Application of the Random Forest Method in Studies of Local Lymph Node Assay Based Skin Sensitization Data

Shengqiao Li,*,†,‡ Adam Fedorowicz,† Harshinder Singh,†,‡,§ and Sidney C. Soderholm†

Health Effect Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, West Virginia 26505, and Department of Statistics, West Virginia University, Morgantown, West Virginia 26506

Received February 10, 2005

The random forest and classification tree modeling methods are used to build predictive models of the skin sensitization activity of a chemical. A new two-stage backward elimination algorithm for descriptor selection in the random forest method is introduced. The predictive performance of the random forest model was maximized by tuning voting thresholds to reflect the unbalanced size of classification groups in available data. Our results show that random forest with a proposed backward elimination procedure outperforms a single classification tree and the standard random forest method in predicting Local Lymph Node Assay based skin sensitization activity. The proximity measure obtained from the random forest is a natural similarity measure that can be used for clustering of chemicals. Based on this measure, the clustering analysis partitioned the chemicals into several groups sharing similar molecular patterns. The improved random forest method demonstrates the potential for future QSAR studies based on a large number of descriptors or when the number of available data points is limited.

INTRODUCTION

Occupational contact dermatitis is the most commonly reported nontrauma related category of occupational illnesses in the United States.[1] Based on etiology, there are two more important types of occupational contact dermatitis: allergic contact dermatitis and irritant contact dermatitis. Irritant contact dermatitis results from contact with irritant substances, while allergic contact dermatitis is a delayed-type immunological reaction in response to contact with an allergen in sensitized individuals. Allergic contact dermatitis develops as a result of repeated skin exposures to allergenic substances.[2] The medical condition underlying allergic contact dermatitis is called skin sensitization, and the causative agents are known as skin sensitizers. Skin sensitization is a complex immunological phenomenon, which involves many biological processes and also depends on other factors such as physicochemical and structural properties of a chemical substance. Many of the metabolic and immune processes required for chemicals to be allergens are still poorly understood.[3] However, it is hypothesized that the skin sensitization potential of a chemical is related to its ability to interact with skin macromolecules and to subsequent recognition of the conjugates by the immune system as foreign.[3−5]

For many years, guinea pig tests have been the most commonly used animal based assays for assessing skin sensitization potential. In the guinea pig tests the evaluation of skin sensitization potential is carried out by measuring the inflammatory reactions of skin as a result of challenge exposure to the allergens. The measurement is often subject to interferences such as testing of dyes and especially the fact that the activity scale depends on the subjective, though often well-trained, assessment of a human observer. In contrast, a mouse assay, the Local Lymph Node Assay (LLNA), which has been recently developed and validated,[6−8] with its unambiguous, well-defined endpoint of 3-fold increase in the population of lymphocytes at the lymph nodes that drain the site of application of an allergen chemical, has the advantage over guinea pig assays in that it is more objective; requires less space for animal housing and less test substance; and is less stressful for animals.

Quantitative Structure−Activity Relationship (QSAR) modeling, which is a combination of methods in statistics and computational chemistry, is based on the examination of measured and calculated molecular descriptors of chemical compounds with known biological activity. The most informative set of descriptors are then related to the target bioactivity. Constructed this way, structure−activity relationship models provide a means of investigating and predicting the toxicological effect of a chemical with yet unknown toxicological activity. Traditional QSAR methods, which assume the continuous character of activity data, are unsuitable to model binary data. However, development of so-called binary QSAR methods, which correlate molecular descriptors with a binary expression of activity (1 = active and 0 = inactive), make it possible to construct proper predicting models of dichotomous biological outcomes.[9] The applications of binary QSAR are mostly based on logistic regression,[10,11] discriminant analysis,[10,12] probabilistic and Bayesian methods,[9] neural networks,[13] nearest neighbor,[13] and random forest methods.[14]

There have been several attempts to construct QSAR models of skin sensitization with the most successful one

* Corresponding author phone: (304)285-5960; fax: (304)285-6112; e-mail: swl4@cdc.gov.
† National Institute for Occupational Safety and Health.
‡ West Virginia University.
§ Deceased February 4th, 2005.

10.1021/ci050049u CCC: $30.25 © 2005 American Chemical Society
Published on Web 04/21/2005

proposed by Enslein et al.,[12] which in fact consists of two separate models developed using guinea pig maximization test data. These models were subsequently included in the commercial package TOPKAT (Accelrys Inc., San Diego, CA). Most of the other QSAR models were developed for a limited set of compounds within one chemical class such as sulfonate esters[15] or aldehydes.[16] This subject has been reviewed recently by Rodford et al.[3] Apart from QSAR studies, expert systems have been deployed in the prediction of skin sensitization. Knowledge base expert systems, such as DEREK for Windows (LHASA Ltd., University of Leeds, Leeds, U.K.) provides rules not based on quantitative structure descriptors but on a wide variety of structural fragments (toxicophores) that were identified in molecules with known activities. Toxicological endpoints of DEREK for Windows include the following: carcinogenicity, mutagenicity, skin sensitization, teratogenicity, irritation, and respiratory sensitization. In a previous report,[17] the performances of these two commercial packages, TOPKAT and DEREK for Windows, were compared with logistic regression models developed using the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) data set.

The relatively new algorithm of random forest has been used in studies relating to astronomy, document classification, and microarray analysis.[18] It has also been used for modeling of a High Throughout Screen (HTS) data set,[19] in prediction of fault proneness modules in a software development process,[20] and in QSAR studies of quantitative and categorical bioactivity of six publicly available QSAR data sets of chemicals.[14]

The random forest algorithm builds a group of dissimilar classification or regression trees from bootstrap samples of the training data set. For each tree, a subset of input variables is randomly chosen. For a new data point, each tree will make its own prediction, called a vote, and the final decision is based on these votes. By default, a majority voting rule is used, but users can also specify their own thresholds. The random forest approach is considered to be an improved method over the classical classification tree.

In this work we continue with a comparison of methods on a dichotomous LLNA data set of 178 compounds from the ICCVAM report[7,8] by applying the classification tree, random forest classification, and clustering analysis for predicting the skin sensitization potential of chemicals based on values of molecular descriptors. The main objectives of the study are (i) to build classification tree models of skin sensitization; (ii) to develop a backward elimination procedure for building an efficient random forest of a set of important chemical descriptors; (iii) to compare the performance of single classification trees with random forests using the ICCVAM skin sensitization data; and (iv) to cluster the compounds into different clusters using a proximity measure between two compounds obtained by using random forest models and compare the structural similarity of compounds in each cluster. We note from the results in the next section that the random forest algorithm provides a good predictive model and gives a substantial improvement over the single classification tree algorithm. The clustering results also verified that the random forest classification is reasonable and promising for a realistic data set of significance in occupational safety and health.

MATERIALS

The evaluation report of the Local Lymph Node Assay by ICCVAM provides LLNA results for 209 chemical compounds.[7,8] From 209 chemicals listed in the ICCVAM report 178 organic compounds were chosen after excluding inorganic salts, natural products, and polymers, as for these chemical substances it was impossible to assign unambiguous molecular structure or they could not be processed by the software applied to calculate molecular descriptors. Furthermore, three organic compounds were excluded as follows: streptozotocin, sodium lauryl sulfate, and benzalkonium chloride because of a continuing discussion about their actual skin sensitization potential.[4,7,8] Of 178 studied compounds, 131 were classified as active skin sensitizers and 47 were nonsensitizers.

The molecular structures of compounds were first encoded using the SMILES notation (Daylight Chemical Information Systems, Inc., Mission Viejo, CA), which was subsequently transformed to three-dimensional coordinates using Cerius[2] (Accelrys Inc., San Diego, CA), providing the initial set of 262 molecular descriptors. An additional 1204 and 747 molecular descriptors were derived using the Dragon (http://www.talete.mi.it/dragon.htm) and Molconn-Z software (eduSoft, LC, Ashland, VA), respectively. After removing repeated, constant, and almost constant variables, the total number of unique molecular descriptors amounted to 1380 per compound. The structure of this data set is such that it has a large number of descriptors but a small number of observations and the two class responses are unbalanced. Standard classification methods such as single classification tree or discriminant analysis may, with this unbalanced data, produce models with low specificity.

METHODS

Classification Tree. A binary classification tree, grown using a recursive partition algorithm, is used for classifying a response into one of two classes based on the values of molecular descriptors. At each node, the algorithm finds the predictor and the cutoff point to split the response into two branches. This is optimal with respect to the splitting criterion used in the algorithm. The three most commonly used splitting criteria are based on minimizing either the *Gini* index, entropy, or misclassification error. The splitting of the nodes is continued until the terminal nodes (leaves) are pure enough. The fully grown tree may overfit the training data and needs to be pruned back by using a criterion which balances the performance of the tree and the tree complexity. Usually the initial set of chemicals is split into two subsets including the following: the training set that is used to grow the classification tree and the testing set that is used to estimate the error rates. If the initial data set is small, application of cross-validation is used for estimating the error rates of the tree.[21]

Classification trees are usually easy to interpret. The structure of the tree can be used to determine which descriptors are important in classifying the response variable. However, the classification tree method has the disadvantage of high instability. Any error introduced in a particular node split is carried further down into the tree. The recently developed random forest algorithm offers a major improve-

ment over a single classification tree as averaging the decisions of many dissimilar trees helps reduce the variance.

Random Forest. The random forest algorithm (RF), which was developed by Leo Breiman,[22] grows a collection, called a forest, of classification trees and uses these for classifying a data point into one of the classes. Two types of randomness, bootstrap sampling and random selection of input variables, are used in the algorithm to make sure that the classification trees grown in the forest are dissimilar and uncorrelated from each other. Growing a forest of trees and using randomness in building each classification tree in the forest leads to better predictions compared to a single classification tree and helps to make the algorithm robust to noise in the data set.

A forest is grown by using *ntree* bootstrapped samples each of size *n* randomly drawn from the original data of *n* points with replacement. This first type of randomization helps in building an ensemble of trees and in reducing dependence among the trees. About two-thirds of the data set are used to grow a classification tree. About one-third of the data are left, called Out Of Bag (OOB) Data. These data are used to obtain unbiased estimates of correct classification rates and variable importance. The second type of randomness is used during building classification trees. For each node of a tree, the RF algorithm randomly selects *mtry* variables and uses only them to determine the best possible split using the *Gini* index as the splitting criterion. This algorithm is fairly robust to the choice of the number *mtry*, the value of which is usually taken to be the square root of the total number of variables.[15] In contradiction to the classification tree approach the random forest trees are grown to the full length and are not pruned back. Predictions for test data are carried out either by the majority vote of classification trees in the forest or are based on a threshold value selected by the user. The number of trees (*ntree*) to be grown in the forest is chosen appropriately to achieve low error rate of convergence.

When applying the random forest algorithm, it is not necessary to set aside a portion of data as test data or to use cross-validation in order to estimate correct classification rates. Instead, the OOB part of the data is used. Each tree in the forest is grown using the bootstrapped sample, and the OOB part is subsequently processed by the grown tree. This gives rise to classification for each point in the OOB part of that bootstrapped sample. This means that about one-third of the trees in the random forest give a prediction for each point in the original data. The final classification of a particular data point is decided on the basis of majority vote or user defined threshold, among the trees predicting that data point. The unbiased estimates of true classification rates are calculated by comparing the OOB set classification made by the forest to the experimentally observed classes to which the data points belong.

The random forest algorithm can also give two measures of importance for the variables used in growing trees. The first measure, Mean Decrease in Accuracy, uses the OOB part of the data set. Values of a variable (X_i) in the OOB part are randomly permuted, and the resulting new OOB data points are processed by the random forest. The margin of a data point is defined as the difference between proportion of the votes for the correct class and the proportion of votes for the other class. The percent decrease in the margin resulting from random permutation in the values of the variable X_i averaged across all data points in the OOB set gives the first measure of importance of variable X_i. The second importance measure, Mean Decrease in Gini, is based on the *Gini* index used also as the splitting criterion. As a result of each split, the value of the *Gini* index decreases based on one of the *mtry* variables chosen randomly. Therefore, the second measure of variable importance for variable X_i can be based on the sum of all decreases of the *Gini* index in the forest. This measure is defined as this sum divided by the total number of trees in the forest. The first measure of importance is considered more reliable than the second measure.[19] In general, a descriptor is considered important if at least one of the above measures has a large value.

Another important feature of Random Forest algorithm is that it does not overfit, i.e., as the number of trees increases, the generalization error almost surely converges, as demonstrated by Breiman.[22] Moreover, using an ensemble of trees grown from bootstrap samples and randomly selected input variables to classify new data can usually reduce very much the prediction variance but not increase too much the bias. Other descriptions of the random forest algorithm can be found in refs 18 and 19.

RESULTS AND DISCUSSION

In this work, ***R,*** an open source statistical computing software from the R Project for Statistical Computing was used to perform data analysis.[23]

Classification Tree. The "*tree*" package in ***R*** was used to fit binary classification trees. The *Gini* index was used as the splitting criterion for choosing a conditional descriptor and the threshold value of the descriptor during the process of building classification trees. The *Gini* index is an indicator of impurity of tree nodes, and it is defined as $Gini = \sum_{j \neq k} p_{ij} p_{ik}$, where p_{ij} and p_{ik} are the proportions of j class and k class in node i, respectively. If the node consists of chemicals of only one class, then it is considered pure, and the index value equals 0. If the node has the same number of observations for each class, which is the most impure possibility, then the index value is close to $1-1/k$, where k is the number of classes. For the binary (active/inactive) classification systems, the *Gini* index can be defined as $Gini = 2p_{i0}p_{i1}$, where p_{i0} is the proportion of inactive compounds and p_{i1} is the proportion of active compounds in node i.

Using the *Gini* index, a full tree with 25 descriptors and 26 leaves, each of whose leaves is either small (<10 compounds) or contains only active or inactive compounds, was constructed. Since a full and complex tree is often difficult to interpret and may be overfitted, it was pruned to the optimal size. The optimal size was determined by the average error rate, i.e., the proportion of misclassified inactive and active chemicals, obtained by 20 repetitions of 10-fold cross-validation. In each 10-fold cross-validation, the data set was randomly partitioned into 10 subsets, and each subset was an independent testing sample for the tree, a cross-validation tree, grown separately from the others. These cross-validation trees were pruned to generate sequences of nested trees. The corresponding 10% of testing data was run through these trees; misclassification rates were collected and then averaged over all runs of cross-validation. Figure 1

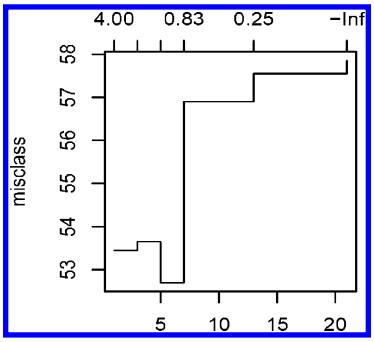


Figure 1. Classification tree cross-validation errors.

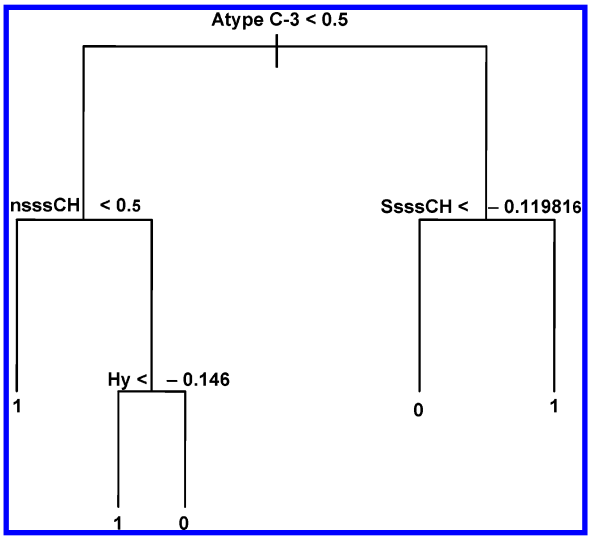


Figure 2. Structure of proposed classification tree.

Table 1. Descriptors Selected by the Classification Tree

descriptor	definition	descriptor class
Atype C-3	count of atom types CHR3	atom type counter
nsssCH	count of atom types	atom type counter
SsssCH	sum of atom type E-state	atom-type EState sum
Hy	hydrophilic factor	molecular property

shows summary misclassification rates along different tree sizes (number of terminal nodes). The labels at the top of the plot are the cost-complexity parameters for pruning.

The minimum value of misclassification rate at tree size around 5 was considered to be the optimal tree size (Figure 1). Subsequently, the aforementioned full tree trained on the whole data set was pruned to this optimal size of 5. The resulting classification tree, presented in Figure 2, has four molecular descriptors, and their short definitions are provided in Table 1. They form the final classification tree model for predicting skin sensitization, which is subsequently denoted as CT4 (**C**lassification **T**ree with **4** descriptors).

The performance of the proposed classification tree model is reported in Table 2. Only one active chemical is misclassified as inactive, and the sensitivity of the classification tree is very high 99.2%. However, the misclassification rate for inactive chemicals is not promising as 31 of them are classified as active ones. This is reflected in a low specificity value of 34%. It seems that for the unbalanced training data set with the number of active chemicals exceeding significantly the number of inactive ones, the proposed classification tree gives corresponding unbalanced performance with high sensitivity and low specificity values. Comparisons

Table 2. Performance Statistics of the Classification Tree

	classified as nonactive	classified as active
nonactive	16	31
active	1	130
sensitivity	99.2%	
specificity	34%	
average error rate[a]	18.0%	
false positive[b]	19.3%	
false negative[c]	5.9%	
cross-validation error	29.6% (52.7/178)	

[a] Average error rate = proportion of misclassified chemicals. [b] False positive rate = proportion of inactive compounds among those which are predicted active compounds. [c] False negative rate = proportion of active compounds among those which are predicted inactive compounds.

between predicted and experimental skin sensitization activity for each individual chemical compound are tabulated in Tables 6 and 7 for inactive and active chemicals along with random forest results, correspondingly.

Random Forest. The "*randomForest*" package in ***R*** was used to build random forest models for classification. When running *randomForest*, one can choose the number of variables randomly selected for tree growing, *mtry*. Usually, it is the square root of the total number of input variables that will be used to grow each single tree. However, this does not lead to any simplification of the model, as all the variables may be used in a random forest. These variables may not be independent of each other, and thus data with a large number of variables usually contains redundant information. The reduced models, obtained by dropping unimportant variables, are almost always desirable in order to save the cost of data collection and enhance prediction correctness. The package *randomForest* has built-in variable importance measures that can be used in a screening process of unimportant variables. However, as the random forest is a recently developed tool there are only a few published applications that describe the variable selection process.[14] There are no statistical tests or importance cutoff points for variable removal. So unlike linear regression models, the random forest technique does not have forward/backward or stepwise procedures of automated variable selection implemented. It seems that one way to select variables is to compare the random forest performance built from different sets of variables. Due to variable interdependency, dropping a variable should change the importance of other variables. Therefore, by simply selecting the most important variables from the full model that is based on the whole set of descriptors may not necessarily result in the best model. For QSAR and other applications, where the number of variables can be quite large, the exhaustive search of all possible models is sometimes hard or even impossible to conduct if prior knowledge about these variables is not available. For instance, for the 1380 variables in this data set, the total number of combinations is $2^{1380} - 1$. Therefore, application of an efficient selection algorithm is crucial to producing a reasonably reduced random forest model. Here, a two-stage backward elimination algorithm is proposed. It is an improved version of an algorithm discussed by Svetnik et al.,[14] and it introduces a second fine-tuned stage.

In the first stage, a full random forest model is built using all available 1380 descriptors. Then half of those descriptors with the lowest importance are dropped from the model. The

```
Random_Forest_Backward_Elimination (data)
{
 stage 1:
        n ← number of variables in data;
        m ← ⌊log2(n)⌋;
        initialize   forest_list[m], perf[m];
        for i from 1 to m loop;
            if i = 1 then rf ← build random forest from all variables of data;
            else rf ← build random forest from 2^(m+1-i) top important variables of rf;
            forest_list [i] ← rf;
            perf [i] ← performance of rf;
        end loop;
        better ← argmax (perf [k]) for k in [1, m];
        pre_better ← better − 1;
        rf ← forest_list [pre_better];          /* This random forest goes to stage 2*/
stage 2:
        reinitialize   forest_list perf;
        n ← number of variables of rf;
        for j from n to 2 loop;
            rf   ← build random forest from j top important variables of rf;
            perf [n + 1 - j] ← performance of rf;
            forest_list [n +1 - j] ← rf;
        end loop;
        best ← argmax (perf [k]) for k in [1, n-1];
        best_rf ← forest_list[best];            /* This gives final random forest model */
        output best_rf;
}
```

Figure 3. A two-stage backward elimination procedure of random forest classification.

remaining descriptors are used as the input set for the formation of a new random forest, from which another half of the least important descriptors are dropped. Subsequently, this process is repeated until only a few descriptors are remaining. At this point, one can quickly find the random forest model with good classification rates and a small number of descriptors.

However, a still better random forest can be located in the close neighborhood of this model. At the second stage, to locate this model one can apply a more fine-tuned backward elimination using the random forest that existed just before the best one found in the stage one. Now only one least important descriptor is discarded at each step.The procedure can be continued to reach a predefined number of descriptors or when all but two descriptors are left because random forest needs at least two variables to work. The algorithm of the full selection procedure is listed in Figure 3.

The algorithm in Figure 3 may not find the best possible model as it may need descriptors that were excluded at the first elimination stage. However, it is expected to result in a model that has a better performance compared to a random forest model based on just a set of descriptors with highest variable importance from the initial random forest of all descriptors. And to some extent, this backward elimination of descriptors also alleviates the possible overfitting problem of a full model.

To evaluate and compare different size forest models, the weighted classification accuracy (WA), which is the average of sensitivity and specificity,[19] was used as a performance

Table 3. AUCs and WAs of Random Forests of Different Number Descriptors

model size	AUC	WA (voting threshold)
15	0.881	80.0 (0.65)
14	0.878	80.9 (0.70)
13	0.878	79.2 (0.55)
12	0.862	79.8 (0.60)
11	0.874	80.0 (0.75)
10	0.877	83.5 (0.70)
9	0.878	80.2 (0.70)
8	0.856	79.3 (0.65)
7	0.858	82.0 (0.60)
6	0.845	80.5 (0.60)
5	0.826	81.1 (0.65)
4	0.828	80.1 (0.65)
3	0.829	77.7 (0.75)

criterion. To drop unimportant variables from a forest, the Mean Decrease in Accuracy Decrement importance measure was used. The "just before best" random forest from the first stage consisted of a set of 32 descriptors. By dropping the least important descriptor incrementally, a total number of 31 random forests were obtained in the second stage. From these forests, one can choose final model(s) using classification performance as a criterion.

Classification of a chemical in the random forest technique is determined by the number of votes from all classification trees in the forest. By changing the threshold of voting, different sensitivity, specificity, and related WA values can be obtained. In this study, a sequence of thresholds from 0 to 1 stepped by 0.05 was used. One can decide to choose a threshold value that produces the best classification results. Using these values of sensitivity and specificity, the receiver operating curves (ROC) plot of sensitivity vs 1-specificity can be constructed. The area under the ROC curve, which is denoted AUC, is often used as an additional performance index. Table 3 lists several random forest models, obtained by using the proposed backward elimination procedure, with a decreasing number of descriptors, along with their AUC and their maximum WA values at the corresponding voting threshold. The model of size 10 is the best one found by the proposed algorithm.

Results presented in Table 3 show that three forests models, having 10 (RF10), 7 (RF7), and 5 (RF5) descriptors, have similar performance. All weighted accuracies of these three forests are over 81%, and their AUCs are significantly high, ranging from 0.83 to 0.88. These three models are much simpler than the full random forest model, and they seem to perform significantly better even than more complex models

Table 4. Random Forest Results Using Different Number Descriptors and Thresholds

no. of descriptors	threshold	specificity (%)	sensitivity (%)	false negative (%)	false positive (%)	error rate (%)
1380	0.7	61.7	65.6	60.8	17.3	35.4
	0.6	48.9	82.4	50.0	18.2	26.4
	0.5	29.8	94.7	33.3	21.0	22.5
	0.4	17.0	99.2	11.1	23.1	22.5
10	0.7	83.0	84.0	35.0	6.8	16.3
	0.6	66.0	88.6	32.6	12.1	17.4
	0.5	51.1	93.9	25.0	15.7	17.4
	0.4	40.4	99.2	5.0	17.7	16.3

with a larger number of molecular descriptors. This follows the main goal of predictive model construction where one tries to develop simpler models that are easier to interpret and implement. From these three forests, the RF10 model got the maximum observed weighted accuracy of 83.5% at threshold of 0.7 with a high AUC value of 0.877. Even more important is that the sensitivity and specificity rates are balanced and significantly high reaching 84% and 83%, respectively. It is important to notice that in the case of the single classification tree these numbers were highly unbalanced. The ROC curves for the RF10, RF7, and RF5 models are shown in Figure 4. If one uses only the 10 descriptors having the highest variable importance from the full model, it results in a model with an AUC of 0.721 and a maximum weighted classification accuracy of 73.5% at 0.65 threshold. Thus, the same size model obtained by the proposed backward elimination procedure has a significantly better performance with an AUC 0.877 and a weighted classification accuracy 83.5% than the previous model. This illustrates the benefit of the proposed backward elimination procedure for building a random forest model.

Table 4 lists the values of sensitivity, specificity, false negative and false positive rates, and average error rate for the majority voting that uses the standard threshold, 0.5, and the threshold that gives rise to the highest total accuracy for that model. This table shows that the reduced model, RF10, is much better than the full models with 1380 descriptors with respect to values of specificity, false negative, and total error rate. By adjusting the threshold, the specificity value increased without significant loss of sensitivity.

Compared with the single classification tree CT4, the error rate of RF10, around 17% is lower than the resubstitution error rate of 18% and much lower than the cross-validation error of 29.6% (Table 2). The calculation of error rates for

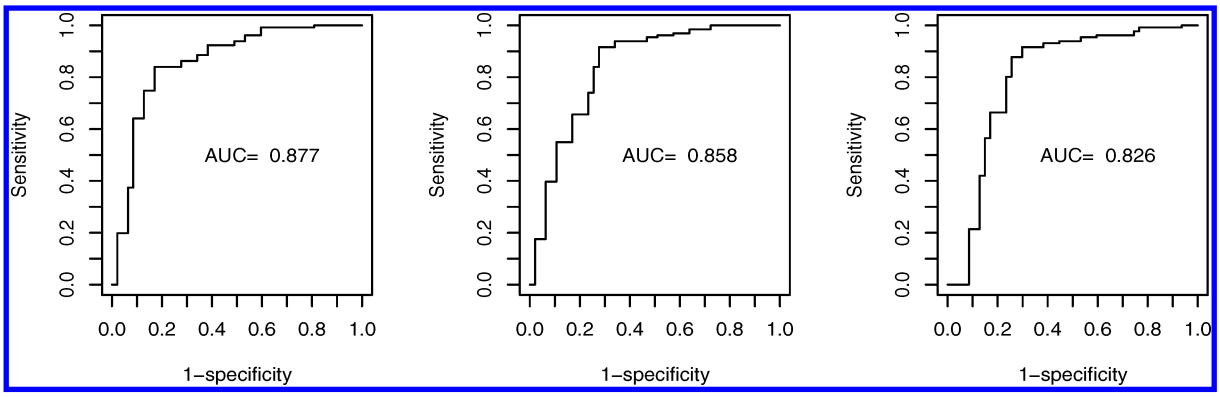


Figure 4. ROC curves of random forest models. Left based on 10 descriptors (RF10), middle based on 7 descriptors (RF7), and right based on 5 descriptors (RF5).

Table 5. Important Descriptors Used in the Proposed Random Forest Models and Their Importance Values

	descriptor			model		
no.	name	definition	class	RF10	RF7	RF5
1	Gmin	minimum E-State	EState	7.00	9.96	12.0
2	HATS1u	leverage-weighted autocorrelation of lag 1/ unweighted	GETAWAY	6.02		
3	HATS4u	leverage-weighted autocorrelation of lag 4/ unweighted	GETAWAY	6.16	8.82	
4	Hmaxpos	maximum positive H E-State	H-Bond EState	7.54	10.2	13.3
5	Mor32p	3D-Morse signal 32/ weighted by atomic polarizabilities	3D-Morse	6.45	7.93	
6	R4e+	R maximal autocorrelation of lag 4/ weighted by Sanderson electronegativities	GETAWAY	6.22	9.13	14.0
7	R4u+	R maximal autocorrelation of lag 4/ unweighted	GETAWAY	6.07		
8	SEigp	eigenvalue sum from polarizability weighted distance matrix	eigenvalue based indices	8.03	10.9	13.4
9	SsssCH	sum of atom-type E-state	atom type EState sums 2	6.21		
10	sumdelI	sum of delta-I values	complexity indices	8.86	11.8	16.3

random forest models are based on the OOB part of data. Because the OOB data are not used in growing of the random forest, therefore, it can be used as a test data set that is independent of the training data set. Consequently, these error rates are unbiased. The resubstitution error rate of CT4 is gained by using all the data points for training as testing data set. They are not independent, and the error is usually overoptimistic. Actually the resubstitution error rates are very small for random forests, although we did not use them for performance assessment. The OOB error of random forest should be compared with the cross-validation error of classification, and it is much improved in this case, 16.5% vs 29.6%.

Short definitions of the molecular descriptors and their importance values evaluated in three proposed random forest models are presented in Table 5. The RF10 model shares only one descriptor, SsssCH, with the classification tree model. The SsssCH descriptor is related to the presence and the topological properties of the tertiary (−CH−) carbon atom. However, this descriptor has a nonzero value for less than 20% of chemicals. It seems that it was selected by the statistical methods used mainly as an "improvement" to an already existing model. This is confirmed by its low importance value of 6.21 (Table 5) and that it is missing in models RF7 and RF5. Tables 6 and 7 list detailed classification information for each studied chemical and compare predicted skin sensitization using all models (three random forests and one classification tree) with experimental LLNA activity.

Clustering Analysis. In addition to classifying chemicals as active or inactive, random forest can also provide a proximity measure defined as the probability of assigning two chemicals to the same node. This can be obtained by measuring how often a given pair of chemicals was assigned to the same node through all classification trees of a random forest. Using this property, all chemicals were grouped using the hierarchical clustering algorithm, *hclust*, in ***R*** into a dendrogram (Figure 5). The average linkage was employed, and the value of *1-proximity* was used as the distance measure. This approach was applied only to the RF10 model as it is considered to be the best model from all proposed random forest models.

Based on the obtained hierarchical dendrogram and using the similarity between molecular structures of neighboring chemicals, rather than a statistically defined cutoff parameter, several branches and their groups were selected. This approach was applied for two reasons as follows: large diversity of chemicals from the ICCVAM data set and to give some insight into how random forests combine structurally similar chemicals. Additionally, this approach helped in highlighting some of the interesting points of the proposed random forest model of skin sensitization. It is worth mentioning that application of this approach led to formation of several interesting groups of chemicals that shared similar molecular patterns and often expressed almost the same skin sensitization activity (Table 8).

The first A group contains mostly halogeno-derivatives of nonane, a hydrocarbon with nine carbon atoms, including nonanoyl chloride. A small sub-branch contains two polychlorophenols (A108, A124), both of which are active. Most of chemicals in the A group are active, and only bromononane is inactive. Most bromoalkanes with longer aliphatic chain are skin sensitizers as it can be seen in group D. Bromoalkanes with a very short alkyl group like bromobutane are mostly inactive. The medium length of the aliphatic chain of bromononane thus may position it between the active and inactive classes. This is supported by experimental data which is just below the threshold used to define chemicals as skin sensitizers.[24] Any chemical above this threshold is considered to be active. Interestingly, inactive bromononane is very often placed together in the same node with active chlorononane. This shows that random forest, and probably other similar algorithms, with the current set of descriptors may have potential difficulties assigning a proper activity to a group of structurally similar chemicals, with different categorical activities, when the full training set is small and very diverse.

The second B branch contains simple phenyl derivatives with only one or two substituents such as −OH, $-NH_2$, −Cl, and −COCl. Interestingly, all three inactive chemicals in this group are in the same sub-branch. Similar to the B group is the E branch, although its phenol or aniline derivatives have more and bulkier substituents attached to the phenyl ring. This branch seems to be an expanded version of the B group. All chemicals in the E group are active.

The C group is a more structurally diverse branch. However, it seems that most of the chemicals here have surfactant-like properties with a polar head, mostly sulfonates or sulfonic acid residues, and a nonpolar tail, which may have from seven to sixteen carbon atoms. The polar head is usually attached to a phenyl ring, which is further connected to the nonpolar tail through ester- or ether-like linkage.

The D group is an interesting collection of compounds from two chemical classes: (i) alkyl and aliphatic acid halides, with the length of the alkyl chain from twelve to sixteen carbon atoms, and (ii) polycyclic aromatic hydro-

Table 6. Classification of LLNA Nonactive Chemicals

code	chemical name	CAS no.	classification			
			CT4	RF10	RF7	RF5
N01	2-acetamidefluorene	53-96-3	0	0	1	1
N02	4-aminobenzoic acid	150-13-0	1	0	0	0
N03	3-(benzenesulfonyloxymethyl)-5,5-dimethyldihydro-2(3H)-furanone	154750-24-0	1	0	0	0
N04	benzocaine	94-09-7	1	0	1	1
N05	benzoyloxy-3,5-benzene dicarboxylic acid	102059-70-1	0	0	0	0
N06	1-bromobutane	109-65-9	1	0	0	0
N07	1-bromohexane	111-25-1	1	0	0	0
N08	1-bromononane	693-58-3	1	1	1	1
N09	4-chloroaniline	106-47-8	1	0	1	1
N10	chlorobenzene	108-90-7	1	1	1	1
N11	3-(chlorobenzenesulfonyloxymethyl)-5,5-dimethyldihydro-2(3H)-furanone	154750-28-4	0	0	0	0
N12	2-chloroethanol	107-07-3	1	1	0	0
N13	2,4-dichloronitrobenzene	611-06-3	1	0	0	0
N14	di-2-furanylethanedione	492-94-4	1	0	0	0
N15	dimethyl isophthalate	1459-93-4	1	0	0	0
N16	5,5-dimethyl-3-(mesyloxymethyl)dihydro-2(3H)-furanone	154750-22-8	1	0	0	0
N17	5,5-dimethyl-3-(methoxybenzenesulfonyloxymethyl)dihydro-2(3H)-furanone	154750-23-9	0	0	0	0
N18	5,5-dimethyl-3-(nitrobenzenesulfonyloxymethyl)dihydro-2(3H)-furanone	154750-29-5	0	0	0	0
N19	5,5-dimethyl-3-(tosyloxymethyl)dihydro-2(3H)-furanone	154060-50-1	1	0	0	0
N20	ethyl methanesulfonate	62-50-0	1	1	1	1
N21	geraniol	106-24-1	1	1	1	1
N22	hexane	110-54-3	1	0	0	1
N23	hydrocortisone	50-23-7	1	0	0	0
N24	4-hydroxybenzoic acid	99-96-7	1	0	0	0
N25	2-hydroxypropyl methacrylate	923-26-2	0	0	0	0
N26	2-propanol	67-63-0	0	0	0	0
N27	kanamycin A	8063-07-8	0	0	0	0
N28	lactic acid	50-21-5	0	0	0	0
N29	6-methylcoumarin	92-48-8	1	1	1	1
N30	methyl salicylate	119-36-8	1	0	0	0
N31	*N*′-(4-methylcyclohexyl)-*N*-(2-chloroethyl)-*N*-nitrosourea	13909-09-6	1	0	0	0
N32	neomycin	1405-10-3	0	0	0	0
N33	2-nitrofluorene	607-57-8	1	1	1	1
N34	octadecylmethane sulfonate	31081-59-1	0	0	1	1
N35	phenol	108-95-2	1	0	0	0
N36	phthalic acid diethyl ether	84-66-2	1	0	0	0
N37	propylene glycol	57-55-6	0	0	1	1
N38	propylparaben	94-13-3	1	0	0	0
N39	resorcinol	108-46-3	1	1	1	1
N40	salicylic acid	69-72-7	1	0	0	0
N41	streptomycin	57-92-1	0	0	0	0
N42	sulfanilamide	63-74-1	1	0	0	0
N43	sulfanilic acid	121-57-3	0	0	0	0
N44	tartaric acid	87-69-4	0	0	0	0
N45	tixocortol-21-pivalate	55560-96-8	0	0	1	1
N46	trimethylammonium-3-tolyl-ϵ-caprolactamide chloride	374680-04-3	1	0	0	0
N47	α-trimethylammonium-4-tolyloxy-4-benzenesulfonate	264869-81-0	1	0	0	0

carbons (PAH) that in a few cases are also halogenated. In general, this group is very homogenic and contains mostly heavy and highly hydrophobic chemicals. The following F group is somehow auxiliary to the previous C and D group. In fact, it has two subgroups of chemicals which share some structural similarity with classes of chemicals seen in previous groups such as PAH, surfactants, highly hydrophobic, and alkyl halides. Interestingly, this group contains three aldehydes, all active, and the aldehyde group is coupled to a double bond R═C−CHO. All chemicals in the F group are active.

The G branch contains mostly chemicals that have two phenyl rings linked together either through ester, ether, methylene, and peroxide or lactone groups. All but one chemical in this group is active. The following H group although more diverse shares some structural similarity with the previous G branch as it contains several chemicals with two linked rings. However, in some cases the second ring is a furanone derivative instead of a phenyl ring. In general, most chemicals in this group have a phenyl ring attached to a generally defined ester group, in addition to other acid derivatives: amide or anhydride. Another major difference between this group and the previous one is that it has a significant fraction of inactive chemicals.

The I branch, although it is the biggest and the most diverse one, contains several distinctive subgroups that share similar molecular patterns like the following: aromatic nitro derivatives; nitro/nitrozo-amines; aliphatic halides with 6−8 carbon atoms in the alkyl chain; amines with short aliphatic chains; catechol derivatives; sulfonates and sulfates with short aliphatic chains; aldehydes coupled to a double bond; thiols; and aromatic halides. Most of the chemicals in this large group are active.

The J branch contains a group of only four low weight, highly hydrophobic chemicals with 4−8 carbon atoms in their structures. Most chemicals are inactive. The L branch has only three surfactants with a very long hydrophobic aliphatic chain of 18−19 carbon atoms.

Table 7. Classification of LLNA Active Chemicals

code	compound	CAS no.	classification			
			CT4	RF10	RF7	RF5
A001	abietic acid	514-10-3	1	1	1	1
A002	2-(*N*-acetoxy-acetamido)fluorene	6098-44-8	1	1	1	1
A003	3-acetylphenyl benzoate	139-28-6	1	1	1	1
A004	4-allylanisole	140-67-0	1	1	1	1
A005	ammonium thioglycolate	5421-46-5	1	1	1	1
A006	2-aminophenol	95-55-6	1	1	1	1
A007	3-aminophenol	591-27-5	0	0	1	1
A008	aniline	62-53-3	1	1	1	1
A009	1,2-benzisothiazol-3(2H)-one	2634-33-5	1	1	1	1
A010	benzopyrene	50-32-8	1	1	1	1
A011	1,4-benzoquinone	106-51-4	1	1	1	1
A012	benzoyl chloride	98-88-4	1	1	1	1
A013	benzoyl peroxide	94-36-0	1	1	1	1
A014	benzyl bromide	100-39-0	1	1	1	1
A015	12-bromo-1-dodecanol	3344-77-2	1	1	1	1
A016	12-bromododecanoic acid	73367-80-3	1	1	1	1
A017	1-bromododecane	143-15-7	1	1	1	1
A018	1-bromoheptadecane	3508-00-7	1	1	1	1
A019	1-bromohexadecane	112-82-3	1	1	1	1
A020	1-bromooctadecane	112-89-0	1	1	1	1
A021	1-bromopentadecane	629-72-1	1	1	1	1
A022	1-bromotetradecane	112-71-0	1	1	1	1
A023	7-bromotetradecane	74036-97-8	1	1	1	1
A024	2-bromotetradecanoic acid	10520-81-7	1	1	1	1
A025	1-bromotridecane	765-09-3	1	1	1	1
A026	1-bromoundecane	693-67-4	1	1	1	1
A027	2,3-butanedione	431-03-8	1	1	1	1
A028	butyl glycidil ether	2426-08-6	1	1	1	1
A029	chloramine T	127-65-1	1	1	1	1
A030	chlorpromazine	50-53-3	1	1	1	1
A031	2-chloromethylfluorene	91679-67-3	1	0	1	1
A032	1-chloromethylpyrene	1086-00-6	1	1	1	1
A033	5-chloro-2-methyl-4-isothiazolin-3-one	26172-55-4	1	1	1	1
A034	1-chlorononane	2473-01-0	1	0	1	1
A035	1-chlorooctadecane	3386-33-2	1	1	1	1
A036	1-chlorotetradecane	2425-54-9	1	1	1	1
A037	cinnamic aldehyde	104-55-2	1	1	1	1
A038	citral	5392-40-5	1	1	1	1
A039	clotrimazole	23593-75-1	1	1	1	1
A040	cocoamidopryl betaine	61789-40-0	1	0	1	1
A041	dodecyl methanesulfonate	51323-71-8	1	1	1	1
A042	dodecyl thiosulfonate	127089-67-2	1	1	1	1
A043	1,2-dibromo-2,4-dicyanobutane	35691-65-7	1	1	1	1
A044	diethyl sulfate	64-67-5	1	1	1	1
A045	diethylenetriamine	111-40-0	1	1	1	1
A046	3,4-dihydrocoumarin	119-84-6	1	1	1	1
A047	dihydroeugenol	2785-87-7	1	1	1	1
A048	2,4-dinitrochlorobenzene	97-00-7	1	1	1	1
A049	2,4-dinitrofluorobenzene	70-34-8	1	0	1	1
A050	2,4-dinitrothiocyanobenzene	1594-56-5	1	0	0	0
A051	7,12-dimethylbenz[*a*]anthracene	57-97-6	1	1	1	1
A052	5,5-dimethyl-3-(bromomethyl)dihydro-2(3H)-furanone	154750-20-6	1	1	1	1
A053	5,5-dimethyl-3-(thiocyanatomethyl)dihydro-2(3H)-furanone	154750-32-0	1	1	1	1
A054	5,5-dimethyl-3-methylenedihydro-2(3H)-furanone	29043-97-8	1	0	0	0
A055	*N*,*N*-dimethyl-1,3-propanediamine	109-55-7	1	1	1	1
A056	dimethyl sulfostearate	99785-70-3	1	0	0	0
A057	dimethyl sulfate	77-78-1	1	1	0	1
A058	disodium 1,2-diheptanoyloxy-3,5-benzenedisulfonate	374678-48-5	1	1	0	0
A059	ethylene glycol dimethacrylate	97-90-5	1	1	1	1
A060	ethylenediamine	107-15-3	1	1	1	1
A061	1-ethyl-3-nitro-1-nitrosoguanidine	4245-77-6	1	1	1	1
A062	*N*-ethyl-*N*-nitrosourea	759-73-9	1	0	0	0
A063	eugenol	97-53-0	1	1	1	1
A064	fluorescein isothiocyanate	25168-13-2	1	1	1	1
A065	formaldehyde	50-00-0	1	0	1	1
A066	glyoxal	107-22-2	1	1	1	1
A067	hexadecanoyl chloride	112-67-4	1	1	1	1
A068	hexyl cinnamic aldehyde	101-86-0	1	1	1	1
A069	hydroquinone	123-31-9	1	1	1	1
A070	hydroxycitronellal	107-75-5	1	0	1	1
A071	2-hydroxyethyl acrylate	818-61-1	1	0	0	0
A072	imidazolidinyl urea	39236-46-9	1	1	1	1

Table 7. (Continued)

code	compound	CAS no.	classification			
			CT4	RF10	RF7	RF5
A073	1-iodohexadecane	544-77-4	1	1	1	1
A074	1-iodohexane	638-45-9	1	0	1	1
A075	1-iodononane	4282-42-2	1	0	1	1
A076	1-iodooctadecane	629-93-6	1	1	1	1
A077	1-iodotetradecane	19218-94-1	1	1	1	1
A078	isoeugenol	97-54-1	1	1	1	1
A079	isononanoyloxybenzene sulfonate	109363-00-0	1	0	0	0
A080	isophorone diisocyanate	4098-71-9	1	0	0	0
A081	2-mercaptobenzothiazole	149-30-4	1	1	1	1
A082	2-methoxy-4-methylphenol	93-51-6	1	1	1	1
A083	4-methylaminophenol sulfate	55-55-0	1	1	1	1
A084	3-methylcatechol	488-17-5	1	1	1	1
A085	4-methylcatechol	452-86-8	1	1	1	1
A086	3-methylcholantrene	56-49-5	1	1	1	1
A087	3-methyleugenol	186743-26-0	1	1	1	1
A088	5-methyleugenol	186743-25-9	1	1	1	1
A089	6-methyleugenol	186743-24-8	1	1	1	1
A090	methyl dodecanesulfonate	2374-65-4	1	1	1	1
A091	methyl hexadec-2-ene sulfonate	54612-23-6	1	1	1	1
A092	methyl methanesulfonate	66-27-3	1	1	1	1
A093	2-methyl-4,5-trimethylene-4-isothiazolin-3-one	82633-79-2	1	1	1	1
A094	3-methoxyphenylbenzoate	5554-24-5	1	1	1	1
A095	1-methyl-3-nitro-1-nitrosoguanidine	70-25-7	1	1	1	1
A096	methylene diphenyl diisocyanate	101-68-8	1	1	1	1
A097	*N*-nitroso-*N*-methylurea	684-93-5	1	1	1	1
A098	nonanoyl chloride	764-85-2	1	1	1	1
A099	α-naphthoflavone	604-59-1	1	1	1	1
A100	*β*-naphthoflavone	6051-87-2	1	1	1	1
A101	4-nitrobenzyl bromide	100-11-8	1	1	1	1
A102	4-nitrobenzyl chloride	100-14-1	1	1	1	1
A103	4-nitroso-*N*,*N*-dimethylaniline	138-89-6	1	1	1	1
A104	octadecanoyl chloride	112-76-5	1	1	1	1
A105	octyl gallate	1034-01-1	1	1	1	1
A106	oxazolone	1564-29-0	1	1	1	1
A107	penicillin G	61-33-6	1	1	1	1
A108	pentachlorophenol	87-86-5	1	1	1	1
A109	phenyl benzoate	93-99-2	1	0	1	1
A110	3-phenylenediamine	108-45-2	1	1	1	1
A111	4-phenylenediamine	106-50-3	1	0	0	0
A112	phthalic anhydride	85-44-9	1	0	0	0
A113	picryl chloride	88-88-0	1	1	1	1
A114	*β*-propiolactone	57-57-8	1	1	1	1
A115	propyl gallate	121-79-9	1	1	1	1
A116	1-propyl-3-nitro-1-nitrosoguanidine	13010-07-6	1	1	1	1
A117	*p*-xylene	106-42-3	1	1	1	1
A118	pyridine	110-86-1	1	1	1	1
A119	sodium 4-(2-ethylhexyloxycarboxy)benzenesulfonate	264869-77-4	1	1	1	1
A120	sodium 4-sulfophenyl acetate	46331-24-2	1	0	1	1
A121	sodium benzoyloxy-2-methoxy-5-benzenesulfonate	159783-19-4	1	1	1	1
A122	sodium benzoyloxybenzenesulfonate	56265-04-4	1	1	1	1
A123	sodium norbornanacetoxy-4-benzenesulfonate	374679-08-0	1	0	1	1
A124	tetrachlorosalicylanilide	1154-59-2	1	1	1	1
A125	tetramethyl thiuram disulfide	137-26-8	1	1	1	0
A126	1-thioglycerol	96-27-5	1	1	1	1
A127	2,4,5-trichlorophenol	95-95-4	1	1	1	1
A128	2,4,6-trichloro-1,3,5-triazine	108-77-0	1	1	1	1
A129	trimellitic anhydride	552-30-7	1	1	1	1
A130	3,5,5-trimethylhexanoyl chloride	36727-29-4	1	1	1	1
A131	4-vinylpyridine	100-43-6	1	1	1	1

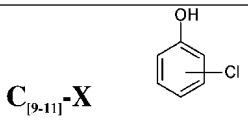
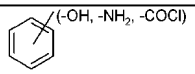
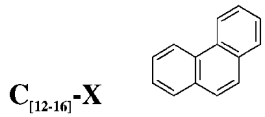
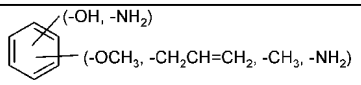
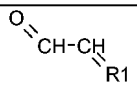
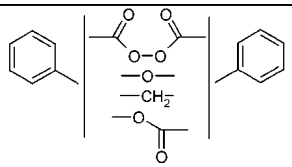
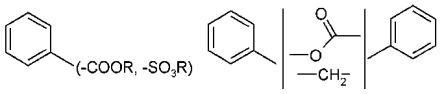
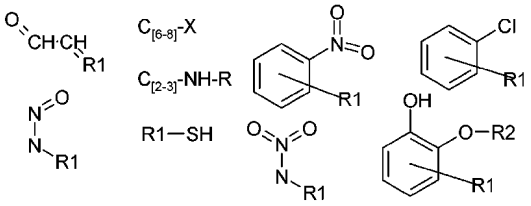
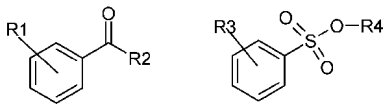
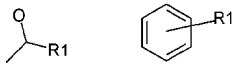
Most of the chemicals in the K group are inactive. This group, in fact, is composed of two separate dendrogram branches that are mostly simple (from the first branch) or more complex (from the second branch) derivatives of benzoic and benzenesulfonic acids. Interestingly, all chemicals with the nonsubstituted carboxyl group are inactive.

The M branch contains large inactive polycyclic chemicals. What is surprising is that this cluster of huge polycyclic molecules, which are all inactive, also contains the smallest chemical—formaldehyde—with just four atoms, which is an active skin sensitizer.

The N branch contains two sets of chemicals. The first set consists of low weight alkyl alcohols, all of them inactive. The second set contains phenyl derivatives, mostly active, with molecular structures sharing some similarity with chemicals from branches A, B, E, H, and I.

In summary, analysis of proximity measure data from the RF10 skin sensitization model provides additional informa-

Table 8. Leading Structural Motifs in Clusters Generated from Dendrogram (Figure 5)

Group	Chemical Code	Ratio +/-	Leading Structural Motifs
A	A034, N08, A098, A026, A076, A108, A124	**6/1**	
B	N09, N35, N39, A069, A111, A012, A008	**4/3**	
C	A090, A119, A105, A002, A107, A091, A058, A079	**8/0**	**$C_{[7-16]}$-(O-Ph)-(SO_3)**
D	A067, A100, A036, A024, A042, A039, A023, A010, A022, A025, N10, A020, A077, A099, A015, A104, A073, A018, A086, A076, A035, A019, A017, A032, A131, A081	**25/1**	
E	A009, A006, A110, A089, A088, A087, A083, A078	**8/0**	
F	A121, A041, A003, A072, A051, A118, A011, A066, A030, A068, A016	**11/0**	
G	A129, N33, A106, A013, A096, A064, A122, A115	**7/1**	
H	A112, N20, A094, N15, A127, N23, A031, A109, N19, N03, N31	**5/6**	
I	A063, A082, A114, A027, A004, N12, A103, A060, A055, A038, A005, A101, A084, A126, A052, A085, A033, A037, A028, N21, A047, A048, A044, A053, A102, A097, A057, A116, A092, A095, A113, A061, A059, A128, A045, A029, N29, A074, A130, A014, A070, A046, A001, A093, A054, A043, A125, N13	**44/4**	
J	N07, N22, A117, N06	**1/3**	**Low weight hydrocarbons, $C_{[4-6]}$-Br**
K	A080, N16, N38, A062, N30, N40, N24, N42, N02, N36, N04, A120, A071 ---------- A050, N44, N11, N17, N18, N43, N05, N47, N14, N46	**5/18**	
L	A056, N34, A040	**2/1**	**$C_{[16-18]}$-(SO_3CH_3//COOH)**
M	N27, N45, N32, N41, A065	**1/4**	**Polycyclic**
N	A123, A007, N01, A049, N25, N28, N37, N26	**3/5**	

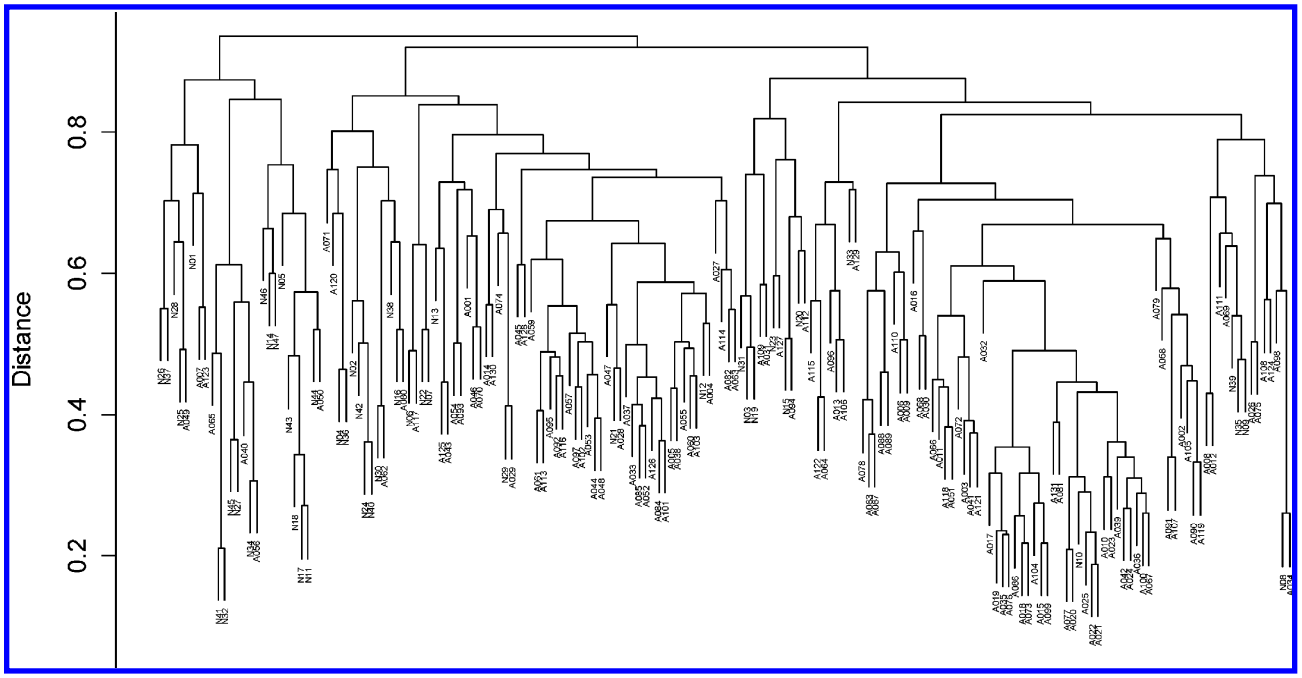


Figure 5. Hierarchical clustering dendrogram.

tion about the structure of the ICCVAM data set used in this study and the model itself. Random forest is usually considered as a "black box" approach where a user does not have much knowledge about the mechanism behind the predictions being made. However, the "black curtain" of the random forest can be partially raised by the analysis of proximity measure. In our case, one can notice that the random forest method tries to separate inactive chemicals from the active ones by forming groups or branches that are significantly enriched with either of the classes. What is more important, chemicals in most of these branch structures often share some level of molecular similarity: e.g. group A, which has four halogeno-derivatives of nonane, or group D, which has several alkyl halides with a long aliphatic chain ranging from twelve to sixteen carbon atoms. This separation of chemicals into structurally similar groups shows a future potential for the application of the random forest method. Furthermore, it also shows the great diversity and complexity of the ICCVAM data set and thus indicates some potential trouble spots such as small subgroups of structurally similar molecules that have opposite activities.

CONCLUSION

The random forest method showed improved prediction performance compared to a single classification tree. The proposed backward elimination method for selecting random forest variables produced remarkably predictive and yet simple models. Through the backward elimination process, random forest models with 10, 7, and 5 descriptors were quickly located from the data set of a larger number of input descriptors. This new approach enhanced the ability of the random forest algorithm to generate simple models and to find the important variables from a large set of variables. These models have a much better performance compared to models built by simply using the descriptors having the highest variable importance in the full model. Through adjusting the voting threshold, balanced performance of random forest models were obtained for the unbalanced data. The proximity measure given by the random forest model with 10 descriptors was used to measure the similarity between molecules and the resulting clustering gave rise to a reasonable partitioning of chemicals by formation of groups that, apart from expressing similar structural motifs, often have an increased ratio of active or inactive skin sensitizers. Presented results show future potential for the application of the random forest algorithm in QSAR studies of categorical responses.

REFERENCES AND NOTES

(1) Lushniak, B. D. The importance of occupational skin diseases in the United States. *Int. Arch. Occup. Environ. Health* **2003**, *76*, 325−330.
(2) *Handbook of occupational dermatology;* Elsner, L. P., Wahlberg, J. E., Maibach, H. I., Eds.; Springer-Verlag: New York, LLC: New York, 2000.
(3) Rodford, R.; Patlewicz, G.; Walker, J. D.; Payne, M. P. Quantitative structure−activity relationships for predicting skin and respiratory sensitization. *Environ. Toxicol. Chem.* **2003**, *22*, 1855−1861.
(4) Ashby, J.; Hilton, J.; Dearman, R. J.; Kimber, I. Streptozotocin: inherent but not expressed skin sensitizing activity. *Contact Dermatitis* **1995**, *33*, 165−167.
(5) Dupuis, G.; Benezra, C. *Allergic contact dermatitis to simple chemicals: a molecular approach;* Marcel Dekker Inc.: New York, 1982.
(6) Andersen, K. E. Occupational issues of allergic contact dermatitis. *Int. Arch. Occup. Environ. Health* **2003**, *76*, 347−350.
(7) Haneke, K. E.; Tice, R. R.; Carson, B. L.; Margolin, B. H.; Stokes, W. S. ICCVAM evaluation of the murine local lymph node assay. Data analyses completed by the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods. *Regul. Toxicol. Pharmacol.* **2001**, *34*, 274−286.
(8) *NIH Publication No. 99-4494: The Murine Local Lymph Node Assay: A test Method for Assessing the Allergic Contact Dermatitis Potential of Chemicals/Compounds;* NIH: 1999.
(9) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary quantitative structure−activity relationship (QSAR) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.
(10) Ren, S.; Schultz, T. W. Identifying the mechanism of aquatic toxicity of selected compounds by hydrophobicity and electrophilicity descriptors. *Toxicol. Lett.* **2002**, *129*, 151−160.
(11) Fedorowicz, A.; Zheng, L.; Singh, H.; Demchuk, E. A QSAR study of skin sensitization using Local Lymph Node Assay data. *Int. J. Mol. Sci.* **2004**, *5*, 55−66.
(12) Enslein, K.; Gombar, V. K.; Blake, B. W.; Maibach, H. I.; Hostynek, J. J.; Sigman, C. C.; Bagheri, D. A quantitative structure-toxicity relationships model for the dermal sensitization guinea pig maximization assay. *Food Chem. Toxicol.* **1997**, *35*, 1091−1098.

(13) Mosier, P. D.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. Predicting the genotoxicity of thiophene derivatives from molecular structure. *Chem. Res. Toxicol.* **2003**, *16*, 721−732.
(14) 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.
(15) Roberts, D. W.; Basketter, D. A. Quantitative structure−activity relationships: sulfonate esters in the local lymph node assay. *Contact Dermatitis* **2000**, *42*, 154−161.
(16) Patlewicz, G.; Basketter, D. A.; Smith, C. K.; Hotchkiss, S. A.; Roberts, D. W. Skin-sensitization structure−activity relationships for aldehydes. *Contact Dermatitis* **2001**, *44*, 331−336.
(17) Fedorowicz, A.; Singh, H.; Soderholm, S.; Demchuk, E. Structure−Activity Models for Contact Sensitization. *Chem. Res. Toxicol.* **2005**. In press.
(18) Breiman, L. RF/Tools: A class of two eyed algorithm. 2003, http://oz.berkeley.edu/users/breiman/siamtalk2003.pdf.
(19) Remlinger, K. Introduction and Application of Random Forest on High Throughput Screening Data from Drug Discovery. 2004, http://www4.ncsu.edu/∼ksremlin/Katja%20Remlinger%20Random-Forest.pdf.
(20) Guo, L.; Ma, Y.; Cukic, B.; Singh, H. Robust Prediction of Fault-Proneness by Random Forests. *Proceedings of the 15th International Symposium on Software Reliability Engineering* 2004, 417−428.
(21) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(22) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(23) *R*; *the R Development Core Team*, http://www.r-project.org.
(24) Basketter, D. A.; Roberts, D. W.; Cronin, M.; Scholes, E. W. The value of the local lymph node assay in quantitative structure−activity investigations. *Contact Dermatitis* **1992**, *27*, 137−142.

CI050049U