

ARTICLES

Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors

Qing-You Zhang and João Aires-de-Sousa*

CQFB and REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Received November 28, 2005

Fast-to-calculate empirical physicochemical descriptors were investigated for their ability to predict mutagenicity (positive or negative Ames test) from the molecular structure. Fast methods are highly desired for the screening of large libraries of compounds. Global molecular descriptors and MOLMAP descriptors of bond properties were used to train random forests. Error percentages as low as 15% and 16% were achieved for an external test set with 472 compounds and for the training set with 4083 structures, respectively. High sensitivity and specificity were observed. Random forests were able to associate meaningful probabilities to the predictions and to explain the predictions in terms of similarities between query structures and compounds in the training set.

INTRODUCTION

The relevance of reliable methodologies for the structure-based prediction of chemical carcinogenicity and mutagenicity cannot be overemphasized. Regulatory bodies continue to use rodent carcinogenicity and bacterial mutagenicity to predict human health risks, even though there is little evidence that these model organisms have much predictive power. The standard bioassay in rodents that is used to assess the carcinogenic potential of chemicals is extremely long and costly and requires the sacrifice of large numbers of animals. Alternative short-term tests have been proposed that can provide some indications.¹ Among these, the mutagenicity in *Salmonella typhimurium* as determined by the Ames test has become the standard test for mutagenicity determinations. Although simple, even such experimental tests cannot be performed on the large virtual libraries of molecules currently screened for drug discovery including compounds yet to be synthesized. At the same time, early detection of possible drug safety problems is highly desired. In a different context, assessment of the toxicological properties of chemicals imported or produced within sophisticated economic spaces is increasingly mandatory, which demands prioritizing lists of compounds for experimental testing. Fast computer programs that can screen large sets of compounds and assign a probability of mutagenicity are thus of the utmost importance in all those situations. Additionally, mutagenicity tests are usually associated with high levels of uncertainty, which makes theoretical predictions a potentially useful way to identify suspect experimental results.

A large number of programs and methods were developed for predicting mutagenicity, and the state of the art was reviewed by Patlewicz et al. in 2003,² and by Benigni in 2005.³ Molecular structures have been analyzed or repre-

sented in terms of explicit structural fragments,^{4,5} physicochemical descriptors calculated by ab initio⁶ or semiempirical procedures,⁷ 2D molecular descriptors based on molecular graphs,^{8–10} or 3D descriptors.¹⁰ A number of computer approaches have been tried for establishing relationships between structure representations and mutagenicity including linear models,^{7,10} *k*-nearest neighbors,^{7,8} artificial neural networks,^{7–9} support vector machines,¹¹ inductive logic programming,¹² decision trees,¹³ or decision forests.⁸

Recent developments have focused on the exploration of larger data sets of non-congeneric molecules (ca. 2000–4000 compounds),^{4,8} newer machine learning algorithms,^{5,11} and the association of mutagenicity models with data on metabolic reactions.^{14,15} Kazius et al.⁴ assembled a data set of 4337 compounds and derived toxicophores for mutagenicity obtaining an error rate of 18% for the training set and 15% for an external 535-compounds test set. Votano et al.⁸ reported the application of three quantitative structure–activity relationship (QSAR) methods, artificial neural networks, *k*-nearest neighbors, and decision forest, to a data set of 3363 diverse compounds represented by molecular connectivity indices, electrotopological state indices, and binary indicators. An average concordance of 82% was obtained for an independent test set of 400 compounds, and a superior performance to that of the DEREK and MULTI-CASE systems was claimed for therapeutic drugs. Mekenyan et al.¹⁴ proposed different procedures for compounds that were positive without metabolic activation, and for compounds exhibiting mutagenicity only in the presence of metabolic activation. Common reactivity patterns were used to identify structural features responsible for mutagenicity. Compounds classified as nonmutagenic without activation were submitted to a metabolism simulator, and the potential metabolites were classified in terms of mutagenicity.

QSAR methods have been suggested to have a potential to perform better with new compounds, particularly when

*Corresponding author tel.: (+351) 21 2948300; fax: (+351) 21 2948550; e-mail: jas@fct.unl.pt.

based on molecular descriptors that do not explicitly encode structural fragments. Electronic parameters calculated at the semiempirical level, usually combined with other molecular descriptors, have been shown to yield simple useful models of mutagenicity.³ Particularly related to mutagenicity are the electronic properties of chemical bonds, because chemical reactivity toward DNA is the main mechanism of mutagenicity.

Chemical reactivity, being related to the ability to make and break bonds, is primarily determined by properties of the bonds available in a molecule. Properties of bonds, calculated by *ab initio* methods, have been used for structure–mutagenicity relationships within congeneric series of triazines and hydroxyfuranones.⁶ Bond properties could be used even for compounds that are mutagenic only after metabolic transformation, as the properties of the original compound implicitly encode its ability for metabolic activation.⁶ Gasteiger et al.¹⁶ proposed that seven empirical physicochemical properties are particularly relevant for representing bonds and for modeling chemical reactivity: σ electronegativity, π atomic charge, total atomic charge, bond polarity, mean bond polarizability, resonance stabilization, and bond dissociation energy. Such empirical physicochemical properties can nowadays be calculated extremely fast for large data sets of compounds. In order to use all that information for an entire molecule, and at the same time have a fixed-length representation, we have proposed to map all the bonds of a molecule into a fixed-length 2D self-organizing map—a *MOLMAP* (*MOL*ecular *MAP* of Atom-level *Properties*).^{17,18}

A self-organizing map (SOM)¹⁹ must be trained beforehand with a diversity of bonds from different structures (each bond described by the seven bond properties). Then, all the bonds of one molecule are submitted to the trained SOM; each bond activates one neuron, and the pattern of activated neurons is a map of reactivity features of that molecule (*MOLMAP*)—a fingerprint of the bonds available in that structure.

In this paper, we used the PETRA software package²⁰ to calculate empirical physicochemical properties for molecules and for bonds, and *MOLMAP* descriptors were generated on the basis of the bond properties. Random forests (RFs)^{21,22} were trained with the Bursi Mutagenicity data set⁴ to predict mutagenicity from *MOLMAP* descriptors and molecular descriptors. The prediction of mutagenicity was here considered a binary classification problem (positive or negative Ames test result). RFs have several advantageous features concerning current requirements for QSAR models, as they can model linear as well as nonlinear relationships, can account for different mechanisms, assign a probability to each prediction, yield a similarity measure between a query structure and structures in the training set, assess the relative importance of the descriptors in the model, include an intrinsic internal cross-validation test, and perform well with large sets of descriptors and large sets of training examples. After the training, the RFs were evaluated both by out-of-bag (OOB) estimation (internal cross-validation of the training set) and with an external test set of 472 compounds, previously used by Kazius et al.⁴

METHODS

Data Set. A data set of 4083 organic structures and the corresponding Ames test results (2308 mutagens and 1775

nonmutagens) from the Bursi Mutagenicity data set⁴ were used for establishing structure–mutagenicity relationships. According to the source paper,⁴ this data set was constructed from a diversity of data sources, namely, the Chemical Carcinogenicity Research Information System database, TOXNET, National Toxicology Program database, Carcinogenic Potency database, and EPA/IARC Genetic Activity Profile database. Inorganic compounds, organometallic compounds, and additional occurrences of enantiomers and diastereoisomers had been removed from this data set, and the structures were not desalted. Data were restricted to standard Ames test results of *Salmonella typhimurium* strains required for the regulatory evaluation of drug approval, either with or without metabolic activation. A compound was listed as mutagenic if at least one Ames test result was positive and as nonmutagenic if no Ames test result was positive. The data set is currently available from <http://www.chem-informatics.org>. Mutagenicity data in an independent set of 472 structures collected by Young et al.¹³ was used for external validation, which is composed of 305 mutagens and 167 nonmutagens. These are essentially the same sets used by Kazius et al.⁴ with some structures excluded from the original data sets when they were not accepted by PETRA 3.2.

Molecular Descriptors. Constitutional and empirical physicochemical descriptors were calculated by PETRA 3.2²⁰ for molecules and for bonds. The global molecular descriptors were used as such. The bond descriptors for all the bonds in a molecule were incorporated into a *MOLMAP* descriptor for that molecule.

Seventeen global molecular descriptors were calculated: number of atoms, number of bonds, molecular weight, number of aromatic atoms, polarizability, number of NH groups, number of NH₂ groups, number of heavy atoms, number of hydroxyl groups, number of oxygen atoms, number of nitrogen atoms, minimum partial atomic charge, maximum partial atomic charge, minimum partial atomic charge on hydrogen, maximum partial atomic charge on hydrogen, aromatic delocalization energy, and ring strain energy.

MOLMAP descriptors were generated with Kohonen SOMs.¹⁹ SOMs reduced to 2D the dimension of chemical bonds, represented by seven bond properties: resonance stabilization, difference between the σ electronegativity of the two bonded atoms, difference between the total charge of the two bonded atoms, difference between the π charge of the two bonded atoms, mean bond polarizability, bond dissociation energy, and bond polarity.^{23,24} As some properties depend on the orientation of the bond, each bond was represented twice (as A–B and B–A). In order to focus on regions around functional groups, only bonds were considered that include a heteroatom or an atom belonging to a π system.

SOMs learn by unsupervised training, revealing similarities between objects. A Kohonen SOM consists of a grid of so-called neurons, each containing as many elements (weights) as the number of input variables. Here, the objects are bonds, and the input variables are the seven properties of bonds (Figure 1). Before the training starts, the weights take random values. During the training, each individual bond is mapped into the neuron that contains the most similar weights compared to its properties. This is the central neuron, or

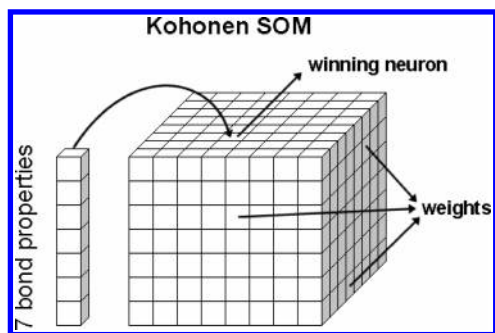


Figure 1. Representation of a Kohonen SOM for processing chemical bonds. Every small box of the block represents a weight. The Kohonen SOM is trained by the iterative presentation of objects (bonds described by seven properties).

winning neuron. It is said that the winning neuron is *excited* by the bond, and its weights are then adjusted to make them even more similar to the properties of the presented bond. Not only does the winning neuron have its weights adjusted but also the neurons in its neighborhood. The extent of adjustment depends, however, on the topological distance to the winning neuron—the closer a neuron is to the winning neuron, the larger is the adjustment of its weights. The objects of the training set are iteratively fed to the map, the weights are corrected, and the training is stopped when the predefined number of cycles is attained. A trained Kohonen SOM reveals similarities in the objects of a data set in the sense that similar objects (similar bonds) are mapped into the same or closely adjacent neurons.

SOMs of sizes $15 \times 15 = 225$, $20 \times 20 = 400$, and $25 \times 25 = 625$ were trained with a random subset of 4999 bonds extracted from the 4083 structures of the training set. SOMs were implemented with in-house-developed software based on JATOON Java applets.^{25,26}

The bonds existing in a molecule can be represented as a whole by mapping all the bonds of the same molecule onto the SOM previously trained with a diversity of bonds. The pattern of activated neurons can be interpreted as a fingerprint of the available bonds in the molecule, and it was used as a descriptor (MOLMAP). For numerical processing, each neuron got a value equal to the number of times it was activated by bonds of the molecule. The map was then transformed into a vector by concatenation of the columns. In order to account for the relationship between the similarity of bonds and proximity in the map, a value of 0.3 was added to each neuron multiplied by the number of times a neighbor was activated by a bond. A detailed description of MOLMAP generation can be found in refs 17 and 18. For structures composed of more than one fragment, for example, salts, the MOLMAPs of all the fragments were summed.

MOLMAP descriptors (representing bond properties) and the 17 global molecular descriptors were explored separately and in combination to predict mutagenicity.

Random Forests.^{21,22} A RF is an ensemble of unpruned classification trees created by using bootstrap samples of the training data and random subsets of variables to define the best split at each node. It is a high-dimensional nonparametric method that works well on large numbers of variables. Prediction is made by majority vote of the individual trees. It has been shown that the method is extremely accurate in a variety of applications.²² Additionally, performance is internally assessed with the prediction error for the objects

left out in the bootstrap procedure (internal cross-validation or out-of-bag estimation). The method quantifies the importance of a variable by the increase in misclassification occurring when the values of the variable are randomly permuted, or by the decrease in a node's impurity every time the variable is used for splitting. RFs also assign a probability to every prediction on the basis of the number of votes obtained by the predicted class. A measure of similarity between two objects can be calculated from the number of trees that classify the two objects in the same terminal node. Therefore, such a comparison relies on the descriptors that were chosen by the forest to build the model. In this study, RFs were grown with the R program, version 2.0.1,²⁷ using the randomForest library.²⁸ RFs were trained to classify molecules as mutagenic or nonmutagenic (dependent variable) on the basis of their MOLMAP and/or global molecular descriptors (independent variables). The number of trees in a RF was set to 1000.

Classification Tree.²⁹ A single classification tree was investigated to predict mutagenicity. This was grown with the CART algorithm,²⁹ differently from the trees in RFs. A classification tree is sequentially constructed, partitioning objects from a parent node into two child nodes. Each node is produced by a logical rule, usually defined for a single variable, where objects below a certain variable's value fall into one of the two child nodes, and objects above fall into the other child node. The prediction for an object reaching a given terminal node is obtained by a majority vote of the objects (in the training set) reaching the same terminal node. The entire procedure comprises three main steps. First, an entire tree is constructed by data splitting into smaller nodes; each produced split is evaluated by an impurity function which decreases as long as the new split permits the child node's content to be more homogeneous than the parent node. Second, a set of smaller, nested trees is obtained by the obliteration (pruning) of certain nodes of the tree obtained in the first step. The selection of the weakest branches is based on a cost-complexity measure that decides which subtree, from a set of subtrees with the same number of terminal nodes, has the lowest (within node) error. Finally, from the set of all nested subtrees, the tree giving the lowest value of error in cross validation (where the set of objects used to grow the tree is different from the prediction set) is selected as the optimal tree. In this study, a classification tree was grown with the R program, version 2.0.1,²⁷ using the RPART library with the default parameters.

RESULTS AND DISCUSSION

RFs were trained to predict mutagenicity from (a) MOLMAP descriptors of different sizes, (b) global molecular descriptors, and (c) a combination of MOLMAPs and global molecular descriptors. The results for internal cross-validation (out-of-bag estimation) are presented in Table 1. The OOB estimation is a reliable indication of the robustness of the model—every tree of the forest is grown with a (random) subset of the training set, and predictions are obtained for the objects left out. All these predictions for all the trees are combined to calculate the OOB estimation.

Very good predictions were observed, considering the average intrinsic error of the assembled data set (11%) and the experimental error of Ames tests in general (15%).⁴ The

Table 1. Correct Predictions (Internal Cross-Validation) of Random Forests Trained with the Whole Training Set Consisting of 4083 Compounds Represented by MOLMAP Descriptors of Varying Size and 17 Global Molecular Descriptors

descriptors	correct out-of-bag estimation
17 molecular descriptors	81.0%
MOLMAPs (15 × 15)	82.4%
MOLMAPs (20 × 20)	82.3%
MOLMAPs (25 × 25)	83.3%
MOLMAPs (15 × 15) + 17 molecular descriptors	83.7%
MOLMAPs (20 × 20) + 17 molecular descriptors	83.7%
MOLMAPs (25 × 25) + 17 molecular descriptors	84.1%

Table 2. Random Forest Predictions of Mutagenicity from MOLMAP Descriptors of Size of 25 × 25 and 17 Global Molecular Descriptors, for Specific Classes of Compounds in the Training Set (OOB Estimations)

class	% correct predictions	sensitivity ^a	specificity ^b
three-member heterocycles	88	0.89	0.79
cyclopropanes	61	0.44	0.80
chlorides	80	0.81	0.78
N=O (no formal charge on N)	92	0.96	0.17
nitro	90	0.98	0.38
N=N	78	0.89	0.26

^a Ratio of true positives to the sum of true positives and false negatives. ^b Ratio of true negatives to the sum of true negatives and false positives.

predictions slightly improved with increasing resolution of the MOLMAPs. Interestingly, the 17 global molecular descriptors alone obtained comparable results to MOLMAPs. The best results could be achieved by a combination of MOLMAPs and molecular descriptors. The random forests exhibited high robustness against initialization of the random procedures. When 10 different seeds were used, the OOB error only varied between 15.5% and 16.2%. The results compare well with the predictions obtained by Kazius et al. for approximately the same data set by applying toxicophores (18% error percentage).⁴

An inspection of the assigned relative importance of descriptors for the model based on MOLMAPs (25 × 25) + 17 global molecular descriptors revealed that nine out of the top 10 descriptors are global molecular descriptors. Of particular relevance are the maximum and minimum partial atomic charges of hydrogen atoms, ring strain energy, and descriptors related to size (number of bonds, number of atoms, and molecular weight). The ring strain energy accounts for the known mutagenicity of small ring epoxides, thioepoxides, and aziridines in general. Votano et al.⁸ reported a statistical association between mutagenicity and the maximum hydrogen atom level E-state value in a molecule, which can be related to partial charge.

Table 3. Random Forest Prediction of Mutagenicity from MOLMAP Descriptors of Size of 25 × 25 and 17 Global Molecular Descriptors for a Test Set of 1000 Compounds

descriptors	% correct predictions (OOB training set)	correct predictions (test set)	sensitivity ^a (test set)	specificity ^b (test set)
MOLMAP descriptors	83	815 (82%)	0.85	0.77
17 molecular descriptors	80	803 (80%)	0.85	0.74
MOLMAP + 17 molecular descriptors	83	825 (83%)	0.86	0.77

^a Ratio of true positives to the sum of true positives and false negatives. ^b Ratio of true negatives to the sum of true negatives and false positives.

The most important MOLMAP descriptor (seventh in the global ranking) corresponds to a neuron of the SOM activated by bonds belonging to 81 structures (66 of which are mutagens)—56 C—Cl bonds, 76 N—O bonds of nitro groups, and eight C—H bonds of alkyne groups. The C—Cl bonds are from 35 aliphatic chlorides (91% of which are mutagens), and the N—O bonds are from 38 structures (84% of which are mutagens). Two out of the other eight compounds are mutagens. Nitro compounds are statistically associated with mutagenicity in the training set⁴—86% of the nitro compounds in the data set are mutagenic. Aliphatic halides were also proposed as toxicophores.⁴ This shows the ability of MOLMAPs to reveal interpretable correlations between structural features and mutagenicity without an explicit representation of molecular fragments.

The real importance of a variable *to the problem* may be hidden in RFs by high intercorrelations with other descriptors. The importance of a descriptor, as assessed by the RFs, must be seen as its importance *to the model*. In this particular study, the fact that MOLMAP descriptors are highly intercorrelated (because of the duplicate representation of each bond and the influence of a MOLMAP component on their neighbors) could be a reason for the global descriptors getting the top scores concerning importance. To investigate this point, we excluded intercorrelated descriptors above a Pearson correlation coefficient of 0.8 and trained a new RF with the remaining 494 variables. The accuracy of the prediction was unchanged, and no dramatic change in the ranking of descriptors was observed. Although the most important MOLMAP descriptors became slightly more important (fifth, eighth, and ninth in the new ranking), the two most important variables remained the same, while the fourth became third.

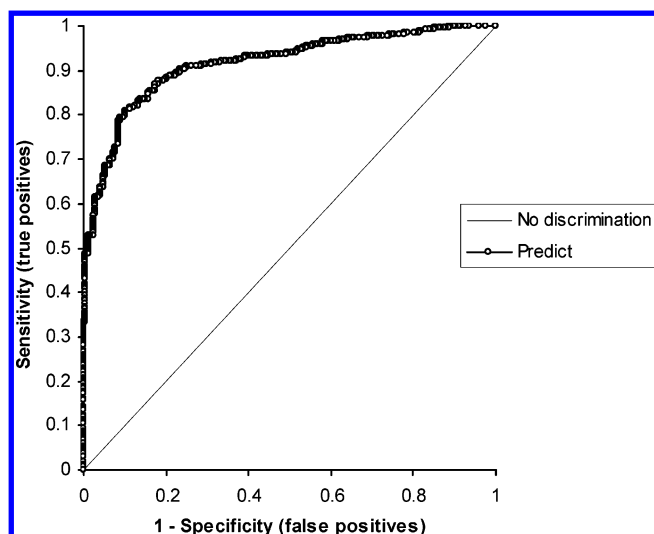
The ability of the RF to predict specific families of compounds, some usually associated with mutagenicity, was analyzed on the basis of the OOB estimations, Table 2. General good global concordance was observed and also high specificity values, except for classes with a reduced fraction of nonmutagens in the data set—nitro, compounds with N=N bonds, and compounds with N=O bonds (no formal charge on N atom). Sensitivity was relatively low for cyclopropanes, which shows that large negative ring strain energy does not necessarily lead to a prediction of mutagenicity.

The robustness of the predictions was confirmed by training RFs with only 3083 random compounds of the training set and leaving 1000 out as a test set. The results are displayed in Table 3.

The ability of such a system to make predictions for new functional groups not used for training was assessed with different partitions between training and test sets. A random forest was trained with all the compounds except those

Table 4. Random Forest Prediction of Mutagenicity for an External Data Set of 472 Compounds.

descriptors	correct predictions	sensitivity	specificity
MOLMAPs 25×25	396 (84%)	0.84	0.84
17 molecular descriptors	374 (79%)	0.76	0.86
MOLMAPs 25 × 25 + 17 molecular descriptors	399 (85%)	0.84	0.86
MOLMAPs 25 × 25 + 17 molecular descriptors (only predictions with probability ≥ 0.7)	91 (296 out of 324 structures)	0.91	0.93

**Figure 2.** ROC for the 472 test compounds predicted by the RF model.

containing a specific functional group. Predictions were then obtained for the objects left out. Percentages of concordance not far from 50% were obtained for compounds with N=O bonds (no formal charge on N). But better results were obtained for compounds with N=N bonds (73% concordance using only molecular descriptors), and for aziridines and thioepoxides together (94% using only molecular descriptors and including epoxides in the training set). Good results were

also obtained for nitro compounds (82% concordance using molecular descriptors and MOLMAPs). An inspection of the reported similar structures in the training set for some of the nitro compounds revealed that in many cases these were correctly classified as mutagenic because of similar nitroso mutagenic compounds (nitroso and nitro bonds activate neighbor neurons) or azoxi mutagenic compounds in the training set. Many nitro compounds in the data set are aromatic, and several were perceived as mutagenic because of similar polyaromatic structures in the training set.

Finally, the RFs trained with the whole data set of 4083 structures were tested with an external set consisting of 472 compounds, Table 4. An advantage of using MOLMAP descriptors is apparent by comparison with the results of molecular descriptors alone. The good predictions are consistent with those obtained by OOB estimation within the training set. They also compare well with the results achieved by Kazius et al.⁴ for approximately the same data set (85% correct predictions) using toxicophores. The external data set includes eight azo compounds listed as mutagenic, 65 compounds with a nitro group (62 listed as mutagenic), and 63 compounds listed as mutagenic bearing a N=O bond (no formal charge on N). All the azo compounds were correctly predicted, as well as 95% of the nitro compounds and 98% of the compounds with a N=O bond (no formal charge on N). The last line of Table 4 illustrates the usefulness of the probabilities associated by RFs to the predictions—a significant improvement of concordance was observed for predictions with high assigned probabilities. Figures 2 and 3 give a more detailed picture on this matter. A receiver operator characteristic (ROC) curve was obtained for the 472 predictions using different thresholds of probability for the prediction of mutagenicity, Figure 2. A model with no predictive ability would yield the diagonal line. The closer the area under the ROC curve is to 1, the greater is the predictive ability of the model.⁸ For the ROC curve in Figure 2, an area of 0.914 was obtained.

Figure 3 graphically represents the probability associated with each of the mutagenic and nonmutagenic compounds. False negatives appeared more as a problem than false positives. All the structures with an assigned probability of

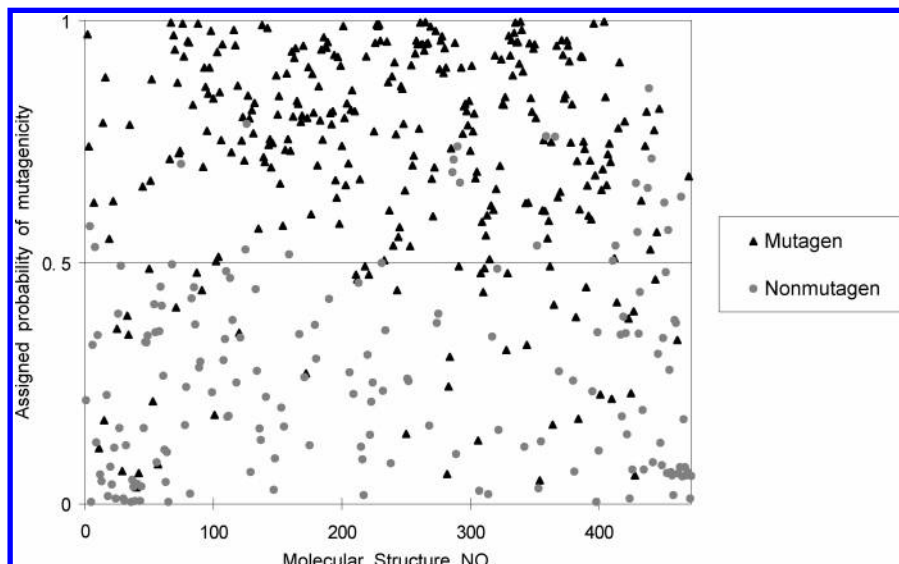
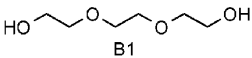
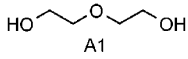
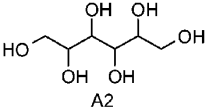
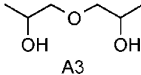
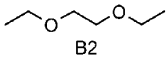
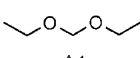


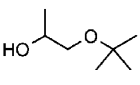
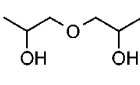


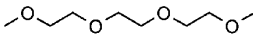
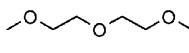
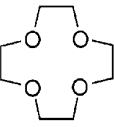
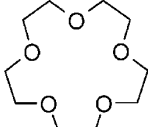
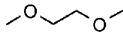
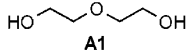

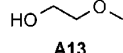
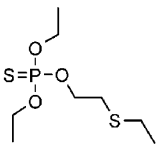
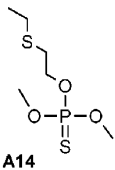
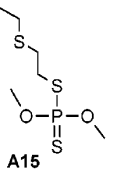
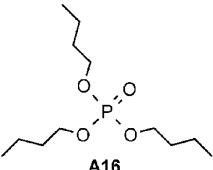
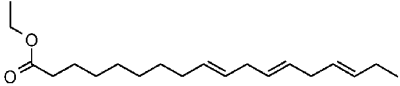
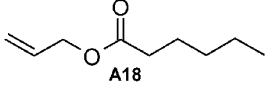
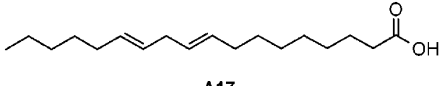
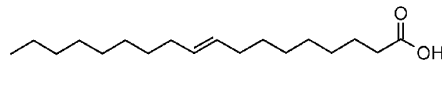
**Figure 3.** Probabilities associated by the random forest to the 472 predictions of the external validation set.

Table 5. “False Negative” Predictions with Associated High Probability (in the External Test Set) on the Basis of Similar Structures in the Training Set.

Compounds predicted nonmutagenic with probability > 0.9 and listed mutagenic	Most similar compounds in the training set (all listed nonmutagenic)
 B1	 A1  A2  A3
 B2	 A4  A5  A6
 B3	 A3  A7  A8
 B4	 A9  A10  A11
 B5	 A1  A12  A13
 B6	 A14  A15  A16
 B7	 A18  A17  A19

mutagenicity higher than 0.86 were correctly predicted. On the other hand, there are seven false negatives with an assigned probability of nonmutagenicity higher than 0.90 (i.e., a probability of mutagenicity lower than 0.10). To find a reason for such wrong predictions, the same RF was queried for the most similar compounds in the training set. Similarity between two compounds is assessed by the number of times they are classified in the same terminal nodes of the RF individual trees. Table 5 represents the seven compounds, **B1–B7**, together with the three most similar to

each of them in the training set. All the retrieved compounds bear striking similarities with the queried compounds and were listed in the training set as nonmutagenic, thus explaining the predictions. Significantly, five out of the seven compounds (**B1–B5**) are glycol ethers or related structures. This family of compounds has been extensively used as solvents in the production of vast numbers of products, and their toxicological properties, namely, their genotoxicity and carcinogenicity, have raised much debate and controversy. For example, *Toxicology Letters* dedicated an entire issue

to the theme in March 2005 and published the Proceedings of the Third International Scientific Symposium on the Health Effects of Glycol Ethers.³⁰ Although glycol ethers are considered as generally nongenotoxic,³¹ which is in agreement with the seven mentioned predictions and the Bursi data set, some of them were reported genotoxic in some tests, and discordant results have appeared for at least one specific structure.³² Compound **B7** is an ester of an unsaturated fatty acid. Although esters of unsaturated fatty acids are nonmutagenic in general, it has been observed that products of their oxidation (autooxidation or photo-sensitized oxidation) exhibit mutagenicity.^{33,34} Such a mechanism could possibly explain how different experimental tests, involving different opportunities for oxidation, could yield different mutagenicity results. All these observations suggest inconsistency in the data rather than a problem with the method itself, or very subtle structural effects that can hardly be learned by a general approach. Inconsistency in the data would not be surprising because of the large size of the set, diversity of sources, inclusion of data obtained with and without metabolic activation, and the experimental error associated with the Ames test. Difficulties can also arise from the binary classification essence of the approach—similar compounds near the threshold of mutagenicity can be differently classified. The similarity analysis illustrates the ability of RFs to give some explanation of the predictions in terms of available data in the training set.

An improvement in computation efficiency was achieved by retraining a RF using only the 100 most important descriptors (as determined by the RF) and 200 trees. Essentially the same results were obtained, with correlation coefficient $r^2 = 0.953$ between the probabilities assigned by the two RFs to the compounds in the external test set. The prediction of mutagenicity for the set of 472 structures took 23 s on an Intel Xeon 3.2 GHz PC with 1 GB of RAM running the Linux kernel 2.6.5 operating system, starting from the connection tables (.sdf file).

In order to compare with RFs, an experiment was performed with a single classification tree grown with the CART algorithm²⁹ on the basis of molecular descriptors and MOLMAP descriptors of size 25×25 . The tree was able to correctly predict 78% of the training set and 83% of the external data set (sensitivity = 0.83, specificity = 0.81). A graphical representation of the tree is available as Supporting Information. Ten descriptors were chosen by the tree. Among these are two of the molecular descriptors identified as the most important by the RFs—ring strain energy and molecular weight. Three MOLMAP descriptors are at the top of the tree. The first (A419) and the third (A249) correspond mostly to C—H bonds of polyaromatic compounds, while the second (A598) corresponds to C—Cl bonds and to C—N bonds in azides. All of these structural features are indeed related to toxicophores.⁴ The first node alone was able to classify as mutagenic 969 compounds out of the 4083 in the training set with a 13% error. In spite of a very good performance, only slightly inferior to RFs in terms of correct predictions, a single tree was not able to provide meaningful probabilities for the predictions of the test set, nor was it able to assess similarities between compounds.

CONCLUSIONS

The results demonstrate that fast-to-calculate physico-chemical descriptors are able to build models for the structure-based prediction of mutagenicity with high accuracy, sensitivity, and specificity as determined for large sets of non-congeneric compounds. MOLMAP descriptors are able to represent the properties of bonds in a molecular structure by a fixed-length code and to compare molecules with no previous alignment or bond-to-bond mapping. They were shown to encode relevant information for mutagenicity prediction and could reveal structural features linked to mutagenicity without explicitly encoding structural fragments. The use of physicochemical descriptors gave the model some ability to make predictions for functional groups not used for training.

Random forests exhibited interesting features for this application, not only in terms of prediction accuracy but also by providing meaningful probabilities and some explanation for the predictions. The predictions could be explained in terms of a similarity measure between query compounds and the compounds in the training set. RFs allow the similarity measure to be based on features identified as relevant for the mutagenicity.

High specificity was observed for classes of compounds generally associated with mutagenicity, except for those with a very low number of negatives in the training set.

ACKNOWLEDGMENT

The authors thank Dr. Roberta Bursi, Dr. Jeroen Kazius, Dr. Stanley Young, and Dr. Christophe Lambert for sharing data sets of mutagenicity. Molecular Networks GmbH (Erlangen, Germany) is acknowledged for access to the PETRA software package. Q.Y.Z. acknowledges Fundação para a Ciência e a Tecnologia (Lisbon, Portugal) for a postdoctoral grant under the POCTI program (SFRH/BPD/14476/2003).

Supporting Information Available: Graphical representation of the classification tree obtained for mutagenicity prediction. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Bajpayee, M.; Pandey, A. K.; Parmar, D.; Dhawan, A. Current Status of Short-Term Tests for Evaluation of Genotoxicity, Mutagenicity, and Carcinogenicity of Environmental Chemicals and NCEs. *Toxicol. Mech. Methods* **2005**, *15*, 155–180.
- (2) Patlewicz, G.; Rodford, R.; Walker, J. D. Quantitative Structure–Activity Relationships for Predicting Mutagenicity and Carcinogenicity. *Environ. Toxicol. Chem.* **2003**, *22*, 1885–1893.
- (3) Benigni, R. Structure–Activity Relationship Studies of Chemical Mutagens and Carcinogens: Mechanistic Investigations and Prediction Approaches. *Chem. Rev.* **2005**, *105*, 1767–1800.
- (4) Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48*, 312–320.
- (5) Helma, C.; Cramer, T.; Kramer, S.; Raedt, L. D. Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Non-congeneric Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402–1411.
- (6) Popelier, P. L. A.; Smith, P. J.; Chaudry, U. A. Quantitative Structure–Activity Relationships of Mutagenic Activity from Quantum Topological Descriptors: Triazines and Halogenated Hydroxyfuranones (Mutagen-X) Derivatives. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 709–718.
- (7) He, L. N.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. Predicting the Genotoxicity of Polycyclic Aromatic Compounds from

- Molecular Structure with Different Classifiers. *Chem. Res. Toxicol.* **2003**, *16*, 1567–1580.
- (8) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q. A.; Tong, W. Three New Consensus QSAR Models for the Prediction of Ames Genotoxicity. *Mutagenesis* **2004**, *19*, 365–377.
- (9) Hall, L. H.; Hall, L. M. QSAR Modeling Based on Structure-Information for Properties of Interest in Human Health. *SAR QSAR Environ. Res.* **2005**, *16*, 13–41.
- (10) Gramatica, P.; Consonni, V.; Pavan, M. Prediction of Aromatic Amines Mutagenicity from Theoretical Molecular Descriptors. *SAR QSAR Environ. Res.* **2003**, *14*, 237–250.
- (11) Mahe, P.; Ueda, N.; Akutsu, T.; Perret, J. L.; Vert, J. P. Graph Kernels for Molecular Structure–Activity Relationship Analysis with Support Vector Machines. *J. Chem. Inf. Model.* **2005**, *45*, 939–951.
- (12) King, R. D.; Muggleton, S. H.; Srinivasan, A.; Sternberg, M. J. E. Structure–Activity Relationships Derived by Machine Learning: The Use of Atoms and Their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 438–442.
- (13) Young, S. S.; Gombar, V. K.; Emptage, M. R.; Cariello, N. F.; Lambert, C. Mixture Deconvolution and Analysis of Ames Mutagenicity Data. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 5–11.
- (14) Mekenyan, O.; Dimitrov, S.; Serafimova, R.; Thompson, E.; Kotov, S.; Dimitrova, N.; Walker, J. D. Identification of the Structural Requirements for Mutagenicity by Incorporating Molecular Flexibility and Metabolic Activation of Chemicals I: TA100 Model. *Chem. Res. Toxicol.* **2004**, *17*, 753–766.
- (15) Sello, G.; Sala, L.; Benfenati, E. Predicting Toxicity: A Mechanism of Action Model of Chemical Mutagenicity. *Mutat. Res.* **2001**, *479*, 141–171.
- (16) Simon, V.; Gasteiger, J.; Zupan, J. A Combined Application of Two Different Neural Network Types for the Prediction of Chemical Reactivity. *J. Am. Chem. Soc.* **1993**, *115*, 9148–9159.
- (17) Gupta, S.; Mathew, S.; Abreu, P. M.; Aires-de-Sousa, J. QSAR Analysis of Phenolic Antioxidants Using MOLMAP Descriptors of Local Properties. *Bioorg. Med. Chem.* **2006**, *14* (4), 1199–1206.
- (18) Zhang, Q. Y.; Aires-de-Sousa, J. Structure-Based Classification of Chemical Reactions without Assignment of Reaction Centers. *J. Chem. Inf. Model.* **2005**, *45* (6), 1775–1783.
- (19) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 1999.
- (20) PETRA can be tested on the Web site <http://www2.chemie.uni-erlangen.de> and is developed by Molecular Networks GmbH (Erlangen, Germany), <http://www.mol-net.de> (accessed Nov 2006).
- (21) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
- (22) 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. Comput. Sci.* **2003**, *43*, 1947–1958.
- (23) Gasteiger, J. Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: Heidelberg, Germany, 1988; pp 119–138.
- (24) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219–3228.
- (25) Aires-de-Sousa, J. JATOON: Java Tools for Neural Networks. *Chemom. Intell. Lab. Syst.* **2002**, *61*, 167–173.
- (26) The JATOON applets are available at <http://www.dq.fct.unl.pt/staff/jas/jatoon> (accessed Nov 2006).
- (27) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2004; ISBN 3-900051-07-0, URL <http://www.R-project.org> (accessed Nov 2004 and Jan 2006).
- (28) Fortran original by Leo Breiman and Adele Cutler; R port by Andy Liaw and Matthew Wiener (2004).
- (29) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman & Hall/CRC: Boca Raton, Florida, 2000.
- (30) Lewis, S. A.; Tyler, T. R.; Kelsey, J.; Waugh, K. P. Proceedings of the Third International Scientific Symposium on the Health Effects of Glycol Ethers – 2002, 17–18 October 2002, Maison de la Chimie Paris, France. *Toxicol. Lett.* **2005**, *156*, 1–2.
- (31) Multigner, L.; Catala, M.; Cordier, S.; Delaforge, M.; Fenaux, P.; Garnier, R.; Rico-Lattes, I.; Vasseur, P. The INSERM Expert Review on Glycol Ethers: Findings and Recommendations. *Toxicol. Lett.* **2005**, *156*, 29–37.
- (32) Fastier, A.; Herve-Bazin, B.; McGregor, D. B. INRS Activities on Risk Assessment of Glycol Ethers. *Toxicol. Lett.* **2005**, *156*, 59–76.
- (33) MacGregor, J. T.; Wilson, R. E.; Neff, W. E.; Frankel, E. N. Mutagenicity Tests of Lipid Oxidation Products in *Salmonella typhimurium*: Monohydroperoxides and Secondary Oxidation Products of Methyl Linoleate and Methyl Linolenate. *Food Chem. Toxicol.* **1985**, *23*, 1041–1047.
- (34) Burchman, P. C. Genotoxic Lipid Peroxidation Products: Their DNA Damaging Properties and Role in Formation of Endogenous DNA Adducts. *Mutagenesis* **1998**, *13*, 287–305.

CI050520J