0.933,\ r2 = 0.871,\ s = 0.398,\ F = 47.4,\ P < 0.001$$

where BRM = log (MW/TD50)mouse, BRR = log (MW/TD50)rat, TD50 is the daily dose required to halve the probability for an experimental animal of remaining tumorless to the end of its standard life span, ΣMR2,6, is the sum of molar refractivity of substituents in the ortho-positions of the aniline ring, MR3 is the molar refractivity of substituents in the meta-position of the aniline ring, Es(R) is the Charton's substituent constant for substituents at the functional amino group, I(monoNH2) = 1 for compounds with only one amino group, I(diNH2) = 1 for compounds with more than one amino group, I(Bi) = 1 for biphenyls, I(I(BiBr) = 1 for biphenyls with a bridge between the phenyl rings, I(RNNO) = 1 for compounds with the group N(Me)NO, and I(F) = 1 for aminofluorenes.

Mutagenic Activity in *S. thyphimurium* TA98 and TA100 (+S9).[30]

TA98

$$w = -0.34\ HOMO + 0.86\ LUMO - 0.28\ MR5 + 0.48\ MR6 + 0.67\ Idist \quad \textbf{(QSAR5)}$$

where *w*(mean; class1) = 1.68 N1 = 25 (nonmutagens), and *w*(mean;class2) = −0.49 N2 = 86 (mutagens). The squared canonical correlation of the model is 0.46. The equation correctly reclassified 89.2% (accuracy) of the compounds [class1, nonmutagens, 88.0% (specificity); class2, mutagens, 89.5% (sensitivity)].

TA100

$$w = -0.67\ HOMO + 0.75\ LUMO + 0.39\ MR2 + 0.38\ MR3 + 0.44\ MR6 + 0.62\ Idist \quad \textbf{(QSAR6)}$$

where *w*(mean; class1) = 1.21 N1 = 47 (nonmutagens), and *w*(mean;class2) = −0.89 N2 = 64 (mutagens). The squared canonical correlation of the model is 0.52. The equation correctly reclassified 87.4% (accuracy) of the compounds [class1, nonmutagens, 95.7% (specificity); class2, mutagens, 81.3% (sensitivity)]. N1 = number of nonmutagens (class1), N2 = number of mutagens (class2), HOMO is the energy of the highest occupied molecular orbital, LUMO is the energy of the lowest unoccupied molecular orbital, MR2, MR3, MR5, and MR6 are the molar refractivity contributions of substituents in positions 2, 5, and 6 to the amino group, and Idist is 1 for compounds with crowded substituents on the positions 3′-, 4′-, and 5′- of 4-aminobiphenyl.

Carcinogenicity of Aromatic Amines in Rodents.[31]

$$w = -2.86\ L(R) + 2.65\ B5(R) - 1.16\ HOMO + 1.76\ LUMO + 0.40\ MR3 + 0.58\ MR5 + 0.54\ MR6 - 1.55\ I(An) + 0.74\ I(NO2) - 0.55\ I(BiBr) \quad \textbf{(QSAR7)}$$

where *w*(mean,class1) = −1.56 N1 = 13, and *w*(mean, class2) = 0.38 N2 = 53. The squared canonical correlation of the model is 0.38. The equation correctly reclassified 87.9% (accuracy) of the compounds [class1, noncarcinogens, 84.6% (specificity);class2, carcinogens, 88.7% (sensitivity)].

$$w = -3.42L(R) + 3.11\ B5(R) - 1.57\ HOMO + 2.19\ LUMO + 0.66\ MR3 + 0.65\ MR5 + 0.54\ MR6 - 1.64\ I(An) + 0.57\ I(NO2) - 0.63\ I(BiBr) \quad \textbf{(QSAR8)}$$

where *w*(mean,class1) = −2.04 N1 = 12, and *w*(mean,class2) = 0.47 N2 = 52. The squared canonical correlation of the model is 0.50. The equation correctly reclassified 93.7.% (accuracy) of the compounds [class1, noncarcinogens, 92.7% (specificity);class2, carcinogens, 94.2% (sensitivity)]. N1 = number of noncarcinogens (class1), N2 = number of carcinogens (class2), L(R) is the sterimol length, B5(R) is the sterimol maximal width, HOMO is the energy of the highest occupied molecular orbital, LUMO is the energy of the lowest unoccupied molecular orbital, MR3, MR5, and MR6 are the molar refractivity contributions of substituents in positions 3, 5, and 6 to the amino group, I(An) is 1 for anilines, I(NO2) is 1 for the presence of a NO2 group, and I(BiBr) is 1 for biphenyls with a bridge between the phenyl rings.

Mutagenicity of Aromatic and Heteroaromatic Nitro Compounds in *S. typhimurium* Strain TA98.[32]

$$\log \mathrm{TA98} = 0.65\ (\pm 0.16)\ \mathrm{logP} - 2.90\ (\pm 0.59)\ \log(\beta\ 10^{\mathrm{logP}} + 1) - 1.38\ (\pm 0.25)\ \mathrm{LUMO} + 1.88\ (\pm 0.39)\ \mathrm{Il} - 2.89\ (\pm 0.81)\ \mathrm{Ia} - 4.15\ (\pm 0.58) \quad \textbf{(QSAR9)}$$

$n = 188$, $r = 0.900$, $s = 0.886$, logP0 = 4.93, $\log\beta = 5.48$, F1,18148.6

where Il is 1 for compounds with 3 or more fused rings, and Ia is 1 for 5 substances of the set that are much less active than expected.

Mutagenicity of Nitroarenes in *S. typhimurium* TA100, without Metabolic Activation.[32]

$$\log \mathrm{TA100} = 1.20\ (\pm 0.15)\ \mathrm{logP} - 3.40\ (\pm 0.74)\ \log(\beta\ 10^{\mathrm{logP}} + 1) - 2.05\ (\pm 0.32)\ \mathrm{LUMO} - 3.50\ (\pm 0.82)\ \mathrm{Ia} + 1.86\ (\pm 0.74)\ \mathrm{Iind} - 6.39\ (\pm 0.73) \quad \textbf{(QSAR10)}$$

$n = 117$, $r = 0.886$, $s = 0.835$, logP0 = 5.44 (± 0.24), $\log\beta = -5.7$, F1,110 = 24.7

where Ia is 1 for compounds where acenthrylene ring is present, and Iind is 1 for the 1- and 2-methylindazole derivatives.

Mutagenicity of α-*β* Unsaturated Aldehydes in *S. typhimurium* TA100.[33]

$$\mathrm{Negatives} = -47.13331 + 38.24641\ \mathrm{MR} - 31.77763\ \mathrm{logP} + 30.46.799\ \mathrm{LUMO}$$

$$\mathrm{Positives} = -20.52153 + 25.41469\ \mathrm{MR} - 21.45102\ \mathrm{logP} + 19.77513\ \mathrm{LUMO} \quad \textbf{(QSAR13)}$$

$n = 20$; 100% correct reclassification; 3/20 errors in cross-validation.

REFERENCES AND NOTES

(1) Hansch, C. On the Predictive Value of QSAR. In *Biological Activity and Chemical Structure*; Keverling Buisman, J. A., Ed.; Elsevier: Amsterdam, 1977.

(2) Bristol, D. W.; Wachsman, J. T.; Greenwell, A. The NIEHS Predictive-Toxicology Evaluation Project:Chemcarcinogenicity Biossay. *Environ. Health Perspect.* **1996**, *104*, 1001–1010.

(3) Zeiger, E.; Ashby, J.; Bakale, G.; Enslein, K.; Klopman, G.; Rosenkranz, H. S. Prediction of Salmonella Mutagenicity. *Mutagenesis* **1996**, *11*, 474–484.

(4) Benigni, R.; Zito, R. The Second National Toxicology Program Comparative Exercise on the Prediction of Rodent Carcinogenicity: Definitive Results. *Mutat. Res. Rev.* **2004**, *566*, 49–63.

(5) Eriksson, L.; Jaworska, J. S.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.

(6) Golbraikh, A.; Tropsha, A. Beware of Q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

(7) Kubinyi, H. Validation and Predictivity of QSAR Models. In *QSAR and Molecular Modelling in Rational Design of Bioactive Molecules-.The 15th European Symposium on Quantitative Structure-Activity Relationships and Molecular Modelling*; Aki-Sener, E., Yalcin, I., Eds.; CADDS: Ankara, 2005.

(8) Benigni, R.; Netzeva, T. I.; Benfenati, E.; Bossa, C.; Franke, R.; Helma, C.; Hulzebos, E.; Marchant, C. A.; Richard, A. M.; Woo, Y. T.; Yang, C. The Expanding Role of Predictive Toxicology: an Update on the (Q)SAR Models for Mutagens and Carcinogens. *J. Environ. Sci. Health, Part Cqqq* **2007**, *25*, 53–97.

(9) Benigni, R.; Bossa, C.; Netzeva, T. I.; Worth, A. P. Collection and evaluation of (Q)SAR models for mutagenicity and carcinogenicity. EUR 22772 EN. 2007; Office for the Official Publications of the European Communities. EUR - Scientific and Technical Research Series: Luxenbourg, 2007; pp 1−118.

(10) Benigni, R.; Bossa, C. Structural Alerts of Mutagens and Carcinogens. *Curr. Comput.-Aided Drug Des.* **2006**, *2*, 169–176.

(11) Benigni, R.; Giuliani, A. Quantitative Modeling and Biology: the Multivariate Approach. *Am. J. Physiol.* **1994**, *266*, R1697–R1704.

(12) Benigni, R.; Bossa, C.; Richard, A. M.; Yang, C. A Novel Approach: Chemical Relational Databases, and the Role of the ISSCAN Database on Assessing Chemical Carcinogenicity. *Ann. Ist. Super. Sanità* **2008**, *44*, 48–56.

(13) Benigni, R.; Andreoli, C.; Giuliani, A. QSAR Models for Both Mutagenic Potency and Activity: Application to Nitroarenes and Aromatic Amines. *Environ. Mol. Mutagen.* **1994**, *24*, 208–219.

(14) Benigni, R. Structure-Activity Relationship Studies of Chemical Mutagens and Carcinogens: Mechanistic Investigations and Prediction Approaches. *Chem. Rev.* **2005**, *105*, 1767–1800.

(15) Enslein, K. The Future of Toxicity Prediction With QSAR. *In Vitro Toxicol.* **1993**, *6*, 163–169.

(16) Richard, A. M. Structure-Based Methods for Predicting Mutagenicity and Carcinogenicity: Are We There Yet. *Mutat. Res.* **1998**, *400*, 493–507.

(17) Greene, N.; Judson, P. N.; Langowski, J. J.; Marchant, C. A. Knowledge-Based Expert Systems for Toxicity and Metabolism Prediction: DEREK, StAR and METEOR. *SAR QSAR Environ. Res.* **1999**, *10*, 299–314.

(18) Klopman, G. Multicase 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176–184.

(19) Rosenkranz, H. S. SAR in the Assessment of Carcinogenesis: the MultiCASE Approach. In *Quantitative Structure-Activity Relationship (QSAR) Models of Chemical Mutagens and Carcinogens*; Benigni, R., Ed.; CRC Press: Boca Raton, 2003; Chapter 6, pp 175−206.

(20) Provost, F.; Fawcett, T. Robust Classification for Imprecise Environment. *Machine Learn. J.* **2001**, *42*, 5–11.

(21) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Khan, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G.; Perkins, R.; Roberts, D. W.; Schult, T. W.; Stanton, S. T.; van de Sandt, J. J. M.; Tong, W.; Veith, G. D.; Yang, C. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. *ATLA-Altern. Lab. Animals* **2005**, *33*, 1–19.

(22) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.

(23) Benigni, R.; Gallo, G.; Giorgi, F.; Giuliani, A. On the Equivalence Between Different Descriptions of Molecules: Value for Computational Approaches. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 575–578.

(24) Piegorsch, W. W.; Zeiger, E. Measuring intra-assay agreement for the Ames Salmonella assay. In *Statistical methods in toxicology;* Hothorn, L., Ed.; Springer-Verlag: Berlin, 1991; Vol. 43, pp 35−41.

(25) Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. D. Chem-Bioinformatics: Comparative QSAR at the Interface Between Chemistry and Biology. *Chem. Rev.* **2002**, *102*, 783–812.

(26) Unger, S. H.; Hansch, C. On Model Building in Structure-Activity Relationships. A Reexamination of Adrenergic Blocking Activity of Beta-Halo-Beta-Arylalkylamines. *J. Med. Chem.* **1973**, *16*, 754–749.

(27) Yang, C.; Richard, A. M.; Cross, K. P. The Art of Data Mining the Minefields of Toxicity Databases to Link Chemistry to Biology. *Curr. Comput.-Aided Drug Des.* **2006**, *2*, 135–150.

(28) Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella Typhimurium TA98 and TA100. *Environ. Mol. Mutagen.* **1992**, *19*, 37–52.

(29) Benigni, R.; Giuliani, A.; Franke, R.; Gruska, A. Quantitative Structure-Activity Relationships of Mutagenic and Carcinogenic Aromatic Amines. *Chem. Rev.* **2000**, *100*, 3697–3714.

(30) Benigni, R.; Bossa, C.; Netzeva, T. I.; Rodomonte, A.; Tsakovska, I. Mechanistic Qsar Of Aromatic Amines: New Models For Discriminating Between Homocyclic Mutagens And Nonmutagens, And Validation Of Models For Carcinogens. *Environ. Mol. Mutagen.* **2007**, *48*, 754–771.

(31) Franke, R.; Gruska, A.; Giuliani, A.; Benigni, R. Prediction of Rodent Carcinogenicity of Aromatic Amines: a Quantitative Structure-Activity Relationships Model. *Carcinogenesis* **2001**, *22*, 1561–1571.

(32) Debnath, A. K.; Lopez de Compadre, R. L.; Shusterman, A. J.; Hansch, C. Quantitative Structure-Activity Relationship Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 2. Mutagenicity of Aromatic and Heteroaromatic Nitro Compounds in Salmonella Typhimurium TA100. *Environ. Mol. Mutagen.* **1992**, *19*, 53–70.

(33) Benigni, R.; Conti, L.; Crebelli, R.; Rodomonte, A.; Vari′, M. R. Simple and α-*β* -Unsaturated Aldehydes: Correct Prediction of Genotoxic Activity Through Structure-Activity Relationship Models. *Environ. Mol. Mutagen.* **2005**, *46*, 268–280.

(34) Pearl, G. M.; Livingstone-Carr, S.; Durham, S. K. Integration of Computational Analysis As a Sentinel Tool in Toxicologic Assessments. *Curr. Top. Med. Chem.* **2001**, *1*, 247–255.

(35) Benigni, R. The First U.S. National Toxicology Program Exercise on the Prediction of Rodent Carcinogenicity: Definitive Results. *Mutat. Res.* **1997**, *387*, 35–45.

(36) ECETOC. (Q)SARs: Evaluation of the commercially available software for human health and environmental endpoints with respect to chemical management applications; Technical Report 89; ECETOC: Brussels, 2003.

(37) Snyder, R. D.; Pearl, G. M.; Mandakas, G.; Choy, W. N.; Goodsaid, F.; Rosenblum, I. Y. Assessment of the Sensitivity of the Computational Programs DEREK, TOPKAT, and MCASE in the Prediction of the Genotoxicity of Pharmaceutical Molecules. *Environ. Mol. Mutagen.* **2004**, *43*, 143–158.

(38) Hulzebos, E. M.; Posthumus, R. (Q)SARs: Gatekeepers Against Risk on Chemicals. *SAR QSAR Environ. Res.* **2003**, *14*, 285–316.

(39) Crettaz, P.; Benigni, R. Prediction of Rodent Carcinogenicity of 60 Pesticides by the DEREKfW Expert System. *J. Chem. Inf. Model.* **2005**, *45*, 1864–1873.

(40) Cariello, N. F.; Wilson, J. D.; Britt, B. H.; Wedd, D. J.; Burlinson, B.; Gombar, V. K. Comparison of the Computer Programs DEREK and Topkat to Predict Bacterial Mutagenicity. *Mutagenesis* **2002**, *17*, 321–329.

(41) Greene, N. Computer Systems for the Prediction of Toxicity: an Update. *Adv. Drug Delivery Rev.* **2002**, *54*, 417–431.

(42) Prival, M. J. Evaluation of the TOPKAT System for Predicting the Carcinogenicity of Chemicals. *Environ. Mol. Mutagen.* **2001**, *37*, 55–69.

(43) Mueller, F.; Simon-Hettich, B.; Kramer, P. J. An Evaluation of the Predictability of the Ames-Test Mutagenicity Using the TOPKAT TM Program. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **2000**, *Suppl. to vol. 361*, R173.

CI8000088