Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity

Pavel G. Polishchuk,[†] Eugene N. Muratov,*[,†,‡] Anatoly G. Artemenko,[†] Oleg G. Kolumbin,[§] Nail N. Muratov,[||] and Victor E. Kuz'min[†]

Laboratory on Theoretical Chemistry, A.V. Bogatsky Physical-Chemical Institute NAS of Ukraine, Odessa 65080, Ukraine, Laboratory of Molecular Modeling, School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599, Department of Chemistry, Pridnestrovskij State University, Tiraspol, MD-3300, Chemical-Technological Department, Odessa National Polytechnic University, Odessa 65000, Ukraine

Received June 7, 2009

This work is devoted to the application of the random forest approach to QSAR analysis of aquatic toxicity of chemical compounds tested on *Tetrahymena pyriformis.* The simplex representation of the molecular structure approach implemented in HiT QSAR Software was used for descriptors generation on a two-dimensional level. Adequate models based on simplex descriptors and the RF statistical approach were obtained on a modeling set of 644 compounds. Model predictivity was validated on two external test sets of 339 and 110 compounds. The high impact of lipophilicity and polarizability of investigated compounds on toxicity was determined. It was shown that RF models were tolerant for insertion of irrelevant descriptors as well as for randomization of some part of toxicity values that were representing a "noise". The fast procedure of optimization of the number of trees in the random forest has been proposed. The discussed RF model had comparable or better statistical characteristics than the corresponding PLS or KNN models.

INTRODUCTION

Toxic environmental chemicals may be damaging to the environment and human health, and therefore, they represent a considerable danger to society. Unfortunately, there is a great gap in the number of chemical compounds with measured physical−chemical properties and toxicity and the ones that still require these data. In the 1990s, the U.S. Environmental Protection Agency, Office of Toxic Substances (OTS) listed approximately 70000 industrial chemicals. About 1000 chemicals have been added each year. However, even simple toxicological experiments have not been carried out.

Chemical toxicity can be associated with many hazardous biological effects such as gene damage, carcinogenicity, or induction of lethal rodent or human diseases. It is important to evaluate the toxicity of all commercial chemicals, especially the high production volume (HPV) compounds as well as drugs or drug candidates, before releasing them into the market.[1] To address this need, standard experimental protocols have been established by the chemical industry, pharmaceutical companies, and government agencies to test chemicals for their toxic potential.

Although such experimental protocols for toxicity testing have been developed for many years and the cost of testing of an individual compound has been reduced significantly, computational chemical toxicology continues to be a viable approach to reduce the amount of efforts and cost of experimental toxicity assessment.[1] Surely, the computation modeling will be a good alternative to experimental testing. In the near future and further on it will result in significant savings that could be achieved if the potential property of a new chemical could be predicted before its synthesis and biological testing. To address this challenge, many quantitative structure activity relationship (QSAR) studies have been conducted and reported for different toxicity endpoints.[1−5]

The European Union (EU) has recently approved REACH (Registration, Evaluation, and Authorization of Chemicals) regulation that will create a list of chemicals used in the EU. This law requires assessment of physical−chemical properties and adverse effects (e.g., carcinogenic and mutagenic properties) of all compounds, which are produced in excess of 1 ton/year, which will lead to the registration of more than 30000 compounds. The implementation of REACH requires demonstration, by means of experimental tests, of the safety of chemicals manufacturing and safety of their usage throughout the supply chain. The total cost of tests required for the registration of compounds is estimated to be 5 billion € during the next 11 years.[6] REACH recommends the usage of nonanimal testing methods, e.g., QSAR/QSPR approaches, in order to decrease the number and costs of animal tests. For example, the REACH system requires that nonanimal methods should be used for the majority of tests in the 1−10 ton band of chemicals produced in large volumes.

It is well-known[7−9] that the most critical limitation of many QSAR studies[10−12] is their low external predictive power, i.e., their ability to predict accurately the underlying endpoint toxicity for compounds that were not used for model development. The low external prediction accuracy of QSAR models in spite of the high accuracy of the training set models is a well-known Kubinyi paradox named by van Drie in 2003[13] but formulated by the author in the last century.[9]

* Corresponding author phone: +380682642192; e-mail: murik@email.unc.edu.
[†] A.V. Bogatsky Physical-Chemical Institute NAS of Ukraine.
[‡] University of North Carolina.
[§] Pridnestrovskij State University.
[||] Odessa National Polytechnic University.

10.1021/ci900203n CCC: $40.75
Published on Web 10/27/2009

There could be many reasons for this discrepancy between the internal (cross-validation) and external predictive power of QSAR models. The most common is that training set models are based on data interpolation, and therefore, they inherently have limited applicability in the chemical space, whereas any external prediction implies inherent and frequently excessive extrapolation of the training set models. The poor external predictive power of the QSAR models could be due to the incorrect usage or lack of external validation during the modeling process. An even more important issue is the quality of the experimental data, which must be carefully curated prior to modeling. Moreover, each statistical method used in QSAR studies has its own specific advantages, weaknesses, and practical constraints, so it is important, although to a lesser degree than data curation, to select the most suitable QSAR methodology for each specific toxicity endpoint. As it was shown in ref 1, consensus modeling can be the most appropriate for such kind of investigations. Thus, the first international virtual collaboration consisting of six independent groups with shared interests in computational chemical toxicology was formed in 2007[1] in order to unite the efforts in obtaining predictive QSAR models in the mentioned area. An aqueous toxicity data set containing 1093 unique compounds tested in the same laboratory over a decade against *Tetrahymena pyriformis* have been compiled.[1] The training set consists of 644 compounds, and 2 external test sets consist of 339 and 110 compounds. Every participating group used their own QSAR tools for model development. In total, 15 different QSAR models of aquatic toxicity were developed and resulted in a consensus model by averaging the predicted aquatic toxicity for every compound using all 15 models during this virtual collaboration. The power of collaborative and consensual approach to QSAR model development has been shown by that study.[1]

Earlier it was shown that random forest (RF)[14] could be a very effective solution of QSAR tasks,[15] but this method has not been widely used yet.[16−21] RF methodology seems to be very helpful because every forest represents a consensus nonlinear model derived from a large number of single models (trees). Moreover, RF has the following important advantages: (i) RF models are quite resistant to overfitting. (ii) RF does not require rather complicated and time-consuming processes of variable selection. (iii) Compounds with various mechanism of actions could be studied within the same (single) training set.

The simplex representation of molecular structure (SiRMS) QSAR approach implemented in HiT QSAR Software[22] has been used. This method showed good results in previous QSAR studies.[23−26] One of the main advantages of SiRMS is a good interpretability of obtaining models.

Thus, the aims of investigation were the (i) development of robust, externally predictive and interpretable models on the basis of simplex descriptors and a RF statistical approach; (ii) comparison of RF models with partial least squares or projections to latent structures (PLS) and *k*-nearest-neighbors (KNN) consensus models developed using simplex descriptors and a collaborative consensus model from a joint study;[1] and (iii) investigation of influence of tuning parameters and noise insertion on RF models behavior and optimization of RF model development.

MATERIALS AND METHODS

Investigated Compounds. The growth inhibition of the aquatic ciliate *Tetrahymena pyriformis* is a commonly accepted toxicity screening tool. The *Tetrahymena pyriformis* toxicity data set used in this study has been taken from ref 1 and represents a comprehensive toxicity data set consisting of 1093 compounds. Toxicity of each compound has been expressed as the inverse logarithm of 50% inhibition of *Tetrahymena pyriformis* growth concentration ($pIGC_{50}$) values. The work set has been divided into three parts according to original study:[1] (i) the training set of 644 compounds; (ii) first external validation set consisting of 339 compounds, and (iii) second external validation set of 110 compounds. All three data sets consist of similar fractions of compounds with low, intermediate, and high toxicity values. A complete list of the compounds and their observed and predicted toxicity values for all three data sets is represented in Table 1 of the Supporting Information.

Simplex Representation of Molecular Structure. Bounded and unbounded two-dimensional (2D) simplex descriptors (tetratomic fragments with fixed composition and topological structure) were used for molecular structure representation. Not only atom type but also other physical−chemical characteristics of an atom, e.g., partial charge,[27] lipophilicity,[28] refraction,[29] and the ability for an atom to be a donor or acceptor in hydrogen bond formation were used for atom differentiation in SiRMS. For these atom characteristics, which have real values (charge, lipophilicity, and refraction), the transition from the values range to definite discrete groups was carried out at the preliminary stage. In the given work, the atoms were divided into groups corresponding to their (i) partial charge $A \le -0.05 < B' 0 < C \le 0.05 < D$, (ii) lipophilicity $A \le -0.5 < B \le 0 < C \le 0.5 < D$, and (iii) refraction $A \le 1.5 < B \le 3 < C \le 8 < D$. For the atom H-bond characteristic, the atoms have been divided into three groups: A (acceptor of hydrogen in H-bond), D (donor of hydrogen in H-bond), and I (indifferent atom). The usage of diverse variants of differentiation of simplex vertexes (atoms) represents the principal feature of the offered approach. The main advantages of SiRMS are the opportunity of analysis of molecules with noticeable structural differences as well as the possibility to reveal individual molecular fragments (simplex combinations) promoting or interfering with investigated activity. Because SiRMS has been well-described earlier, more detailed information about the methodology can be found in ref 22.

Log P[30] and molecular refraction[29] were used as two integral descriptors additionally to more than 6000 calculated 2D simplex descriptors.

Statistical Approaches and Characteristics. All QSAR models despite the approach used were developed on the basis of the training set only. Various (depending on the method) selection criteria were used. Selected models were united in corresponding consensus ensembles. Only after that were obtained consensus models applied to test set compounds in a "blind prediction" mode.

Random Forest. Random forest models were constructed according to the described original RF algorithm.[14] RF is an ensemble of single decision trees. This ensemble produces a corresponding number of outputs. Outputs of all trees are aggregated to obtain one final prediction. In regression, a

task final predicted value is the average of the individual tree predictions. Each tree has been grown as follows: (i) A bootstrap sample, which will be a training set for the current tree, is produced from the whole training set of N compounds. Compounds which are not in the current tree training set are placed in an out-of-bag (OOB) set (OOB set size is $\sim N/3$). (ii) The best split by CART algorithm[31] among the m randomly selected descriptors from whole set of M ones in each node is chosen. The value of m is just one tuning parameter for which RF models are sensitive. (iii) Each tree is grown to the largest possible extent. There is no pruning.

Because RF possesses its own reliable statistical characteristics (on the basis of OOB set prediction), which could be used for validation and model selection, no cross-validation has been performed. It was shown[15] that prediction accuracy of an OOB set and a 5-fold external cross-validation procedure is near the same. The major criterion for estimation of the predictive ability of the RF models and model selection is the value of R^2_{OOB}.

Descriptor Importance and RF Specific Domain Applicability Approach. Estimation of domain of applicability (DA) is based on the distance matrix calculated with consideration of descriptors importance. Descriptors importance was calculated by following procedure:(i) prediction of OOB set on the basis of original descriptor values, (ii) randomization of selected descriptor values for all training set molecules, (iii) prediction of OOB set on the basis of randomized descriptor values, and (iv) calculation of differences between the OOB set prediction values with and without randomization for each tree and averaging of the obtained values.

$$E_i = \begin{cases} = 0, if\ (A^i_{OOB} - \bar{A}^i_{OOB}) <= 0 \\ = (A^i_{OOB} - \bar{A}^i_{OOB}), if\ (A^i_{OOB} - \bar{A}^i_{OOB}) > 0 \end{cases} \quad (1)$$

E_i is the decreasing of the OOB set prediction accuracy of i-th tree; A^i_{OOB} is the OOB set prediction accuracy of i-th tree without descriptor values randomization; and $\bar{A}^i_{OOB}$ is the OOB set prediction accuracy of i-th tree after descriptor values randomization.

According to formula 1, if the OOB set prediction accuracy increases after descriptor values randomization value of E_i sets to zero, such a descriptor can be considered as unimportant for property prediction.

$$IMP = \frac{1}{T}\sum_{i=1}^{T} E_i \quad (2)$$

IMP is the descriptor importance; T is the number of trees in the random forest model; and E_i is the decreasing of the OOB set prediction accuracy of i-th tree according to the randomization procedure.

The most important descriptors have the largest IMP values. Ranking of all descriptors by their importance allows one to define their impact to the target property.

The tree approach[32] has been used for DA estimation (Figure 1). It consists of the following steps: (i) calculation of distances between all training set molecules, taking into account descriptors importance according to formula 3; (ii) building of an extreme short distances tree by iterative detection of the shortest distance in the obtained distances matrix until all molecules in the obtaining tree will be connected each other by one edge; (iii) calculation of the average distance $Dist_{av}$ and its root-mean-square deviation (σ) from the whole tree. Evidently, such distance is the characteristic of the average density of the molecules distribution in the structural space.

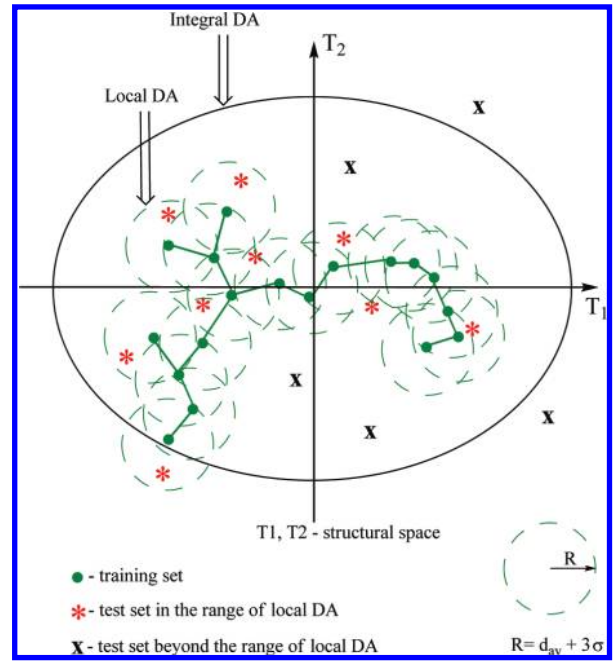

Figure 1. Tree DA approach.

$$Dist_{i,j} = \sum_{k=1}^{n} IMP_k^2 (D^i_k - D^j_k)^2 \quad (3)$$

$Dist_{i,j}$ is the distance in descriptors space between the i-th and j-th molecules; n is the number of descriptors with nonzero importance values; IMP_k is the k-th descriptor importance; and D^i_k and D^j_k are the values of the k-th descriptor for the i-th and j-th molecules, respectively.

Then, all of the points corresponding to test set molecules have been taken into account in the mentioned structural space. If any of the test set molecules has been situated on the distance larger than $Dist_{av} + 3\sigma$ from the nearest training set point, it means that this test set molecule is situated outside the DA. Respectively, molecules belonging to DA are situated on the distance less than $d_{av} + 3\sigma$ from the training set points.

Such an approach for DA estimation is similar to some extent to methods described in ref 33. Opposite to integral approaches,[24] where usually the convex region (polyhedron, ellipsoid) that could contain vast cavities has been determined in the structural space, the offered approach is local. Actually, the space of structural parameters has been analyzed locally, i.e., regions around every training set point are analyzed. The presence of cavities in the structural space, which corresponds to DA, is undesirable, and it has been eliminated in the given approach.

PLS and KNN. During the calculation, the training set (644 molecules) was additionally divided into modeling and internal test sets by the different ways described in ref 22 for the PLS method and in ref 34 for the KNN method. All the models selected for participation in consensus modeling satisfy the following criteria: ($R^2 \geq 0.81$; $Q^2 \geq 0.75$; and $R^2_{test} \geq 0.75$).

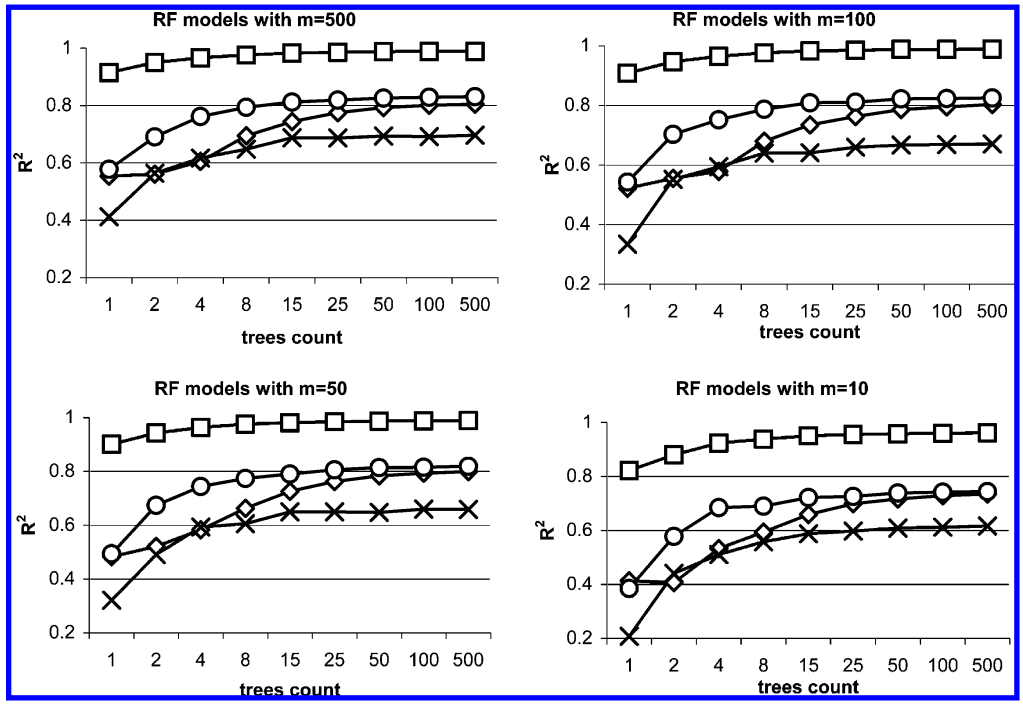

Figure 2. Average (per 10 models) values of R^2 for the training set (□), OOB set (◇), test set 1 (○), and test set 2 (×) for random forest models.

The PLS method[35,36] was used for statistical model development as a most popular linear statistical method. The genetic algorithm (GA),[37] trend−vector method,[38−40] and automatic variable selection (AVS) strategy[22] based on interactive[41] and evolutionary[42] variables selection were used for descriptors selection in PLS.

Briefly, this scheme can be represented in the following way: elimination of nonsignificant and highly correlated ($|r| = 0.9$) descriptors → TV procedure → AVS ↔ GA → AVS → selected QSAR models. Selection of the best QSAR models on every stage of the scheme was carried out according to maximum of fitness function (FF) criterion, where $FF = R^2 + 2Q^2$ and FF → max, i.e,. the best selected QSAR model is the model with the maximum FF value. For a more detailed description of the PLS method, please see refs 35 and 36.

The KNN method[34] employs the *k*-nearest-neighbors classification principle and the variable selection procedure. Briefly, a subset of *n* descriptors is selected randomly at the beginning of calculations. The number of used descriptors is a tuning parameter, and the training set models are developed with leave-one-out cross-validation, where each compound is eliminated from the training set and its activity is predicted as the average activity of *k* most similar molecules, where the value of *k* is optimized as well ($k = 1$ to 5). The similarity is characterized by the Euclidean distance between compounds in multi-dimensional descriptor space. A method of simulated annealing with the Metropolis-like acceptance criteria is used to optimize the selection of variables. The objective of this method is to obtain the best model selected by maximal Q^2 leave-one-out value by optimizing the number of descriptors and nearest neighbors. A more detailed description of the KNN method can be found in ref 34.

Consensus Modeling. In modern QSAR analysis, the most effective prognoses are realized as the result of usage of consensus approaches,[43] i.e., when not one but several single models have been used. Actually, the prognosis of an activity or property has been developed on the basis of averaging (by different schemes) the results of an ensemble of QSAR model applications. The success of the consensus approach depends on how the models selected are supplemented to each other, mutually compensating the errors. The combination of different QSAR models, evidently, allows ones to solve this task more successfully than any single QSAR approach. Seemingly, models included in this combination must be different in order to describe various response surfaces. This is why our models obtained, even using the same statistical approach and initial set of simplex descriptors, actually represent a consensus ensemble. Predicted toxicity for test set molecules using a united consensus model was calculated as a simple nonweighted average of individual predictions by all three models (PLS, KNN, and RF).

In order to make a comparison of the quality of our model with the results of the virtual collaboration study,[1] identical statistical parameters have been chosen to describe the predictivity of the model. There were:[1]

R^2_{abs} (determination coefficient) for each external validation sets and for the OOB set

$$R^2_{abs} = 1 - \sum_n (Y - \hat{Y})^2 / \sum_n (Y - \bar{Y})^2 \qquad (4)$$

where Y is the observed toxicity value; $\hat{Y}$ is the predicted toxicity value; $\bar{Y}$ is the average observed toxicity value for the training set; and n is the number of compounds in the corresponding test set.

MAE (mean absolute error of prediction)

$$MAE = \sum_Y |Y - \hat{Y}|/n \qquad (5)$$

where Y is the observed toxicity value; $\hat{Y}$ is the predicted toxicity value; and n is the number of compounds in the corresponding set.

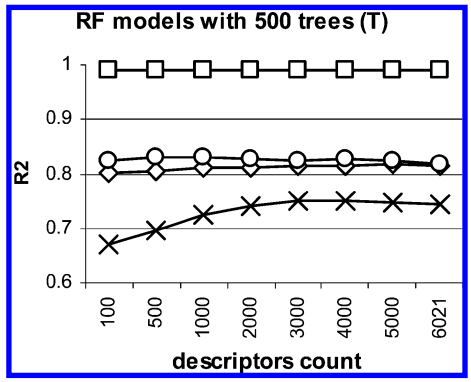

Figure 3. Dependence of $R2$ for the training set (□),OOB set (◊), test set 1 (○), and test set 2 (×) from the number of descriptors used in each splitting in RF models.

Many other statistical characteristics can be used to evaluate model performance; however, according to ref 1, we restricted ourselves to these three parameters that provide minimal but sufficient information concerning any ability of the model to reproduce the trends in experimental data for the test sets as well as mean accuracy of predicting all experimental values.

RESULTS AND DISCUSSION

Several RF model series with constrained m values in each series were obtained. Determination coefficients for the training set (R^2) and for the out-of-bag set (R^2_{OOB}) were considered as two major characteristics of the model. Ten RF models for each pair of values T (trees count) and m (descriptors count) were obtained. Average values of R^2 and R^2_{OOB} are depicted in Figure 2. In all cases, models with 500 trees have almost the same values of R^2_{OOB} than corresponding models with $T = 100$. Because the internal predictive ability becomes almost constant, it is possible to make a conclusion that value $T = 500$ is optimal for RF models development for a given task. The values of R^2 for the OOB set and for both test sets have close values, which were quite close in all cases (Figure 2). Analogous results were obtained by Svetnik et al.[15] So it can be suggested to use prediction statistics of the OOB set for a fair estimation of the predictive ability of RF models. These statistics can be very useful in analysis of small sets of compounds when it is undesirable to move compounds from the training set to the external validation set because of severe exhaustion of the former.

There is an empirical rule that $m = M/3$ (where M is the whole number of descriptors used) can be an optimal value in regression tasks.[15,19] We have investigated statistics of RF models for the following m values − 1000 (~$M/6$), 2000 (~$M/3$), 3000 (~$M/2$), 4000 (~$2M/3$), 5000 (~$5M/6$), and 6021 (M, all descriptors, bagging). As shown in Figure 3, prediction of compounds from the second test set was improved as the descriptors number used in each splitting was increased until m reached 3000. On the other side, prediction accuracy for compounds from the first test set was slightly decreased with an increasing m value. All models with $m \geq 3000$ had almost identical statistical values. Values of m from the range 2000−3000 could be considered as optimal, and on the whole, it was confirmed an aforementioned empirical rule.

Table 1. Statistical Characteristics of Obtained QSAR Models

	R^2_{ts1}	MAE_{ts1}	R^2_{ts2}	MAE_{ts2}
PCI_KNN	0.83	0.32	0.68	0.43
PCI_PLS	0.80	0.33	0.69	0.41
PCI_RF	0.83	0.30	0.74	0.38
PCI_(RF+KNN+PLS)	0.84	0.29	0.77	0.38
combinatorial consensus model[1]	0.85	0.29	0.67	0.39

For obtaining the final model, the following settings were chosen: number of trees $T = 500$ (defined earlier) and number of descriptors used for splitting at each node $m = 6021/3 \approx 2000$ (define according to the rule $m = M/3$). The obtained RF model (PCI_RF, Table 1), with $R^2 = 0.99$, $R^2_{OOB} = 0.81$, $R^2_{ts1} = 0.83$, and $R^2_{ts2} = 0.74$, is comparable or better than the combinatorial consensus model from ref 1 (Table 1). Increasing the number of trees in the last RF model to 750 does not effect the model quality, i.e., as mentioned above the optimal number of trees for this task is 500. Thus, the following procedure could be recommended for determination of the optimal number of trees: choosing a small enough number of descriptors (m) and models obtained with permanently increasing the number of trees until the statistical values for the OOB set would not be significantly changed. This procedure will be able to reduce the overall calculation time for the processing of large data sets.

Evaluation of local DA for the selected RF model allowed the determination of three compounds of test set 1 (nonylphenol, pentabromophenol, and *p*-phenylene diisothiocyanate) and one compound of test set 2 (4-phenyltoluene), which were outside of DA. Exclusion of these compounds had no significant influence on the predictive ability of the model.

Additionally 7 PLS and 84 KNN models ($R^2 \geq 0.81$, $Q^2 \geq 0.75$, and $R^2_{test} \geq 0.75$) have been obtained and selected as a base for two separate consensus models, PCI_PLS and PCI_KNN, respectively (Table 1). The PCI_RF model is comparable or better than both of them. Then successful united consensus model PCI_(RF+KNN+PLS) based on PCI_RF, PCI_KNN, and PCI_PLS QSAR models has been obtained. The obtained consensus model PCI_(RF+KNN+PLS) possesses near the same quality (Table 1) as the PCI_RF model and complicated combinatorial reference model from ref 1, nevertheless the latter was based on various statistical methods (KNN, PLS, MLR, SVM, and ASNN) and descriptors pools (MolConnZ, Dragon, CODESSA, ISIDA, electrotopological state (*E*-state) indices introduced by Hall and Kier, and "inductive" descriptors IND_I, which were developed in a series of papers by Cherkasov and coauthors). Moreover, according to the Fischer test, they are

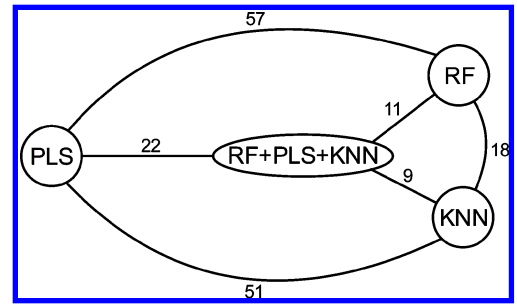

Figure 4. Dissimilarity measures between PCI_RF, PCI_KNN, PCI_PLS, and PCI_(RF+KNN+PLS) models.

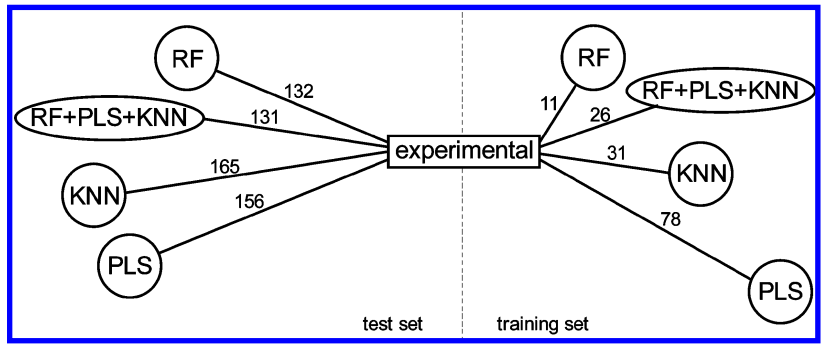

Figure 5. Dissimilarity measures between PCI_RF, PCI_KNN, PCI_PLS, and PCI_(RF+KNN+PLS) models and experimental toxicity values for the training set (right) and for combined test sets (left).

not statistically different. In our opinion, it can be explained by the good quality of a combi-QSAR reference model and a not comprehensive enough representation of the molecular structure in PCI_(RF+KNN+PLS). Really, linear (PLS in the PCI model and MLR and PLS in the reference model) and nonlinear (KNN and RF in the PCI model and KNN, ASNN, and SVM in the reference one) statistical approaches were used in both models. But there were only fragmental simplex descriptors in all PCI models against a vast variety of fragmental and integral descriptor pools (MolConnZ, Dragon, CODESSA, ISIDA, *E*-state indices, etc.).

Parameter ${}^{k}d_{ij} = (1 - r_{ij}) \times 10^3$ was used in the current study as a characteristic of the QSAR models dissimilarity in a space of investigated compounds. Here, r_{ij} is the correlation coefficient predicted by models i and j toxicity values for the k-th set of compounds (training set, test set, etc.).

Analysis of obtained QSAR models similarity and dissimilarity in a space of training set molecules shows that KNN and RF nonlinear models are the closest (Figure 4). They are also the closest to the consensus model. The multilinear PLS model is quite far from all of them as well as from the experimental toxicity values. It indicates an essential nonlinearity of response (toxicity) surface in structural space. The RF model is the closest to the experimental one in the training set space. Obviously, this model is better fitted than others (including consensus one). At the same time, the RF model is not overfitted because in the test set space it is the closest to the vector of experimental values along with the consensus model (Figure 5). Thus, the RF model is comparable with the consensus one by goodness-of-fit and predictivity. Probably, it related with the consensus nature of the RF model, where the final result is the averaging of predictions made by a large number of single decision trees.

The robustness of the obtained RF model to informational noise in the training set was also investigated. Noise was modeled by insertion of irrelevant descriptors or shuffling of toxicity values for some part of the training set compounds (*Y* randomization).

For 5% and 10% of randomly selected training set compounds, toxicity values were shuffled. Ten models were constructed for each case, and obtained statistics were averaged. This procedure leads to significant deterioration of OOB set statistics, but the prediction ability of new models was not damaged so dramatically. Thus, RF models could be considered as quite robust to errors in values of investigated activity.

Table 2. Statistical Characteristics for RF Models ($T = 500$, $m = M/3$) before and after *Y*-Randomization

Y-randomization part	R^2	R^2_{OOB}	R^2_{ts1}	R^2_{ts2}
0% (initial model)	0.99	0.81	0.83	0.74
5%	0.98	0.71	0.80	0.70
10%	0.98	0.62	0.78	0.68
100% (*Y*-scrambling)	0.97	−0.15	−0.15	−0.09

Table 3. Statistical Characteristics for RF Models ($T = 500$, $m = M/3$) with and without Irrelevant Descriptors

irrelevant descriptors part	R^2	R^2_{OOB}	R^2_{ts1}	R^2_{ts2}
0% (initial model)	0.99	0.81	0.83	0.74
5%	0.99	0.77	0.79	0.71
10%	0.99	0.76	0.78	0.71
20%	0.99	0.75	0.77	0.70

The *Y*-scrambling procedure[43] was used in order to confirm the absence of chance correlations in the initial RF model. As expected, no satisfactory models were obtained (Table 2).

In addition, RF models were obtained using irrelevant descriptors, which were represented by random numbers in the range [0; 1]. All obtained models possess 500 trees and $m = M/3$. As shown in Table 3, considerable degradation of model quality was observed when 5% of noise descriptors (from the whole number of descriptors) were inserted into the initial model. Further noise insertion only slightly decreased the performance of the models. Thus, RF models can be considered as tolerant to irrelevant descriptors.

Descriptor importance in the PCI_RF model was calculated by the randomization procedure.[14] This procedure was carried out twice, and identical results were obtained. It resulted in selection of 380 descriptors important for toxicity variation of investigated compounds. On the basis of only of the selected descriptors another RF model has been obtained ($T = 500$, $m = 380$). The new model has identical statistical characteristics ($R^2 = 0.99$, $R^2_{OOB} = 0.82$, $R^2_{abs1} = 0.83$, and $R^2_{abs2} = 0.74$) compared to those of the initial model PCI_RF (Table 1). This can be additional evidence that selected descriptors are the most influenced and can substantially determine toxicity variation of training and test sets compounds.

Calculated descriptors importance within each group of simplexes (by charge, lipophilicity, refraction, atom individuality, and donor−acceptor properties) were summarized, and the general influence of some physical−chemical properties of investigated compounds on toxicity variation was evaluated. Within the PCI_PLS model, such influence

was studied in the same way as described in ref 22. According to the PCI_RF model, lipophilicity and polarizability have the biggest impact on toxicity (31% and 29%, respectively). The PCI_PLS model results indicate a high influence of lipophilicity (~50%). Electrostatic factors and atom individuality are also important but to a lesser degree (~20% each). The obtained results are in agreement with the results of previous *Tetrahymena pyriformis* studies.[44]

CONCLUSIONS

An adequate PCI_RF QSAR model based on simplex descriptors and a RF statistical approach was developed for *Tetrahymena pyriformis* set and successfully validated on two external test sets. The obtained model is comparable or better than our consensus models obtained using KNN and PLS methods as well as the combinatorial consensus model from ref 1. The consensus model based on all three statistical approaches (PLS, KNN, and RF) is insignificantly better than any separate model.

It was shown that very often statistical values of RF models for the OOB set and for external test sets are quite similar. Although they cannot be used as a measure of the predictivity of models, the R^2_{OOB} value can serve as better criterion for model selection than, for example, Q^2. It can be especially useful when one has a small data set and removing compounds from the training set to the external test set is undesirable. The tolerance of RF models to noise (irrelevant descriptors and errors in target property values) was clearly demonstrated. The importance of hydrophobic factors for toxicity variation was determined.

Supporting Information Available: Compound names, smiles, and observed and predicted toxicity values for all models with and without DA consideration as well as observed versus predicted plot for the PCI_consensus model. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008**, *48*, 766–784.

(2) Dilley, J. V.; Tyson, C. A.; Spanggord, R. J.; Sasmore, D. P.; Nexwell, B. W.; Dacre, J. C. Short-term oral toxicity of a 2,4,6-trinitrotoluene and hexahydro-1,3,5-trinitro-1,3,4-triazine mixture in mice, rats, and dogs. *J. Toxicol. Environ. Health.* **1982**, *9*, 587–610.

(3) Donlon, B. A.; Razo-Flores, E.; Field, J. A.; Lettinga, G. Toxicity of N-substituted aromatics to acetoclastic methanogenic activity in granular sludge. *Appl. Environ. Microbiol.* **1995**, *61*, 3889–3893.

(4) Kuz'min, V. E.; Muratov, E. N.; Artemenko, A. G.; Gorb, L. G.; Qasim, M.; Leszczynski, J. The effect of nitroaromatics' composition on their toxicity in vivo: Novel, efficient non-additive 1D QSAR analysis. *Chemosphere* **2008**, *72*, 1373–1380.

(5) Verma, R. P.; Hansch, C. Chemical Toxicity on HeLa Cells. *Curr Med Chem* **2006**, *13* (4), 423–48.

(6) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Tropsha, A.; Zhu, H.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR Models to predict environmental toxicity against *Tetrahymena pyriformis*: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.

(7) Golbraikh, A.; Tropsha, A. Beware of Q2. *J. Mol. Graph. Model.* **2002**, *20*, 269–276.

(8) Kubinyi, H. Quantitative structure-activity relationships (QSAR) and molecular modeling in cancer research. *J. Cancer Res. Clin. Oncol.* **1990**, *116*, 529–537.

(9) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem.* **1998**, *41*, 2553–2564.

(10) Isayev, O.; Rasulev, B.; Gorb, L.; Leszczynski, J. Structure−toxicity relationships of nitroaromatic compounds. *Molecular Diversity* **2006**, *10*, 233–245.

(11) Kulkarni, S. A.; Raje, D. V.; Chakrabarti, T. Quantitative structure−activity relationships based on functional and structural characteristics of organic compounds. *SAR & QSAR in Env. Res.* **2001**, *12*, 565–591.

(12) Wei, D. B.; Wu, C. D.; Wang, L. S.; Hu, H.-Y. QSPR-based prediction of absorption of halogenated aromatics on yellow-brown soil. *SAR & QSAR in Env. Res.* **2003**, *14*, 191–198.

(13) van Drie, J. H. Pharmacophore discovery: Lessons learned. *Curr. Pharm. Des.* **2003**, *9*, 1649–1664.

(14) Breiman, L. Random forests. *Machine Learning* **2001**, *45* (1), 5–32.

(15) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **2003**, *43* (6), 1947–1958.

(16) Debeljak, Z.; Skrbo, A.; Jasprica, I.; Mornar, A.; Plecnko, V.; Banjanac, M.; Medic-Saric, M. QSAR study of antimicrobial activity of some 3-nitrocoumarins and related compounds. *J. Chem. Inf. Model.* **2007**, *47* (3), 918–926.

(17) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and log P. *J. Chem. Inf. Model.* **2008**, *48* (1), 220–232.

(18) Lombardo, F.; Obach, R. S.; DiCapua, F. M.; Bakken, G. A.; Lu, J.; Potter, D. M.; Gao, F.; Miller, M. D.; Zhang, Y. A hybrid mixture discriminant analysis; random forest computational model for the prediction of volume of distribution of drugs in human. *J. Med. Chem.* **2006**, *49* (7), 2262–2267.

(19) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random forest models to predict aqueous solubility. *J. Chem. Inf. Model.* **2007**, *47* (1), 150–158.

(20) Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J. Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *J. Med. Chem.* **2007**, *50* (14), 3173–3184.

(21) Zhang, Q.-Y.; Aires-de-Sousa, J. Random forest prediction of mutagenicity from empirical physicochemical descriptors. *J. Chem. Inf. Model.* **2007**, *47* (1), 1–8.

(22) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. Hierarchical QSAR technology on the base of simplex representation of molecular structure. *J. Comp. Aid. Mol. Des.* **2008**, *22*, 403–421.

(23) Artemenko, A. G.; Muratov, E. N.; Kuz'min, V. E.; Kovdienko, N. A.; Hromov, A. I.; Makarov, V. A.; Riabova, O. B.; Wutzler, P.; Schmidtke, M. N′,N-(bis-5-Identification of individual structural fragments of nitropyrimidyl)dispirotripiperazine derivatives for cytotoxicity and antiherpetic activity allows the prediction of new highly active compounds. *J. Antimicrob. Chemother.* **2007**, *60* (1), 68–77.

(24) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N.; Volineckaya, I. L.; Makarov, V. A.; Riabova, O. B.; Wutzler, P.; Schmidtke, M. Quantitative structure−activity relationship studies of [(biphenyloxy)propyl]isoxazole derivatives: Human rhinovirus 2 replication inhibitors. *J. Med. Chem.* **2007**, *50*, 4205–4213.

(25) Kuz'min, V. E.; Muratov, E. N.; Artemenko, A. G.; Gorb, L. G.; Qasim, M.; Leszczynski, J. The effects of characteristics of substituents on toxicity of the nitroaromatics: HiT QSAR study. *J. Comp. Aid. Mol. Des.* **2008**, *22*, 747–759.

(26) Muratov, E. N.; Artemenko, A. G.; Kuz'min, V. E.; Lozitsky, V. P.; Fedchuk, A. S.; Lozitska, R. N.; Boschenko, Y. A.; Gridina, T. L. Investigation of anti-influenza activity using hierarchic QSAR technology on the base of simplex representation of molecular structure. *Antivir. Res.* **2005**, *65* (3), A62–A63.

(27) Jolly, W. L.; Perry, W. B. Estimation of atomic charges by an electronegativity equalization procedure calibration with core binding energies. *J. Am. Chem. Soc.* **1973**, *95*, 5442–5450.

(28) Wang, R.; Fu, Y.; Lai, L. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615–621.

(29) Ioffe, B. V. *Chemistry Refractometric Methods*, 3 ed.; Himiya: Leningrad, 1983; p 350.

(30) Wang, R.; Fu, Y.; Lai, L. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comp. Sci.* **1997**, *37*, 615–621.

(31) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*. Wadsworth: Belmont, 1984; p 368.

(32) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N.; Polischuk, P. G.; Ognichenko, L. N.; Liahovsky, A. V.; Hromov, A. I.; Varlamova, E. V. Virtual Screening and Molecular Design Based on Hierarchical QSAR Technology. In *Recent Advances in QSAR Studies*; Puzyn, T.; Cronin, M.; Leszczynski, J., Eds.; Springer: New York, 2009; DOI 10.1007/978-1-4020−9783−6_5.

(33) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *Altern. Lab. Anim.* **2005**, *33*, 445–459.
(34) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure−property relationship approach based on *k*-nearest-neighbor principle. *J. Chem. Inf. Model.* **2000**, *40*, 185–194.
(35) Lindgren, F.; Geladi, P.; Rannar, S.; Wold, S. Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms. *J. Chemom.* **1994**, *8*, 349–363.
(36) Rannar, S.; Lindgren, F.; Geladi, P.; Wold, S. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *J. Chemom.* **1994**, *8*, 111–125.
(37) Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel anatgonists. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306–310.
(38) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: Definition and application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(39) Kuz'min, V. E.; Artemenko, A. G.; Kovdienko, N. A.; Tetko, I. V.; Livingstone, D. J. Lattice model for QSAR studies. *J Mol Model* **2000**, *6*, 517–526.
(40) Vitiuk, N. V.; Kuz'min, V. E. Mechanistic models in chemometrics for the analysis of multidimensional data of researches. Analogue of dipole-moments method in the structure(composition)−property relationships analysis. *Z. Anal. Khimii* **1994**, *49*, 165–167.
(41) Lindgren, F.; Geladi, P.; Rännar, S.; Wold, S. Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms. *J. Chemom.* **1996**, *8* (5), 349–363.
(42) Kubinyi, H. Evolutionary variable selection in regression and PLS analyses. *J. Chemom.* **1996**, *10*, 119–133.
(43) Tropsha, A.; Gramatica, P.; Gombar, V. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.
(44) Schultz, T. W.; Netzeva, T. I., Development and Evaluation of QSARs for Ecotoxic Endpoints: The Benzene Response−Surface Model for *Tetrahymena* Toxicity. In *Modeling Environmental Fate and Toxicity*, Cronin, M. T. D.; Livingstone, D. J., Eds.; CRC Press: Boca Raton, 2004; pp 265−284.

CI900203N