Journal of Computer-Aided Molecular Design **17:** 811–823, 2003.

In silico models for the prediction of dose-dependent human hepatotoxicity

Ailan Cheng[a,b] & Steven L. Dixon[a,*]
[a]*ADMET R&D, Accelrys, CN5375, Princeton, NJ 08543-5375, USA;* [b]*Present address: 548 Westgate Dr., State College, PA 16803, USA; E-mail: cheng189@adelphia.net*

Received 3 June 2003; accepted in revised form 15 November 2003

Key words: 2D descriptors, classification, data mining, ensemble recursive partitioning, hepatotoxicity, liver, variable selection

Summary

The liver is extremely vulnerable to the effects of xenobiotics due to its critical role in metabolism. Drug-induced hepatotoxicity may involve any number of different liver injuries, some of which lead to organ failure and, ultimately, patient death. Understandably, liver toxicity is one of the most important dose-limiting considerations in the drug development cycle, yet there remains a serious shortage of methods to predict hepatotoxicity from chemical structure. We discuss our latest findings in this area and present a new, fully general *in silico* model which is able to predict the occurrence of dose-dependent human hepatotoxicity with greater than 80% accuracy. Utilizing an *ensemble recursive partitioning* approach, the model classifies compounds as toxic or non-toxic and provides a confidence level to indicate which predictions are most likely to be correct. Only 2D structural information is required and predictions can be made quite rapidly, so this approach is entirely appropriate for data mining applications and for profiling large synthetic and/or virtual libraries.

Introduction

The liver is the first organ that comes into contact with most of the products of digestion, and as such, it is very vulnerable to the effects of xenobiotics, such as drugs and environmental chemicals. This propensity for injury is directly related to the liver's central role in metabolism and disposition of nutrients and foreign agents. Xenobiotics and their metabolites are often concentrated in the liver, and occasionally these concentrations reach toxic levels, causing various types of acute and chronic hepatocellular injuries (steatosis, necrosis, cirrhosis), cholestatic injuries, neoplasia, and elevated levels of hepatobiliary enzymes (aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase) [1, 2].

*To whom correspondence should be addressed.
Present address: Schrödinger, 120 W. 45th Street, 32nd Floor, New York, NY 10036, USA. Fax: +1-646-366-9550; E-mail: dixon @schrodinger.com.

Hepatotoxicity continues to be one of the costliest and most dangerous forms of toxicity encountered in the drug development cycle. Evidence of liver injury has caused the discontinuation of countless clinical studies, rejection of new drug applications and even the withdrawal of marketed drugs. Drug-induced hepatotoxicity is now the most common cause of acute liver failure, which often leads to patient death or the immediate need for a liver transplant [3].

Hepatotoxic compounds can be divided into intrinsic toxins and those causing idiosyncratic adverse reactions [1, 2]. An intrinsic toxin causes injury either by way of the original parent compound or indirectly by its toxic metabolites. The incidence rate is high and reproducible and the injury is dose-related. In contrast, idiosyncratic toxins are tolerated by a vast majority of the population, causing injury in only a very small fraction of subjects, typically less than 1/1000. However rare, injuries can be very severe and are often fatal. Idiosyncratic toxic potential cannot be predicted by pre-clinical animal studies and it is

neither dose-related nor generally reproducible. Individual genetic variations are believed to be the most important cause for idiosyncratic susceptibility. Other factors, such as age, gender, nutrition, pre-existing disease, environmental factors and exposure to other foreign compounds can also play a role.

Liver toxicity is clearly a major concern in both pharmaceutical research and environmental sciences, and there is no question as to the value of a model that could provide accurate *in silico* predictions of hepatotoxicity. Commercial software is available for the prediction of a host of other toxic endpoints (e.g., mutagenicity, carcinogenicity, developmental and reproductive toxicity, skin and eye irritancy, etc.). However, prediction of organ toxicity is still in its infancy, primarily due to the complex nature of toxic mechanisms and the lack of *in vivo* and *in vitro* data. Yet, there is a great need for such models because more therapeutic candidates fail due to organ toxicity than due to other toxicities for which predictive methodologies exist.

Developing models of idiosyncratic hepatotoxicity based on chemical structure may not be a realistic goal since these types of liver injuries are often linked to individual genetic polymorphisms and are not readily identified with particular structural features in toxic compounds. Genomic technologies may ultimately provide the most appropriate solution, possibly by identifying genetic profiles of susceptible populations.

Intrinsic hepatotoxicity, by contrast, is often linked to the presence of certain structural motifs, and thus a model based on chemical structure seems feasible. We have developed *in silico* models for the prediction of dose-dependent hepatotoxicity potential in humans. This includes compounds exhibiting either direct or indirect intrinsic toxicity. We are exploring novel, robust methods of prediction that do not attempt to subdivide compounds into different training sets based on common structural features. We are also predicting all manifestations of hepatotoxicity simultaneously, primarily because so many compounds cause multiple injuries. With appropriate data, however, it may be possible at some point to develop individual models for different types of liver injuries.

What follows are the details of the model development and validation procedures. We summarize how the hepatoxicity data were collected from the literature, we provide an explanation of the methodology behind the model, and we present various validation results, which indicate that the incidence of hepatotoxicity may be predicted with an accuracy of greater than 80%.

Data collection

Various books [1, 2, 4–6], public compilations [7–11] and journal articles (see, e.g., [12–22]) were consulted to identify compounds that are either known to cause human dose-dependent intrinsic hepatotoxicity or are generally accepted as safe. Whenever possible, the original literature sources cited in the compilations were reviewed to ensure the accuracy of the information. In addition, extensive cross-referencing among the different sources was done to eliminate any candidates for which inconsistent information was reported.

This process yielded a toxic or 'positive' data set of 149 compounds that cause dose-dependent hepatocellular, cholestatic, neoplastic and other liver injuries, or that trigger dose-related elevated aminotransferase levels in more than 10% of the population, even in the absence of overt hepatic injuries. Note that compounds reported to cause idiosyncratic injuries were included as long as there was also unequivocal evidence of dose-dependent behavior.

A non-toxic or 'negative' data set was collected under the requirement that the compounds show no evidence of causing liver injury in either humans or other animals. Even rare idiosyncratic adverse reactions and rare abnormal liver enzyme elevation in humans served as grounds for exclusion. This procedure yielded a total of 233 negatives. While there are undoubtedly many thousands of natural and synthetic compounds that do not cause liver toxicity, finding confirmation of this fact for pure substances (as opposed to mixtures and extracts) is not always a trivial matter. Moreover, inconsistent scales are frequently applied which can lead to unreasonable conclusions about toxicity. For example, ethanol is often cited as a textbook case for liver toxicity, but effects are normally only seen in patients who consume on the order of 100 grams or more on a daily basis for many years. This is far from the standard used to determine the toxicity or safety of other substances.

In assembling this training set, every effort was made to build in diversity by including compounds causing all types of liver injuries and compounds spanning a wide range of chemical families. The resulting set of 382 includes drug and drug-like compounds of various therapeutic classes, environmental chemicals, nutrients, dietary supplements, flavorings, preservatives, emulsifiers and a variety of other natural and

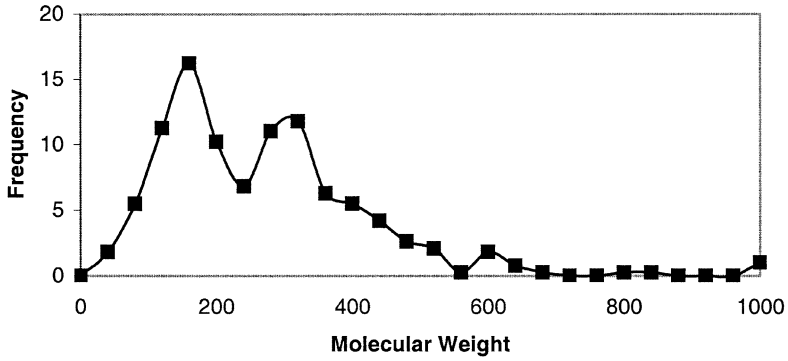

Figure 1. Molecular weight distribution for the 382 compounds in the training set.

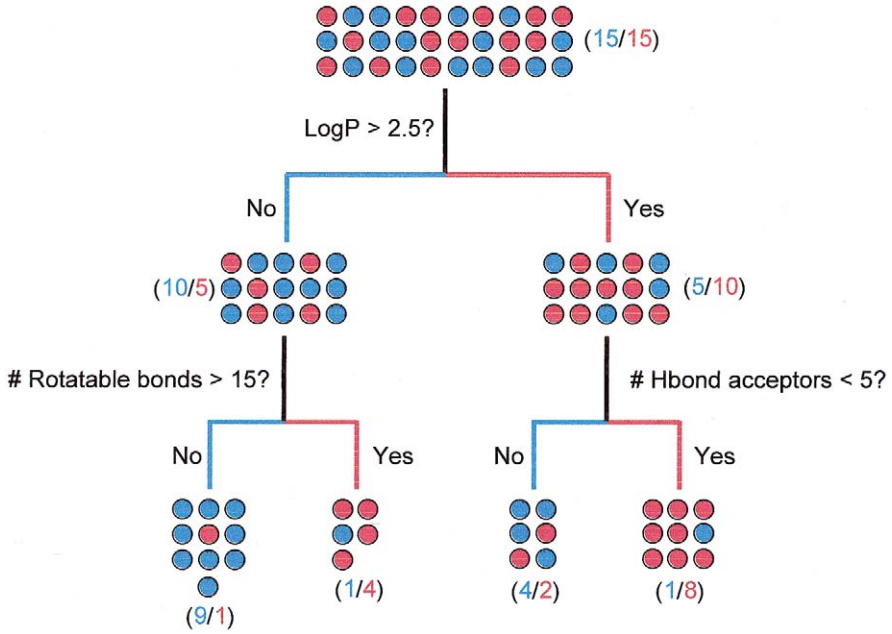

Figure 2. A decision tree that splits a training set of 30 compounds into four groups (leaves) that are purer in composition than the overall training set.

synthetic compounds. Figure 1 contains the molecular weight distribution of this training set.

With regard to the endpoint and an appropriate *in silico* model, it should be noted that consistent quantitative information on liver toxicity is not always readily available. Maximum tolerated dose and dose-limiting dose are reported for some compounds, but the actual endpoint value is dependent upon the route of administration, treatment regimen, gender and age of the subjects. Since many of the original references are case-study reports, the clinical conditions vary from one investigation to another. Moreover, it is frequently the case that only qualitative information is available. Therefore, development of a model that attempts to predict hepatotoxicity on a continuous scale is not really feasible. It is more realistic to focus on answering a simple yes/no question about dose-dependent hepatotoxicity. As detailed below, we have adopted this sort of approach in our *in silico* models, with the additional feature of a confidence level on each active versus inactive prediction.

In silico *models*

After deciding on the nature of the endpoint prediction (continuous vs. categorical), the first issue that must be addressed is whether or not to subdivide the training set into fairly homologous groups of compounds and

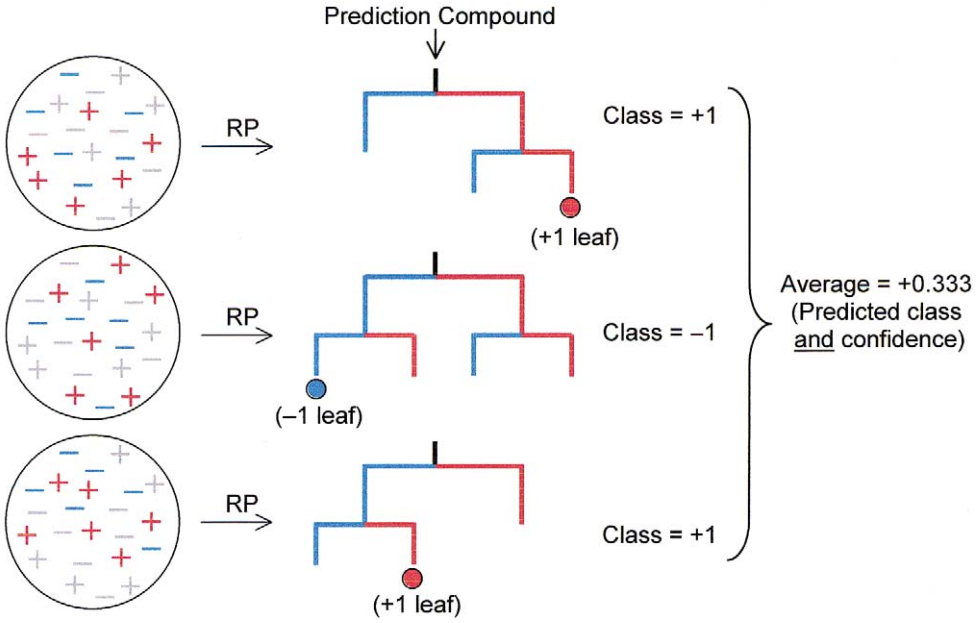

Figure 3. Ensemble approach to recursive partitioning. The red '+' and blue '−' characters indicate the members of the training set that are randomly selected and used to build each tree.

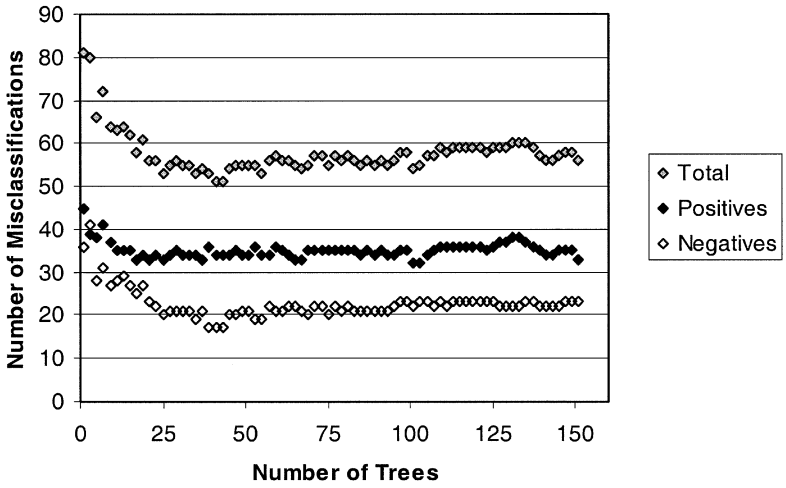

Figure 4. Ensemble LOO misclassification rates as a function of the number of trees.

develop individual models on these subsets. Because it is generally easier to derive a model from chemically related compounds, the use of subsetting is quite common in both quantitative and semi-quantitative models of toxicity and other pharmacological endpoints. However, while this procedure can provide better fits of the training set data, it can dramatically decrease the accuracy of predictions for new compounds, especially those that do not seem to fall clearly into any one particular model subset. Moreover, subdividing the training set increases the chances of over-fitting, simply because models are developed on smaller, less diverse sets of compounds, where chance correlations can play a much more significant role.

To avoid these pitfalls, we derive our model using all of the training set data together, despite the fact that many compounds have little or no structural similarity to others. This apparent difficulty can be overcome by using an *ensemble approach* [23, 24], wherein the overall model is actually an average of numerous models developed from random subsets of the training set. Combining the information from many different

training subsets automatically attenuates the effects of noisy data and over-fitting, and it results in a very general model that is applicable to extremely broad classes of compounds. An ensemble approach may be applied to any of various model-building techniques [23], but we choose recursive partitioning [25, 26] in the present case because it does not assume any particular functional relationship between the endpoint and the independent variables, and as such it is one of the most powerful available methods of classification.

Recursive partitioning (RP) involves the use of a decision tree, which poses a series of yes/no questions about the values of various independent variables in a manner that splits a training set into progressively smaller groups that have a higher 'purity' with respect to some property of interest. For hepatotoxicity, the goal is to break up the training set into groups of compounds that are either predominantly toxic or predominantly non-toxic. Figure 2 illustrates a hypothetical decision tree that might be built to purify a training set of 30 compounds falling into one of two categories (red, blue).

At each splitting point, or node, the RP algorithm searches a pool of independent variables (i.e., descriptors) and identifies a single variable and corresponding splitting value that best purifies the group of compounds entering the node. The splitting process continues until either no further improvement can be achieved, or the numbers of compounds in each purified group have reached some threshold below which further splitting is not statistically justified. At that point, the training set has been filtered into a final collection of terminal nodes or *leaves*. Each of these leaves is now identified with the class that predominates in the leaf population.

Once a decision tree is constructed from a training set, it can be applied to any other compound for which the necessary descriptors are known. The query compound is ultimately filtered into one of the leaves, and a prediction of its class (red/blue, toxic/non-toxic, etc.) is made according to the identity assigned to the leaf when the tree was built.

While quite powerful, RP can be readily abused. The method is very adept at finding descriptors and splitting criteria that purify a training set, but if conditions are not carefully controlled, the resulting decision tree can have very low predictivity when applied to new compounds. Some contributing factors include the use of descriptor pools that are too large, allowing leaves to become too small, and employing training sets that are overly biased toward one class. One other unsettling observation is that two training sets which differ by as little as one compound can yield totally different decision trees. This is because the addition or loss of a single observation may cause a different descriptor to be chosen at some early splitting point, which can completely change the structure of the tree from that node forward. Both trees may still exhibit high predictivity, but they might not share the same patterns of misclassifications.

The techniques we employ address all of these issues and provide predictions that not only indicate the class of a compound, but also a level of confidence about that prediction. Figure 3 illustrates the basic principles behind the ensemble RP approach. Each iteration of the model-building algorithm selects from the training set a random subset containing equal numbers of positives ($+1$) and negatives (-1), where, again, positives are hepatotoxic compounds. A tree is built using only those compounds and it is stored for later use. Once a sufficiently large number of trees has been built, they can be applied sequentially to any query compound, yielding a series of ± 1 predictions, depending on whether the query is filtered into a $+1$ leaf or a -1 leaf. If these ± 1 values are summed and divided by the number of trees, an average class is afforded. The algebraic sign indicates the category into which the compound should be placed (positive/negative), and the absolute value indicates the level of confidence.

The greatest advantage of this approach is that it automatically filters out noise and erroneous characteristics of models. There is no doubt that a given tree is partially invalidated by errors in the experimental data, failure of certain descriptors to accurately model the property being predicted, and incomplete coverage of chemical space by the training set. However, these factors tend to bias trees in random ways, so when an ensemble approach is employed, the biases effectively cancel each other out, leaving an average model that is largely free of defects due to statistical artifacts. This sort of model can be expected to have general applicability and high predictivity.

Our approach may be compared to that of *bagging* [27], but with the distinction that each sampled training set contains equal numbers of positives and negatives, and that no observation appears more than once in the sample. Bagging assigns all observations an equal probability of being selected, so the sampled populations will, on average, have the same distribution of positives and negatives as the overall training set. Bagging also utilizes sampling *with replacement*,

Table 1. Filtered pool of 25 structural descriptors.

Descriptor	Definition
Atype_C_17 Atype_C_18 Atype_C_30 Atype_N_78	AlogP atom type counts [29]
S_aaaC	Sum of E-State values for carbons forming 3 aromatic bonds [29]
V-DIST-mag	Vertex-based information index [29]
SIM1D_1 (Furosemide)* SIM1D_2 (Indicine-N-oxide)* SIM1D_3 (Tannic acid)* SIM1D_4 (Kepone)* SIM1D_5 (Carbon tetrachloride)* SIM1D_6 (2,2′-Methylene-bis(4-chlorophenol)* SIM1D_7 (Flavaspidic acid)* SIM1D_8 (Ethyl carbamate)* SIM1D_9 (2378-TCDD)* SIM1D_10 (Seneciophylline)* SIM1D_11 (Carbamazepine)* SIM1D_12 (Lorazepam) SIM1D_13 (Anagrelide) SIM1D_14 (Clonidine) SIM1D_15 (Proline) SIM1D_16 (Acetaldehyde) SIM1D_17 (Ethyl acetate) SIM1D_18 (Isopentyl acetate) SIM1D_19 (Isopentyl butyrate)	1D similarity to specific training set compounds [30]

*Hepatotoxic compound.

so a given training subset may have replicate observations.

Boosting [28] is another related technique, although it is considerably more complex than either bagging or the current ensemble approach. Boosting creates a sequence of sampled training sets whose compositions are adaptively modified so that poorly predicted observations are more likely to be selected with each subsequent training set. In doing so, a series of *weak learners* is said to be transformed into *strong learners*.

Computational details

Initially, each compound was characterized by a set of several hundred descriptors computed from 2D structural information alone. These included Cerius^2 topological, electrotopological and physicochemical parameters, various fragment keys based on AlogP atom types, [29] and 1D molecular similarity scores [30] computed against each member of the training set. As noted previously, RP is prone to generating decision trees with little or no predictivity if the variable pool is too large, so this rather enormous collection of descriptors was pared down to a manageable set of 25 using an in-house Monte Carlo linear regression algorithm applied to a ± 1 hepatotoxicity dependent variable.

In this approach, an initial set of 25 descriptors was chosen using simple forward stepwise selection, a procedure which involved identification of the best 1-variable model, followed by sequential addition of 24 descriptors, with no deletions. In other words, the n^{th} descriptor added was the one that yielded the best n-variable fit, given that the previous $n-1$ descriptors remained in the model. This provided a starting point for the Monte Carlo simulated annealing algorithm.

Table 2. Summary of leave-one-out (LOO) predictions.

RP method	Correct classifications		
	Positives	Negatives	All
Ensemble	116/149 (78%)	210/233 (90%)	326/382 (85%)
Standard	84/149 (56%)	169/233 (73%)	253/382 (66%)

Table 3. Summary of leave-10%-out predictions.

RP method	Correct classifications		
	Positives	Negatives	All
Ensemble	114/149 (76%)	175/233 (75%)	289/382 (76%)
Standard	89/149 (60%)	167/233 (72%)	256/380 (67%)

Here, individual descriptors in an evolving 25-variable model were selected at random and flagged for possible replacement by a descriptor in the unused portion of the large pool. The candidate chosen from the pool was the descriptor that yielded the best 25-variable fit when combined with the other 24 descriptors still in the model. So whereas a random scheme was used to select the descriptor for ejection, a greedy scheme was used to identify its possible replacement. If the new model produced a lower standard deviation of regression, the replacement was made; if not, the Monte Carlo test was applied. The temperature was reduced over the course of 1000 steps in a log-linear fashion. Upper and lower limits were set at $0.5\sigma_y$ and $0.05\sigma_y$, respectively, where σ_y was the observed standard deviation in the ± 1 hepatotoxicity variable.

Ultimately, a set of 25 linearly independent descriptors is identified, with each descriptor exhibiting a mild statistical relationship to hepatotoxicity. One may argue that since models will ultimately be constructed using RP, applying a linear regression filter to the descriptor pool is not an optimal approach. However, because RP is so adept at finding relationships among data, be they robust or spurious relationships, we cannot overstress the importance of pre-filtering the variable pool using a non-RP technique.

The filtered pool of 25 descriptors was submitted to both the ensemble RP approach and the standard, single-tree approach. All decision trees were generated with the aid of the S-PLUS statistical package [31], using default settings unless otherwise noted. No pruning was performed, but, as discussed below, node sizes were controlled according to the results of leave-one-out experiments.

In the ensemble tree-generation procedure, subsets containing 135 positives and 135 negatives were selected randomly from the overall training set. A total of 151 subsets were selected this way, yielding 151 different trees. Each of these trees typically utilizes 8–12 of the 25 descriptors, and while some trees are observed to share characteristics with others, no single pattern is generally discernable. An odd number of trees was employed so that the average class would never be exactly 0.0. We determined that 151 was also sufficient to yield satisfactory convergence in the average class values.

Three sets of experiments were carried out to validate the methodology and models. The first two experiments involved the leave-n-out technique, wherein n observations were removed from the 382-member training set and the hepatotoxicities of these n compounds were predicted using models built from the remaining $382-n$ compounds. This process was repeated for distinct, non-overlapping sets of n, until each compound was predicted once. We investigated both n=1 and n=10% (38 or 39), using two different conditions on descriptor pool selection. In the leave-one-out case, a single pool of 25 descriptors was selected as described above, using the Monte Carlo regression technique applied to the entire training set of 382. Thus each compound being held out for prediction was in fact present when the descriptor pool was selected and, therefore, had some impact on the composition of the pool. By contrast, the leave-10%-out experiments involved selecting a new set of 25 descriptors for each 90% training set (344 or 343 compounds), ignoring all information from the 38 or 39 compounds being withheld for prediction.

Note that the leave-n-out tests represent *internal* validation, and that an ensemble or standard RP model built to predict a given set of n compounds will generally differ from a corresponding model built to predict a different set of n compounds. These tests, therefore, do not validate any single hepatotoxicity model *per se*, rather, they suggest whether the overall technique can produce hepatotoxicity models with predictive ability. We note that internal validation may also be used to arrive at appropriate settings for parameters that control model architecture and performance. For example, the minimum allowed node size was optimized by monitoring its effect on the accuracy of the leave-one-out predictions.

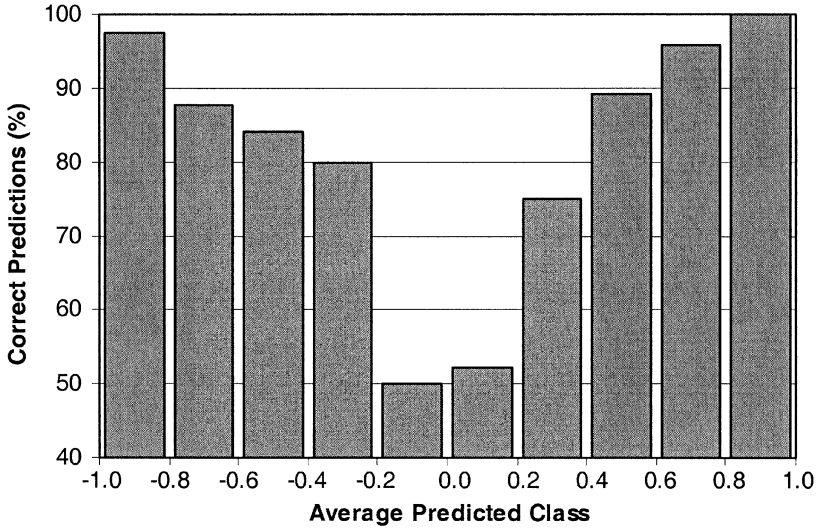

Figure 5. Accuracy of ensemble LOO class predictions (toxic versus nontoxic) as a function of average predicted class value on the interval [−1, 1].

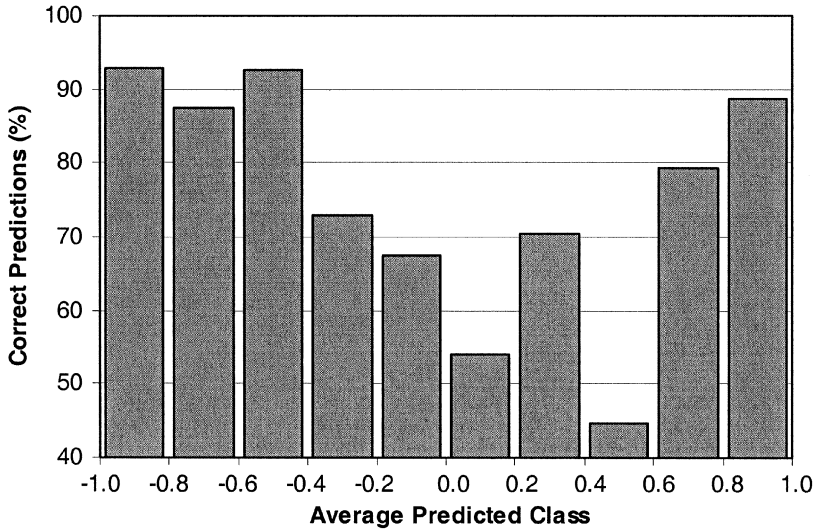

Figure 6. Accuracy of ensemble leave-10%-out predictions as a function of average predicted class.

While internal tests are instrumental in assessing the overall validity of the present methodology, such tests should be supplemented with *external* validation, wherein a final model is tested on a new set of compounds that were never considered in any of the decision making or model building steps. Toward this end, a separate set of 54 compounds (23 positives, 31 negatives) was assembled subsequent to development of the final hepatotoxicity models (standard and ensemble). This external validation set was collected from the same compilations as the training set, using the same criteria for inclusion/exclusion. These compounds hold no particular distinction from those included in the training set, except that the original literature sources were less readily available to us, so somewhat more work was required to verify the toxicity/safety of each compound.

Results and discussion

The Monte Carlo filtering process resulted in the pool of 25 descriptors summarized in Table 1. This set was used to generate all LOO results as well as the final

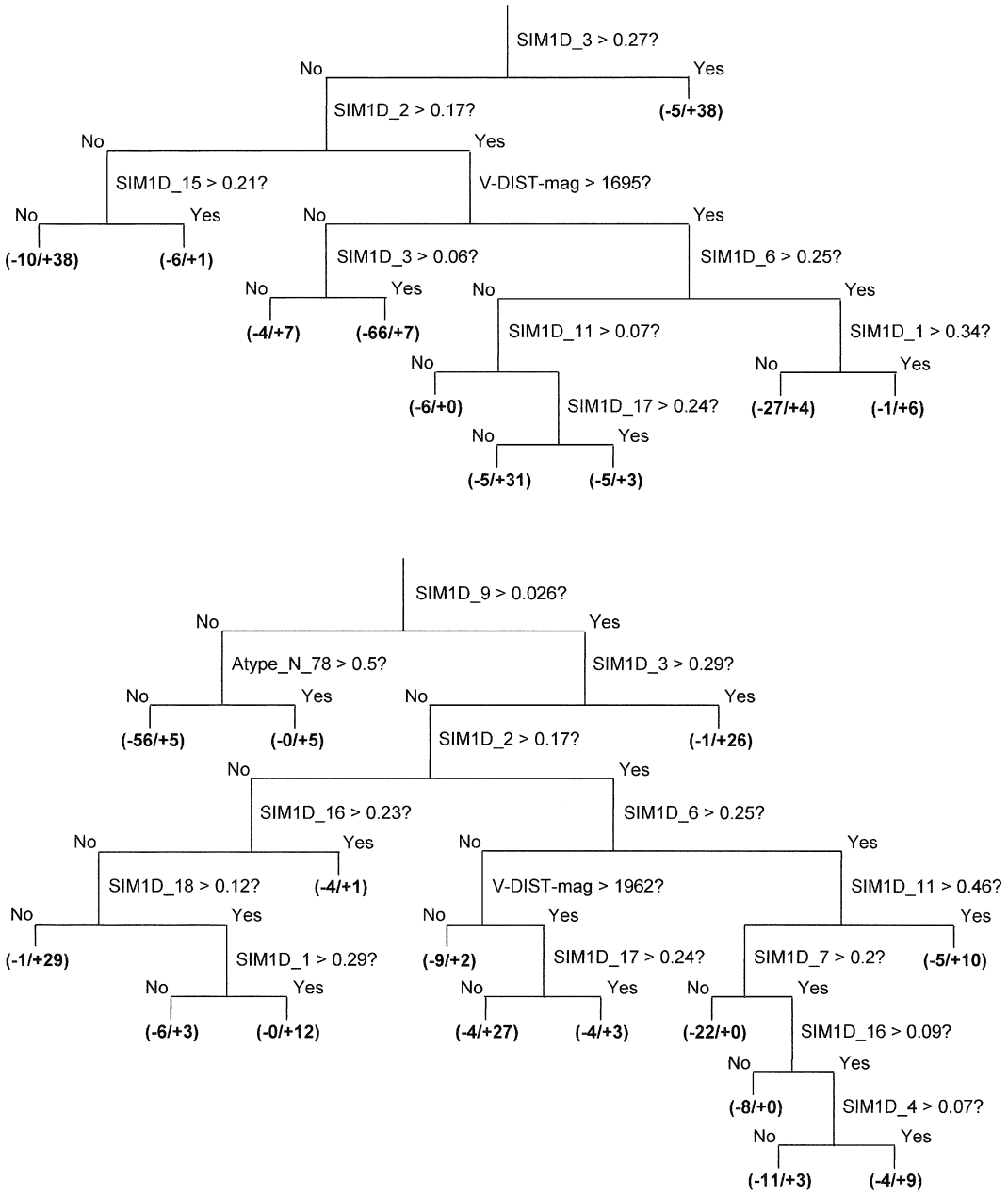

Figure 7. Example decision trees from the final ensemble model.

standard and ensemble models that were applied to the external validation compounds. Each leave-10%-out experiment involved a different pool of 25 descriptors, which we omit here for brevity. Note that 19 of the 25 descriptors are 1D similarities computed against specific training set compounds. Many of these similarities are probably encoding structural features that are critical determinants of hepatotoxicity. While one might expect that this list of 19 would consist only of hepatotoxic compounds, in fact 8 of them are known to be safe. Obviously, structural similarity to a safe compound cannot, in and of itself, guarantee safety, so the decisions involving these compounds are not expected to be quite so cut-and-dry.

Systematic variation of the minimum allowed node size in the leave-one-out (LOO) experiments indicated

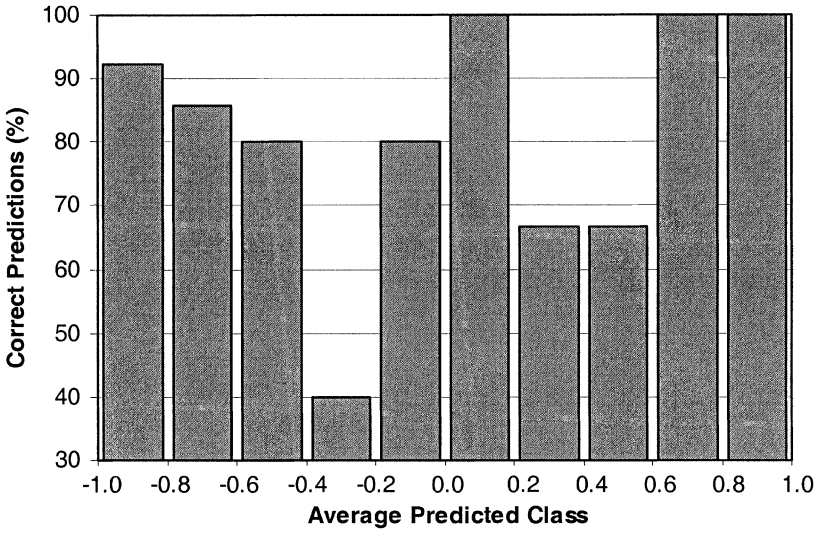

Figure 8. Accuracy of ensemble external predictions as a function of average predicted class.

that near optimal predictivity occurred when nodes with fewer than 20 compounds were not subjected to further splitting (S-PLUS: MINSIZE = 20). In fact, we observed that a range of values around this limit afforded essentially the same quality of LOO predictions, whether using the ensemble or standard RP approach. Consequently, there was no compelling reason to pick any particular value other than 20 for all subsequent work.

Table 2 summarizes the results of the LOO validations using both the ensemble and standard RP approaches. Ensemble predictions are seen to be considerably more accurate for both positives (78% vs. 56%) and negatives (90% vs. 73%). While neither method is able to identify positives as consistently as negatives, the ensemble approach does reduce the relative disparity between the two accuracy rates. This phenomenon has been observed previously [23] and it is partially due to the fact that the ensemble training subsets contain equal numbers of positives and negatives. Thus the natural bias favoring the majority class (i.e., negative predictions) is reduced.

Figure 4 illustrates the behavior of the ensemble LOO predictions as the number of trees is increased from 1 to 151. A general trend toward lower misclassification rates is observed as more trees are incorporated into the model, suggesting that the ensemble approach is in fact attenuating the effects of noise and erroneous models. Ultimately, a flattening of the curves occurs, indicating that the predictions, and hence the underlying model, have essentially converged. At this point, misclassification counts change by no more than one or two, and these fluctuations are due not to significant changes in the model, but rather to alternation of algebraic signs for certain average class values which are very close to zero.

One of the greatest advantages of the ensemble approach is that the average class can be used as a measure of confidence in the prediction. Figure 5 contains a histogram that illustrates LOO prediction accuracy as a function of average class. A classic u-shaped pattern results, indicating higher rates of accuracy (>95%) as the average class value tends towards the limits +1 and −1. Predictions near the middle are less accurate, and this is an extremely valuable piece of information for anyone using such a model. The ensemble method can clearly be used to prioritize compounds both in terms of their predicted hepatotoxicity, and with respect to the confidence in these predictions.

LOO validations provide a valuable assessment of the methodology, but it must be remembered that each compound being predicted was included when the descriptor pool (Table 1) was selected, so the potential for bias exists. The leave-10%-out validations, however, involve no such bias, as each 10% group withheld for prediction was also excluded from the descriptor selection process. Table 3 contains the combined classification results from these 10-fold validation experiments. Overall, the ensemble method still outperforms the standard, single-tree approach, but a fairly substantial drop is observed in the number of correct negative classifications achieved by the

Table 4. Descriptor statistics from 151 decision trees used in the final ensemble model.

Descriptor	Number of decisions[a]	Negative leaves		Positive leaves	
		Population[b]	Distribution[c]	Population[d]	Distribution[e]
Atype_C_17	17	53	−0.906 +0.094	99	−0.131 +0.869
Atype_C_18	19	35	−0.629 +0.371	148	−0.041 +0.959
Atype_C_30	1	0	−0.000 +0.000	5	−0.000 +1.000
Atype_N_78	108	0	−0.000 +0.000	607	−0.000 +1.000
S_aaaC	12	16	−0.562 +0.438	135	−0.030 +0.970
V-DIST-mag	175	1504	−0.801 +0.199	970	−0.138 +0.862
SIM1D_1	120	878	−0.900 +0.100	1087	−0.109 +0.891
SIM1D_2	145	783	−0.811 +0.189	982	−0.094 +0.906
SIM1D_3	218	1084	−0.863 +0.137	1498	−0.136 +0.864
SIM1D_4	114	730	−0.856 +0.144	1191	−0.122 +0.878
SIM1D_5	45	276	−0.877 +0.123	591	−0.032 +0.968
SIM1D_6	136	902	−0.876 +0.124	770	−0.175 +0.825
SIM1D_7	142	1530	−0.899 +0.101	935	−0.102 +0.898
SIM1D_8	59	358	−0.846 +0.154	414	−0.174 +0.826
SIM1D_9	187	739	−0.846 +0.154	703	−0.175 +0.825
SIM1D_10	191	1871	−0.925 +0.075	1429	−0.077 +0.923
SIM1D_11	119	656	−0.825 +0.175	926	−0.257 +0.743
SIM1D_12	136	983	−0.864 +0.136	1414	−0.084 +0.916
SIM1D_13	88	543	−0.880 +0.120	696	−0.108 +0.892
SIM1D_14	138	696	−0.874 +0.126	723	−0.142 +0.858
SIM1D_15	212	1308	−0.906 +0.094	1069	−0.175 +0.825
SIM1D_16	213	1729	−0.909 +0.091	1124	−0.151 +0.849
SIM1D_17	140	899	−0.867 +0.133	1286	−0.096 +0.904
SIM1D_18	107	1986	−0.937 +0.063	634	−0.191 +0.809
SIM1D_19	166	712	−0.860 +0.140	1063	−0.149 +0.851

[a]Total number of decision tree nodes involving the descriptor.
[b]Total number of training set compounds classified into negative leaves predicated on the descriptor.
[c]Population distribution of negative and positive compounds within negative leaves predicated on the descriptor.
[d]Total number of training set compounds classified into positive leaves predicated on the descriptor.
[e]Population distribution of negative and positive compounds within positive leaves predicated on the descriptor.

ensemble model. This is no doubt due in part to an overall reduction in the chemical space covered by each pool of training set compounds. As shown in Figure 6, ensemble prediction accuracy still increases as the average class tends toward the extremes of ± 1, although there is an anomalous decrease in accuracy for compounds whose predicted classes lie between $+0.4$ and $+0.6$. Since only positive compounds should be predicted to lie in this interval, misclassifications here are associated with negative compounds. These incorrect predictions largely account for the aforementioned reduction in overall accuracy with respect to classifying negatives.

All results seen thus far were derived from leave-*n*-out experiments, so discussions have not focused on any particular standard or ensemble model. It is, however, necessary to construct final models using a single pool of training compounds and a single set of descriptors, then validate against a new set of compounds that were never considered during any phase of model development. Accordingly, standard and ensemble models were created using all 382 training set compounds (either directly or by way of $+135/-135$ sampling), and the 25 descriptors in Table 1.

Figure 7 summarizes two of the 151 decision trees that contribute to the final ensemble model. Nearly identical splitting criteria are observed at a few nodes, so the trees obviously share some characteristics. The overall structures are generally quite different, though, and the first tree is considerably less complex than the

second, with 9 decision nodes in the former and 14 in the latter.

One somewhat counterintuitive observation is that many of the 1D similarity splitting values are extremely low, for example, 0.06. Typically, even pharmacologically *unrelated* compounds exhibit similarities in the neighborhood of 0.3 [30], so it would appear that hepatotoxicity is not something that can be readily deduced on the basis of overall structural similarity. This observation is consistent with the notion that a common, possibly small, chemical moiety can be linked to metabolic toxification [1] of very different compounds.

Table 4 contains various descriptor statistics observed across the 151 decision trees. The number of times each descriptor was utilized varies quite dramatically, with Atype_C_30 being involved in only a single decision, while SIM1D_3 was used in 218 decisions. The 'Population' columns indicate the number of training set compounds that were classified into negative or positive leaves based on decisions involving a particular descriptor. 'Distribution' columns indicate the average purity of those leaves. For example, Atype_C_17 was used to classify 53 training set compounds into negative leaves, and 90.6% of these compounds were in fact negative. Observe that because the statistics are pooled from overlapping training sets, these 53 classifications are expected to involve far fewer than 53 *distinct* compounds.

It is worth noting that atom type count Atype_N_78 showed perfect accuracy in classifying hepatotoxic compounds in a large number of trees. Since Atype_N_78 typically flags the presence of an N-nitroso group, this result comes as no surprise. Similarity to the hepatotoxic compound seneciophylline (SIM1D_10), while not a perfect classifier, is arguably one of the more versatile descriptors, with 1871 negative classifications, 1429 positive classifications, and very high purities ($\geq$0.923) in both types of leaves.

Finally, we present the results of external validation, wherein the final models were applied to a set of 54 compounds collected after the models were created. As shown in Table 5, the ensemble method once again outperforms the standard single-tree approach, classifying more than 80% of compounds correctly, compared to only 69% for standard RP. A clear disparity is observed between the two models with regard to identification of positives, as standard RP misclassifies the *majority* of hepatotoxic compounds. This implies a serious deficiency in the standard model, and underscores the possible risk in using a single decision tree to make predictions in a diverse data space.

Table 5. Summary of external predictions using final models

RP method	Correct classifications		
	Positives	Negatives	All
Ensemble	16/23 (70%)	28/31 (90%)	44/54 (81%)
Standard	10/23 (43%)	27/31 (87%)	37/54 (69%)

Figure 8 indicates that ensemble class values near the $\pm$1 limits still coincide with the highest levels of prediction accuracy, although high accuracy is also observed around the center of the class interval. These results, as well as those from previous validation tests, suggest that the ensemble approach is correct about 90% of the time for predictions beyond $\pm$0.6. While no *in silico* model should be used as an absolute indicator of which compounds to abandon on the basis of toxicity, the present model appears to have significant potential as a tool for hepatotoxicity risk assessment.

Conclusions

Organ toxicity poses one of the greatest challenges to the drug development cycle, and there continues to be a serious lack of reliable computer-based methods to predict these complex biological endpoints. The liver is particularly vulnerable to drug-induced toxicity, and it is considered by many to be the system in greatest need of predictive methodologies. Toward this end, a new *in silico* model has been developed to accurately predict the occurrence of dose-dependent human hepatotoxicity. Both direct and indirect toxins have been considered, along with all manifestations of liver toxicity. Using an ensemble approach to recursive partitioning, a single diverse training set has been employed to create a series of decision trees whose combined predictions are more than 80% accurate in answering a yes/no question about hepatotoxicity. Moreover, a confidence level is provided with each prediction, and it is shown to be a reliable indicator of the most accurate yes/no classifications.

The ensemble approach has been observed here and elsewhere [23, 24] to outperform the corresponding single classifier method from which it was derived. This new generation of modeling techniques holds great promise in applications involving struc-

turally diverse compound collections, experimental data obtained from disparate sources, and biological endpoints associated with complex systems.

Acknowledgements

We wish to thank Marvin Waldman for many insightful discussions about statistical sampling approaches and chemical structure characterization. We would also like to acknowledge the significant contributions made by Roberta Susnow toward development of the RP models.

References

1. Farrell, G.C. Drug-Induced Liver Disease. Churchill Livingstone, New York, 1994.
2. Zimmerman, H.J. Hepatotoxicity: The Adverse Effects of Drugs and Other Chemicals on the Liver. Lippincott Williams & Wilkins, Philadelphia, PA, 1999.
3. AASLD/FDA/PhRMA Hepatotoxicity Workshop, Chantilly, VA, 2001.
4. Source Book of Flavors. Chapman & Hall, New York, 1994.
5. Coultate, T.P. Food: the Chemistry of its Components. The Royal Society of Chemistry, Cambridge, UK, 1996.
6. Stricker, B.H.C. Drug-Induced Hepatic Injury. Elsevier, Amsterdam, 1992.
7. Physicians' Desk Reference: Electronic Library. Thomson Micromedex, Inc., Greenwood Village, CO, 2001.
8. Registry of Toxic Effects of Chemical Substances (RTECS), The National Institute for Occupational Safety and Health, Cincinnati, OH.
9. Toxicology Data Network (TOXNET), U.S. National Library of Medicine, Bethesda, MD.
10. FDA Generally Recognized as Safe (GRAS) List, U.S. Food and Drug Administration, Rockville, MD.
11. Fenaroli's Handbook of Flavor Ingredients. CRC Press, Boca Raton, FL, 1995.
12. Gold, E.J., Mertelsmann, R.H., Itri, L.M., Gee, T., Arlin, Z., Kempin, S., Clarkson, B. and Moore, M.A., Cancer Treatment Reports, 67 (1983) 981.
13. Guzzo, C., Benik, K., Lazarus, G., Johnson, J. and Weinstein, G., Arch. Dermatol., 127 (1991) 511.
14. Knip, M., Douek, I.F., Moore, W.P.T., Gillmor, H.A., McLean, A.E.M., Bingley, P.J. and Gale, E.A.M., Diabetologia, 43 (2000) 1337.
15. Lagadic-Gossmann, D., Rissel, M., Le Bot, M.A. and Guillouzo, A., Cell Biol. Toxicol., 14 (1998) 361.
16. Lakhanpal, S., Donehower, R.C. and Rowinsky, E.K., Invest. New Drugs, 19 (2001) 69.
17. O'Brien, J.T., Eagger, S. and Levy, R., Age and Ageing, 20 (1991) 129.
18. Raber, M.N., Newman, R.A., Newman, B.M., Gaver, R.C. and Schacter, L.P., Cancer Res., 52 (1992) 1406.
19. Rowinsky, E.K., Noe, D.A., Ettinger, D.S., Christian, M.C., Lubejko, B.G., Fishman, E.K., Sartorius, S.E., Boyd, M.R. and Donehower, R.C., Cancer Res., 53 (1993) 1794.
20. Ryan, D.P., Supko, J.G., Eder, J.P., Seiden, M.V., Demetri, G., Lynch, T.J., Fischman, A.J., Davis, J., Jimeno, J. and Clark J.W., Clin. Cancer Res., 7 (2001) 231.
21. Von Mehren, M., Giantonio, B.J., McAleer, C., Schilder, R., McPhillips, J. and O'Dwyer, P.J., Invest. New Drugs, 13 (1995) 205.
22. Vogel, C.L., Gorowski, E., Davila, E., Eisenberger, M., Kosinski, J., Agarwal, R.P. and Savaraj, N., Invest. New Drugs, 5 (1987) 187.
23. Dixon, S.L. and Villar, H.O., J. Comput.-Aided Mol. Design, 13 (1999) 533.
24. Susnow, R.G. and Dixon, S.L., J. Chem. Inf. Comput. Sci., 43 (2003) 1308.
25. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. Classification and Regression Trees. Wadsworth International Group, Belmont, CA, 1984.
26. Hawkins, D.M. and Kass, G.V., Topics in Applied Multivariate Analysis, Hawkins, D.H., Ed., Cambridge University Press, Cambridge, UK, 1982, p. 269.
27. Breiman, L., Machine Learning, 24 (1996) 123.
28. Freund, Y., Information and Computation, 121 (1995) 256.
29. Cerius^2, Accelrys, San Diego, CA, 2001.
30. Dixon, S.L., Kenneth M. and Merz, J., J. Med. Chem., 44 (2001) 3795.
31. S-PLUS 6.1 for Windows, Insightful Corporation, Seattle, WA, 2001.