Symbolic, Neural, and Bayesian Machine Learning Models for Predicting Carcinogenicity of Chemical Compounds

Dennis Bahler* and Brian Stone

Artificial Intelligence Laboratory, Department of Computer Science, North Carolina State University, Raleigh, North Carolina 27695-8206

Carol Wellington

Department of Mathematics and Computer Science, Shippensburg University, Shippensburg, Pennsylvania 17257-2299

Douglas W. Bristol

National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709

Received September 9, 1999

Experimental programs have been underway for several years to determine the environmental effects of chemical compounds, mixtures, and the like. Among these programs is the National Toxicology Program (NTP) on rodent carcinogenicity. Because these experiments are costly and time-consuming, the rate at which test articles (i.e., chemicals) can be tested is limited. The ability to predict the outcome of the analysis at various points in the process would facilitate informed decisions about the allocation of testing resources. To assist human experts in organizing an empirical testing regime, and to try to shed light on mechanisms of toxicity, we constructed toxicity models using various machine learning and data mining methods, both existing and those of our own devising. These models took the form of decision trees, rule sets, neural networks, rules extracted from trained neural networks, and Bayesian classifiers. As a training set, we used recent results from rodent carcinogenicity bioassays conducted by the NTP on 226 test articles. We performed 10-way cross-validation on each of our models to approximate their expected error rates on unseen data. The data set consists of physical−chemical parameters of test articles, alerting chemical substructures, salmonella mutagenicity assay results, subchronic histopathology data, and information on route, strain, and sex/species for 744 individual experiments. These results contribute to the ongoing process of evaluating and interpreting the data collected from chemical toxicity studies.

1. INTRODUCTION TO THE PROBLEM

Determining which chemicals in the environment are carcinogenic to humans is of clear public benefit. The rodent bioassay program currently being performed by the U. S. National Toxicology Program (NTP)[19] to determine which chemicals cause cancer in rodents, while clearly important to human risk assessment and public policy,[31] is time consuming and expensive.

It is our hypothesis that inductive analysis of biological information by a variety of machine learning techniques will discover patterns, co-occurrences, and correlations that have not come to light using other techniques. Further, we believe that such inductive analysis could lead to development of information management systems that are useful, first, for assisting researchers in the task of predicting the presence or absence of carcinogenic effects of chemical compounds and, second, in developing mechanistic hypotheses that explain such effects. We are engaged in a long-term project to test our hypothesis.[4−11,36,38,39] Aside from testing machine predictions against the results of ongoing empirical studies, a major goal of the project is to provide predictions that can help guide the selection of chemicals for testing. In the longer term, this approach may also help reduce the use of laboratory animals in such testing. The past few years have seen an explosion of interest in the problem of predicting toxic activity of chemicals without animal testing. Interest has come from both human experts[3,18,37] and from developers of computer-based systems.[16,23,25,34]

* To whom correspondence should be addressed. Telephone: 919-515-3369. E-mail: bahler@ncsu.edu.

2. METHODS

2.1. Training Data. We gathered a set of 904 rodent bioassay experiments conducted on 226 test articles by the NTP and reported in NTP Technical Reports 201−458. Of these experiments, 160 were classified as equivocal evidence and were eliminated in most (but not all) of our experiments. Of the remaining 744 experiments, 276 were classified as no evidence (which we refer to as negative) and 468 were classified as either clear evidence or some evidence. These latter two classifications were combined into a positive class for purposes of training our models.

The information available on the experiments consisted of a set of 264 attributes. (Some attributes in the data set were used for bookkeeping and ignored during learning.) The attributes fell into several categories.

10.1021/ci990116i CCC: $19.00
Published on Web 07/24/2000

Physical Chemical Parameters (20 Attributes). These are properties of a test article, and can be determined either computationally or experimentally. The physical chemical parameters used were electronegativity (K_e) rate (both computed and experimentally determined); octanol−water partition coefficient (log p, computed by Ghose−Crippen; Dixon, and constituent-based methods), highest occupied and lowest unoccupied molecular orbits, molecular weight, pK_a, rectangular area (RA); planar area (PA), *Z*-depth (ZD), the ratio RA/ZD^2 the ratio PA/ZD^2 molecular hardness (computed by two methods, plus an indicator attribute if the two methods produced different results); molecular volume; molecular area; molecular ovality; and dipole moment.

Structural Alerts (21 Attributes). Human expertise has identified certain functional-group substructures of organic molecules that may predispose the parent molecule toward causing chemical mutagenesis and carcinogenesis, because they represent the potential for either entering into electrophilic reaction with DNA or being converted by metabolism into an electrophilic functionality that can react with DNA.[1] An attribute for each of the following structural alerts for DNA reactivity was included in our training set: alkyl esters of either phosphonic or sulfonic acids; aromatic nitro groups; aromatic azo groups; aromatic ring *N*-oxides; aromatic mono- and dialkylamino groups; alkyl hydrazines; alkyl aldehydes; *N*-methylol derivatives; monohaloalkenes; β-haloethyl *N* and *S* mustards; *N*-chloramines; propiolactones and propiosultones; aromatic and aliphatic aziridinyl derivatives; aromatic and aliphatic substituted primary alkyl halides; derivatives of urethane (carbamates); alkyl *N*-nitrosamines; aromatic amines, *N*-hydroxy derivatives and derived esters; aliphatic epoxides and aromatic oxides; Michael reactive centers; halogenated methanes; and aliphatic nitro groups.

Salmonella Mutagenesis Test Result (1 Attribute). The Ames test for mutagenesis, a short-term (typically 14-day) in vitro assay, has been performed on *Salmonella typhimurium* bacteria on most of the test articles in the training set.[2]

Subchronic Histopathology (209 Attributes). These were represented as pairs of organ site and morphology. A total of 38 organs exhibited pathologic change in at least one experiment, and 72 different morphologies were observed at least once.

Sex and Species Exposed (2 Attributes).

Route of Administration (1 Attribute). Routes are feed, water, skin painting, inhalation, and gavage.

Maximally Tolerated Dose (4 Attributes). Using molecular weight and standard conversion formulas, doses were normalized to micromoles per kilogram per day.

2.2. Cross-Validation of the Models. The models were each cross-validated using 10-way cross-validation. In this method, the training set was divided into 10 sets of data with both positive and negative experiments spread as equally as possible between the sets. Each of these sets in turn was set aside while a model was built using the other nine sets. This model was then used to classify the (unseen) test articles in the tenth set and the accuracy computed by comparing these predictions with the actual classes. This process was repeated 10 times, and the results were averaged. Similar procedures were used for all model-building methods.

2.3. Statistics. There are a number of statistics which are used to measure the ability of a model to predict a classification. **Sensitivity** is the percent of positive examples which were correctly classified, and **specificity** is the percent of negative examples which were correctly classified. These give measures of how well the model can predict test articles of each class. **Positive predictivity** is the percent of positive predictions which were correct. **Negative predictivity** is the percent of negative predictions which were correct. These give a measure of how likely it is that a prediction of a specific class is correct. **Accuracy** is the percent of total predictions which were correct. Correlation coefficients and levels of significance (Fisher's exact test) are also informative.

3. TREE AND RULE MODELS

Decision tree models were constructed by computer using a greedy, divide-and-conquer algorithm, which at each step has the goal of selecting from a set of attributes the one whose values best discriminate a set of examples according to the classification.

3.1. Tree Induction Methodology. TIPT (tree induction for predictive toxicology) is a similarity-based classification system based on the tree and rule induction system C4.5.[30] TIPT uses supervised learning over a training set of test article-specific and experiment-specific attributes to devise concept descriptions capable of successfully classifying unseen chemicals.

Each tree was constructed by a greedy, divide-and-conquer algorithm, which at each step has the goal of selecting from a set of attributes the one that best discriminates a set of examples according to the classification. The criterion for deciding which attribute to select was the information−gain ratio, which is computed as follows. Assume we have a set *T* of examples, each example belonging to one of *k* classes C_1, ..., C_k. In this application, *T* is a training set of chemicals and the C_i are the NTP carcinogenicity classifications. The information required to completely discriminate *T* into these classes is given by the entropy formula[33]

$$\mathrm{I}(T) = -\sum_{i=1}^{k}\left[\frac{|T_{C_i}|}{|T|}\log_2\left(\frac{|T_{C_i}|}{|T|}\right)\right]$$

where T_{C_i} is the subset of examples in *T* having classification C_i and $|S|$ is the number of elements in set *S*.

Now suppose *T* is partitioned according to the *n* possible values of some attribute *A*. The amount of information still required for complete discrimination is given by

$$\mathrm{I_A}(T) = \sum_{i=1}^{n}\left[\frac{|T_{(A\leftarrow i)}|}{|T|}\,\mathrm{I}(T_{(A\leftarrow i)})\right]$$

where $(T_{(A\leftarrow i)}$ is the subset of *T* having value *i* on attribute *A*.

The information gained by partitioning the examples in *T* according to their values on *A* is

$$\mathrm{G}(A) = \mathrm{I}(T) - \mathrm{I}_A(T)$$

Consider now the information content of a message pertaining not to a class but to the outcome of a test on an attribute. The expression

$$P(A) = -\sum_{i=1}^{n}\left[\frac{|T_i|}{|T|}\log_2\left(\frac{|T_i|}{|T|}\right)\right]$$

represents the information generated by the mere act of partitioning T into n subsets. Then $GR(A) = G(A)/P(A)$ expresses the proportion of information generated by the partition on A that is useful for classification. To determine which attribute to install at a given position in the tree, the maximum value of GR over all untested attributes is used. This is the gain−ratio criterion. (The gain criterion, an older alternative, maximizes G rather than GR.[28])

In the case of missing data, $I(T)$ and $I_A(T)$ were computed using only known values, $I(A)$ was computed as $K(I(T) - I_A(T))$, where K is the proportion of T with known values of A, and $P(A)$ was computed on $n + 1$ subsets, treating unknown as an additional subset. When partitioning T on attribute A, each example whose value of A is unknown was distributed fractionally among the subsets $T_{(A\leftarrow i)}$ of T in proportion to the membership in the $T_{(A\leftarrow i)}$ of the examples having known values on A. For more details on decision tree induction, see refs 28−30.

3.2. Converting Trees to Rules. Decision tree structures can be large and difficult for humans to understand and can contain redundant subtrees which hide the underlying structure of information. Production rules can avoid these difficulties. Therefore, after the tree induction phase produced 10 trees from each experiment, a single set of production rules was generated from these 10 trees, by converting each path from root to leaf in each decision tree into a corresponding initial rule. This set was then pruned in a process which worked as follows. First, each individual rule was simplified in isolation by removing conditions from its left-hand side that did not discriminate the rule's class (either positive, negative, or equivocal, depending on the rule) from other classes, according to a pessimistic estimate based on contingency table analysis. Then, for each class, all the simplified rules were filtered to remove rules that did not contribute to the accuracy of the rules as a whole. Finally, the rules for each class were ordered to minimize false positives, and the class which contained the most training cases not covered by any rule was chosen as the default class.

3.3. Tree and Rule Results. After cross-validation, overall tree and rule models were then built by using the entire training set of 744 nonequivocal experiments. The results of classification of the training examples by the overall cross-validated tree model are summarized in Table 1. The results of classification of the training examples by the overall cross-validated rule model are summarized in Table 2.

Table 1. Training Set Classification Accuracy (Tree Model ($p < 0.001$))[a]

		pred bioassay		
		pos	neg	tot.
actual bioassay	pos	462	6	468
	neg	66	210	276
	tot.	528	216	744
accuracy			0.90	
sensitivity			0.99	
specificity			0.76	
+ predictivity			0.88	
− predictivity			0.97	
corr coeff			0.80	

[a] Abbreviations: pos, positive; neg, negative; tot., total.

Table 2. Training Set Classification Accuracy (Etracted Rule Model ($p < 0.001$))[a]

		pred bioassay		
		pos	neg	tot.
actual bioassay	pos	468	0	468
	neg	104	172	276
	tot.	572	172	744
accuracy			0.86	
sensitivity			1.00	
specificity			0.62	
+ predictivity			0.82	
− predictivity			1.00	
corr coeff			0.71	

[a] Abbreviations: pos, positive; neg, negative; tot., total.

4. NEURAL AND *M*-OF-*N* RULE MODELS

Having worked on symbolic tree and rule models, we wanted to see if neural network models could be made competitive in terms of both accuracy and understandability. In constructing our neural models, we developed new techniques to simplify the data, simplify the learned models, and explain the models in terms that human experts can understand. First, relevant feature subset selection was used to identify, independent of expert knowledge, which features in the training data were relevant to the prediction task. By using only the relevant features during the training of the neural networks, the predictive accuracy of neural networks on this data set was significantly improved. Second, connection weight pruning was applied to speed up the performance of the network and simplify its internal structure, making rule extraction easier. Connection weight pruning also improved the predictive accuracy of the trained networks. Finally, a new pedagogical method for extracting *M*-of-*N* rules from neural networks was developed and applied to the networks.

4.1. Neural Methodology. A three-layer fully connected feed-forward neural network with four hidden units was used as the basic model for all of the results that follow. Training and testing of the neural networks was done primarily using the Aspirin/MIGRAINES Neural Network Software[24] running on a Sun Sparc workstation. The learning method used throughout was standard error back-propagation[32] using a mean-squared-error function. An inertia term was used to smooth the weight changes over time. No attempt was made to define a formal method for choosing the values for learning rate and inertia.

The toxicology data contain eight different types of data that must be mapped to the input layer of the neural network. Real-valued attributes were normalized to a value between 0 and 1 and mapped to an input node. The values for most of the data are discrete; in those cases, the data were simply mapped to a set of binary input nodes. Input nodes whose input is uniformly 0 in all examples in the data set may simply be removed without changing the behavior of the network. In the case of the 744 examples in the toxicology data, this condition applies to nearly 90% of the data (2527 features). Removing the 2527 noncontributing features leaves a set of 288 features and an input layer in the network of 288 nodes.

The size of the hidden layer was computed by using the number of training examples, the size of the input layer based upon the number of relevant features determined by relevant feature subset selection, and the size of the output layer to guarantee that the network has fewer undetermined parameters than the number of examples in the training set.[35,36] This helps to prevent the network from overfitting the data. An average hidden layer of size 4 resulted in a network computationally very feasible for both training and testing.

4.2. Phases of Training. Using an artificial neural network to generate symbolic rules as a model for chemical carcinogenesis in rodents is a multistep process. A standard back-propagation neural network does not handle the toxicology data well without the aid of critical steps both before and after the back-propagation phase. First, irrelevant and noisy attributes were removed from the training set. Second, the neural network was trained using the remaining input attributes. Third, connection weight pruning was used to simplify and speed up the network model. Finally, symbolic rules were extracted from the network which explain the learned model to the domain experts.

4.3. Neural Results. To estimate the predictive ability of the neural net model described previously, 10-way cross-validation was used. The data were split randomly into 10 sets with the single condition that the proportion of examples with a positive classification to examples with a negative classification be the same in each of the 10 sets. Since there was a total of 744 examples in the data, the average training set size was 670 examples and the average test set size was 74 examples.

It is interesting to note that obtaining perfect prediction results with the toxicology training data is theoretically impossible. There are examples in the data with identical attribute values yet opposite classifications. For these examples, the neural network must simply learn the classification which occurs most frequently and accept being wrong on those examples with the opposite classification. One subset of the data where this occurs often is the examples in which no organ morphologies are present. This subset of nonlearnable examples accounts for almost 7% of the toxicology data, and there are other similar subsets of the data that are equally nonlearnable. In practice, it was discovered that the training accuracy of the neural net reached a maximum at between 90 and 92% accuracy.

Most neural net research recommends that training be stopped before the network reaches 100% accuracy anyway. This practice, called the early stopping method, is used to avoid overtraining the network. The most common observation is that the predictive accuracy of the network will rise along with the training accuracy until some maximum is reached. At that point, the training accuracy will continue to rise while the predictive accuracy falls off sharply. This is the point where the network stops generalizing and begins memorizing the training data.

In the case of the network that was chosen to train on the toxicology data, the size of the hidden layer was chosen to ensure that the network could not overtrain. The expected observation, therefore, is that both the training accuracy and the predictive accuracy will rise to a maximum and then level off without dropping back down. Indeed, while the neural network models were trained, the training accuracy rose slowly toward 90% while the test accuracy leveled off at about 80%. A predictive accuracy of 80% already compares favorably with some of the other machine learning approaches to modeling this data set. However, as the next section will show, feature selection can significantly improve the predictive accuracy of the neural network.

Table 3. Training Set Classification Accuracy after Feature Selection (Neural Model ($p < 0.001$))[a]

		pred bioassay			
		pos	neg	equiv	tot.
actual bioassay	pos	392	60	–	452
	neg	20	272	–	292
	equiv	–	–	–	–
	tot.	412	332	–	744
accuracy				0.89	
sensitivity				0.87	
specificity				0.93	
+ predictivity				0.95	
– predictivity				0.82	
corr coeff				0.78	

[a] Abbreviations: pos, positive; neg, negative; equiv, equivocal; tot., total.

4.3.1. Feature Selection Results. The single hidden unit method[35] was used to find a relevant subset of the 288 attributes that occur in the predictive toxicology data. The first step of assigning the relevance weights is very quick and efficient using this method. Furthermore, finding an optimal value for the threshold required only four passes through the hill climbing algorithm. The optimal threshold was determined to be within 0.2 of the standard deviation of the relevance weights. After the threshold was used to split the 288 features into relevant and irrelevant features, an average of 74 features were labeled as relevant. A network using only the relevant features was then tested using 10-way cross-validation on the complete set of 744 examples from the data. The results are listed in Table 3. The test accuracy of the network prior to using feature selection was only 80%. The estimated predictive accuracy of the network did therefore rise significantly by using the single hidden unit method. In addition, a test accuracy of 89% was considerably better than any of the alternative methods considered for relevant feature subset selection.

The above results demonstrate that the cross-validated accuracy of neural networks trained on all 744 examples can be much improved by using the single hidden unit feature selection. Additional testing also showed that randomly generated features and features which are clearly irrelevant are indeed thrown out by this method of feature selection. These results demonstrate that the single hidden unit method for feature selection not only improved training accuracy through hill-climbing but also eliminated individual features which were not relevant to the classification task.

4.3.2. Connection Weight Pruning Results. The test results for iterative pruning are listed in Table 4. An average of 279.8 (80%) of the connection weights were pruned using this method. Most of the connection weights remaining were incident on the third hidden node.

The cross-validated accuracy was positively affected by iterative pruning. The average estimated prediction accuracy rose almost 0.5%. Considering that the network is roughly 20% the size it was prior to pruning, the rise in accuracy is an excellent result.

Table 4. Training Set Classification Accuracy after Weight Pruning (Neural Model)

av. no. of pruned weights	278.8
av. no. on node 1	84.2
av. no. on node 2	78.9
av. no. on node 3	29.8
av. no. on node 4	85.9
cross-validated test accuracy	90%

Table 5. Training Set Classification Accuracy after Rule Extraction (Neural Model ($p < 0.001$))[a]

		pred bioassay			
		pos	neg	equiv	tot.
actual bioassay	pos	389	74	−	463
	neg	25	256	−	281
	equiv	−	−	−	−
	tot.	414	330	−	744

accuracy	0.87
sensitivity	0.84
specificity	0.91
+ predictivity	0.94
− predictivity	0.78
corr coeff	0.73

[a] Abbreviations: pos, positive; neg, negative; equiv, equivocal; tot., total.

Since the iterative method both improved the predictive accuracy of the network and significantly reduced the number of connection weights, this pruning method was selected from several alternative methods to employ on the trained neural networks prior to weight analysis and rule extraction.

4.3.3. Rule Extraction Results. A revised approach to brute force rule extraction was applied to the best trained and pruned neural network from previous results. Listed in Table 5 are 10-way cross-validated results. Just over 2.5% of predictive accuracy was lost between the neural network and the set of *M*-of-*N* rules. The rule set itself is compact and easy to read and interpret. The rule set was generated from a $74 \times 4 \times 1$ neural network trained on the complete set of 744 training examples. Only 22 rules were required to completely model the behavior of the trained neural network. Of these, two of the rules are used the majority of the time. A few other rules are used four or more times. The bulk of the rules are used only once or twice. This would indicate that there are only a few general rules which can be defined to describe the classifications in the data. To achieve high accuracy, these general rules must be supplemented with many other rules to describe special cases. This requirement makes perfect sense given the nature and source of the data.

The revised method for brute force rule extraction is quick and easy to implement, yet it yields a reasonably sized rule set that is easy to interpret and very closely models the behavior of the neural network. Furthermore, this method for rule extraction, unlike those in ref 15, work regardless of the structure of the neural network or the activation function. The capability of extracting rules from a general trained neural network upgrades the status of neural networks from black boxes to powerful machine learning tools that are capable of explaining their results. In this case, a set of rules has been extracted which is simple and readable and has an estimated predictive accuracy of roughly 87%. The rules also have a very high positive predictivity. This indicates that they could be used as a first step in analysis of test articles for carcinogenity. Test articles that were predicted negative might still require bioassay, but one could be fairly certain that those test articles labeled positive by the rule set were truly positive.

As an example, the following is one of the 22 extracted rules for prediction of chemical carcinogenesis in rodents. Space prohibits inclusion of the complete set.

Sample Rule: Test article is Positive if any 3 or more of the following and no other features are present: positive (+) salmonella; SA type G; SA type Q; above avg. Clogp; above average PA; above average *Z*-depth; forestomach hyperkeratosis; kidney karyomegaly; kidney necrosis; liver hypertrophy; lung histiocytosis; skin erosion/ulceration.

5. BAYESIAN MODELS

The onset of cancer in rodents, or in humans, is almost certainly a multifactorial process, and there is much room for individual variability in response to carcinogenic substances. In order to better accommodate the uncertain and probabilistic nature of these underlying mechanisms, largely still unknown, we then proceeded to experiment with Bayesian models, which explicitly encode conditional probabilistic relationships among our data and between the data and the endpoint of rodent carcinogenesis.

5.1. Bayesian Methodology. We first split the training data into four subsets, one for each sex−species. Each training set contained the test article-specific attributes and the experiment-specific attributes for that sex−species. The classification attribute was the classification NTP gave to that sex−species, not the overall bioassay result. We again used 10-way cross-validation to estimate accuracy. We generated a classification for each example in each of the four training sets. Those classifications were then combined into an overall bioassay prediction using NTP's methodology: if any sex−species was predicted positive, the overall classification was positive; if no sex−species was predicted positive and at least one was predicted equivocal, the overall classification was equivocal; if all sex−species predictions were negative, then the overall classification was negative.

This design matches NTP, but also adds to the intuitive nature of the resulting models. One of our goals is to try to give toxicologists and biochemists some insight into the biological pathways associated with cancer. Multiple experiments are completed, after all, because there is an implicit assumption that biological pathways differ in the various sex−species. If we combine the data for the various sex-species, we may obscure the differences in the pathways.

5.2. Bayesian Classifiers. A Bayesian classifier is a directed graph in which there are nodes which represent each of the attributes and a node which represents the class. An example of such a graph is shown in Figure 1. The node which represents the class is called the **class node** and the other nodes are called **evidence nodes**.

Each node in a Bayesian classifier has a state for each possible value the associated attribute may take on. Saying that the node is in a particular state is equivalent to saying the attribute has the specific value associated with that state of the node. Each node is labeled by a probability distribution over its states. That probability distribution represents the

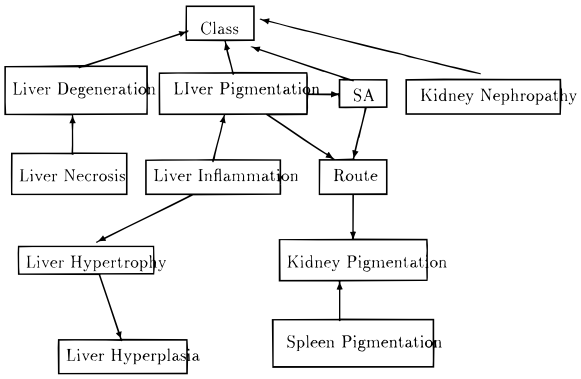

Figure 1. Bayesian classifier for female rat carcinogenesis.

probability that the associated attribute will take on each of its possible values. These are initialized to the probabilities that the attribute will take on each of its possible values in the population being modeled. For example, a node representing the attribute describing if an experiment produced evidence of subchronic liver hypertrophy can have two values: PRESENT and ABSENT. That node would be initialized with a vector of two probabilities: the probability some test article studied in a particular sex/species would show evidence of hypertrophy, and the probability it would not. These probabilities are most often estimated from a population, such as our training set. If a test article is shown to in fact cause the effect, the attribute for that effect is given the value PRESENT, the node labeled with that attribute is in the PRESENT state, and its probability distribution is changed to 1 for the PRESENT state and 0 for the ABSENT state. In this way, Bayesian networks incorporate evidence that may become known only after the model is constructed.

In our Bayesian models there is a path to the class node from every evidence node. The class node is called the **child** of every evidence node, and the evidence nodes are called **parents** (or **ancestors**) of the class node. (A Bayesian classifier is a special case of a Bayesian belief network with this limited structure. For more information on Bayesian belief networks, see refs 20, 22, 26, and 27.)

Each arc in a Bayesian classifier is labeled by a matrix representing the conditional probability that the node at the head of the arc (the child node) will be in a particular state, given the state of the parent node. Specifically, the arc from the node representing the class to the *i*th evidence node will be labeled by a matrix M_i. The entry in that matrix $M_i[j,k]$ will contain the probability that the *i*th evidence node will be in state *j* when the class node is in state *k*.

Once the Bayesian classifier has been built, classifications of new entities can be predicted by specifying the state of each node associated with an attribute whose value is known for the new entity. This is called **instantiating** the evidence nodes and is accomplished by changing the probabilities stored at these nodes so that the current state has a probability of 1 and all of the other states have probability of 0. When instantiation is complete, the probabilities of the states of the class node can be recomputed using Bayes' Law:

$$P(C|e) = P(e|C)\,P(C)/P(e)$$

In this equation, $P(C|e)$ is the probability distribution of the class node given the evidence which has been instantiated about the test article being classified. $P(e|C)$ is the probability of the evidence given the current probability distribution of the class node. This is found from the conditional probabilities which label the arcs of the graph. $P(C)$ is the current probability distribution of the class node, and $P(e)$ is the current probabilities of the evidence nodes.

The process of instantiating the known evidence nodes and recalculating the probability distribution of the class node is called **inference**. The statistics are more complex when there are multiple evidence nodes, but the main idea of inference is contained in Bayes' law. The result of inference is a new probability distribution for the class node representing the probabilities that the test article is in each class.

5.3. Learning Bayesian Models. Our training sets were the input to a sequence of algorithms to learn first the structure and then the probabilities of a Bayesian classifier for each sex−species. The steps in this sequence included real attribute discretization, feature selection, learning a similarity network, constructing the equivalent Bayesian classifier, and learning the probabilities.

5.3.1. Discretizing Continuous Attributes. Each node of a Bayesian classifier must have a finite number of states, so a state of a node representing a continuous valued attribute must be associated with a subrange of the possible values of that attribute. Finding those subranges is called **discretizing** the attribute because it allows a finite number of states to be associated with selected, discrete subranges of the possible values.

Discretization was accomplished using an algorithm based on a minimal entropy heuristic.[12,14] The entropy of a set of instances *S* (in this case, test articles) is

$$\mathrm{Ent}(S) = -\sum_{i=1}^{k} P(C_i,S)\log(P(C_i,S))$$

which roughly measures the amount of information required to specify the classes in *S*. As *S* is more heterogeneous (with respect to class), its entropy will be larger. The goal of the algorithm is to take a set of instances *S* (which in this case is the test articles in the training set) and partition it into subsets based on ranges of the attribute being discretized so that the entropy of the subsets is minimized.

The algorithm takes a set of instances (in this case test articles) and sorts them by the attribute being discretized. The algorithm then looks at each possible partition boundary *T* which will divide *S* into two subsets S_1 and S_2 and measures the **class information entropy of the partition induced by *T***, $E(T,S)$ which is defined as

$$E(T,S) = \frac{|S_1|}{|S|}\mathrm{Ent}(S_1) + \frac{|S_2|}{|S|}\mathrm{Ent}(S_2)$$

The algorithm will discretize the attribute at the partition boundary $T_{\min}$ which minimizes $E(T,S)$ and will then recursively discretize the subsets S_1 and S_2. This recursion will stop when partitioning a subset of the test articles does not result in sufficient entropy gain. Entropy gain of a partition boundary *T* is measured by

$$\mathrm{Gain}(T,S) = \mathrm{Ent}(S) - E(T,S)$$

and recursive partitioning will stop when the value for $\mathrm{Gain}(T_{\min},S)$ passes below a predetermined threshold.

5.3.2. Feature Selection. We used the simple feature selection method associated with APRI learning.[13] In this method, the mutual information between each attribute and the class attribute is measured. Mutual information of a pair of discrete random variables defined as

$$I(X,Y) = \sum P(X,Y) \log(P(X,Y)/P(X)/P(Y))$$

where the sum is overall possible combinations of the values attributes X and Y may take on. Mutual information measures the strength of the correlation between the values of the attribute and the values of the class. The attributes are sorted by this measure and are included in the model in that order until a given percentage of the total mutual information has been included. We used 25−35% for that cutoff.

5.3.3.3. Learning Similarity Networks. Our training sets have more than two classifications, so we have chosen to use learning similarity networks as a stepping stone to learning a Bayesian classifier. Similarity networks are equivalent to Bayesian networks,[17] but allow for more explicit modeling when there are more than two classifications. Our learning methods are described in detail elsewhere.[40]

The first part of a similarity network is a hypergraph where the nodes are the classifications and edges represent "similarity" between classifications. This is called the *similarity hypergraph*. Similarity networks were designed to aid in expert construction of Bayesian networks, and "similarity" was originally measured by human intuition. We used some simple mathematics to quantify that intuition. For each pair of examples with differing classes, we summed the absolute value of the differences in the values of their attributes. The sum of those differences for each pair of classes measured the inverse of the similarity of the classes. In the order given by that measure of similarity, edges are added to the hypergraph until it is connected while remaining acyclic.

The second part of a similarity network is a *local knowledge map* specific to each node of the similarity hypergraph. Each "map" is a Bayesian belief network modeling reasoning as if only the classes associated with that edge exist. In human construction, this allowed experts to distinguish between examples with those classifications. Our hope is to give our learning algorithms that same insight.

In order to learn the structure of the local knowledge maps, we created training sets specific to the associated classifications. These have the same structure as our original training sets, but examples with other classifications (i.e., classifications not in the associated edge of the hypergraph) have been temporarily discarded.

After creating these new training sets, the MDL[21] structure learning algorithm was used to learn the structure of each local knowledge map. MDL learning is a heuristic search for an optimal structure that balances the values of accuracy and compactness of the model.

This method seemed most appropriate to us as it matches our goals. We value accuracy, but, since 219 test articles previously tested are almost certainly not representative of the universe of chemicals, we do not believe that cross-validation will predict true off-training set accuracy. Therefore, we did not want to focus only on the predicted accuracy of our models. We also value models that are small enough to be interpreted by humans who may gain insights from them.

Table 6. Training Set Classification Accuracy (Bayesian Models ($p < 0.001$))[a]

		pred bioassay			
		pos	neg	equiv	tot.
actual bioassay	pos	86	35	3	124
	neg	26	33	0	59
	equiv	12	22	2	36
	tot.	124	90	5	219
accuracy				0.66	
sensitivity				0.71	
specificity				0.56	
+ predictivity				0.77	
− predictivity				0.49	
corr coeff				0.26	

[a] Abbreviations: pos, positive; neg, negative; equiv, equivocal; tot., total.

5.3.4. Constructing the Bayesian Classifier. By the time the similarity network is complete, creating the Bayesian classifier is mechanical. The local networks of the subpopulations are combined with a graph union operation to complete the structure of a Bayesian classifier.

5.3.5. Learning the Probabilities. Because there are probabilities at every node and conditional probabilities at every arc, there are many probabilities which must be specified in building a network. However, once the structure of the network is specified, all of these probabilities can be learned from a set of examples from the population being modeled. The probabilities are learned from the training set using the standard method of Dirichlet priors.[40]

For our experiments, the Bayesian belief network application Netica was used to learn the probabilities. Netica assumes the conditional probabilities being learned are independent and that the prior distribution is Dirichlet. It then uses a β function, parametrized by experience and a probability number, to represent the distribution over possible probabilities.

5.4. Bayesian Results. Our results thus far include Bayesian network models and predicted accuracies for them. We actually constructed four types of models (one for each sex−species). When trying to predict the results for the sex−species experiment, these models had an average cross-validated accuracy of 55%. Interestingly, when the results given by the individual sex−species models are combined into an overall bioassay prediction, the predicted accuracy is not much less. The results of classification of the training examples by the overall cross-validated Bayesian model are summarized in Table 6. For comparison, we also tabulated results for a model which predicted class simply in proportion to the overall occurrence of each class in the training set. See Table 7 for results of this "random guesser".

Since human experts do not predict the equivocal class, we ran some preliminary experiments excluding equivocals. This did not seem to help the accuracy and resulted in lower specificity, so we have not pursued that line of research. In examining these results further, the models have much higher rate of false negatives than false positives. With equivocals excluded, the specificity of the model is 56%, while the

Table 7. Training Set Classification Accuracy (Random Guesser ($p < 0.95$))[a]

		pred bioassay			
		pos	neg	equiv	tot.
actual bioassay	pos	70	33	20	123
	neg	33	16	10	59
	equiv	20	10	6	36
	tot.	123	59	36	218
accuracy				0.57	
sensitivity				0.68	
specificity				0.33	
+ predictivity				0.68	
− predictivity				0.33	
corr coeff				0.01	

[a] Abbreviations: pos, positive; neg, negative; equiv, equivocal; tot., total.

sensitivity is 71%. This means that if the data set is representative, then positive predictions are meaningful.

6. DISCUSSION

In summary, we have developed several techniques for predicting the carcinogenicity of a test article from physical chemical structure, in vitro assays, and short-term in vivo experiments and have shown the resulting models to have reasonably high cross-validated accuracies.

All these results should be considered only preliminary, and much work remains to be done. One vital quality of any useful model must be how amenable it is to understanding by human experts. Decision tree models and rules extracted from trees are relatively transparent, but such models constructed from our data can grow quite large and unwieldy unless a radical pruning strategy is employed. For neural models, on the other hand, a great deal of pre- and postprocessing is necessary to make this approach yield useful models of carcinogenicity, but the resulting *M*-of-*N* rules have so far proven to be of tractable size and of substantial explanatory power. The Bayesian models not only inherently incorporate the uncertain and probabilistic aspects of such a complex biophysical process as carcinogenesis, but the construction method we are using explicitly quantifies the tradeoffs between model accuracy and complexity.

We consider it highly unlikely that any single carcinogenicity model will prove superior to all others in the long run, so we are also engaged in experiments in combining the results of models by various principled schemes.

ACKNOWLEDGMENT

The authors would like to thank Dr. Ray Tennant of NIEHS and Dr. Ann Richard of USEPA for their longstanding cooperation and support of this work, and the anonymous reviewers for their helpful comments on an earlier verion of this manuscript.

REFERENCES AND NOTES

(1) Ashby, J.; Paton, D. The influence of chemical structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogen exposures. *Mutat. Res.* **1993**, *286*, 3−74.

(2) Ashby, J.; Tennant, R. W.; Zeiger, E.; Stasiewicz, S. Classification according to chemical structure, mutagenicity to salmonella, and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. *Mutat. Res.* **1989**, *223*, 73−103.

(3) Ashby, J.; Tennant, R. W. Definitive Relationships among chemical structure, carcinogenicity, and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat. Res.* **1991**, *257*, 229−306.

(4) Bahler, D.; Bristol, D. W. Prediction of Chemical Carcinogenicity in Rodents By Machine Learning of Decision Trees and Rule Sets. *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*; Papers from the 1999 AAAI Spring Symposium, Technical Report SS-99-01, AAAI Press: Menlo Park, CA, 1999; pp 74−77.

(5) Bahler, D.; Bristol, D. W. The Induction of Rules for Predicting Chemical Carcinogenesis in Rodents. In *Intelligent Systems for Molecular Biology;* Hunter, L., Shavlik, J., Searls, D., Eds.; AAAI/MIT Press: Cambridge, MA.

(6) Bahler, D.; Bristol, D. W. A Quantitative Comparison of the Utility of Characteristics for Predicting Chemical Carcinogenesis. *4th Annual Keck Symposium on Computational Biology*, University of Pittsburgh, Pittsburgh, PA, 1993.

(7) Bristol, D. W.; Bahler, D. Summary and recommendations for session B: Activity classification and structure−activity relationship modeling for human health risk assessment of toxic substances. *Toxicol. Lett.* **1995**, *79*, 265−280.

(8) Bristol, D. W.; Bahler, D. Inductive approaches to predicting toxicity. *Proceedings of the 20th Annual Summer Toxicology Forum*; Givin Institute of Pathobiology: Aspen, CO, 1995.

(9) Bristol, D. W.; Bahler, D. Database Analysis to Identify Features, Provide Heuristic Information about Biological Factors, Manage Information, and Classify Chemicals. *Proceedings of the 2nd European Conference on High-Throughput Screening and Molecular Diversity*, European Society of Toxicology: Budapest; 1995.

(10) Bristol, D. W.; Wachsman, J. T.; Greenwell, A. The NIEHS Predictive-Toxicology Evaluation Project. *Environ. Health Perspect.* **1996**, *104* (Suppl. 5), 1001−1010.

(11) Bristol, D.W.; Tennant, R. W.; Bahler, D. Predicting Chemical Carcinogenicity: Progress, Pitfalls, and Promise. *26th Annual Symposium*, Society of Toxicology of Canada: Toronto, December; 1993.

(12) Dougherty, J.; Kohavi, R.; Meharn, S. Supervised and Unsupervised Discretization of Continuous Features. *Mach. Learn.* **1995**, *12*, 194−202.

(13) Ezawa, K. J.; Schuerman, T. Fraud/Uncollectible Debt Detection Using a Bayesian Network Based Learning System: A Rare Binary Outcome with Mixed Data Structures. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*; Morgan Kaufmann: San Mateo, CA, 157−166.

(14) Fayyad, U.; Irani, K. B. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *Procedings of the 12th National Conference on Artificial Intelligence (AAAI-95)*; Morgan Kaufmann: San Mateo, CA, 1995; pp 1022−1027.

(15) Fu, L. Rule Generation from Neural Networks. *IEEE Trans. Syst., Man, and Cybernetics* **1994**, *24* (8, Aug) 1114−1124.

(16) Gini, G.; Lorenzini, M.; Benfenati, E.; et al. Predictive Carcinogenicity: A model for aromatic compounds, with nitrogen-containing substituents, based on molecular descriptors using an artificial neural network. *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 1076−1080.

(17) Heckerman, D. Probabilistic Similarity Networks. *Networks* **1999**, *20*, 607−636.

(18) Huff, J.; McConnell, E. E.; Haseman, J. K.; et al. Carcinogenesis Studies: Results from 398 Experiments on 104 Chemicals from the U.S. National Toxicology Program. *Proc. Natl. Acad. Sci.* N.Y. **1988**, *534*, 1−30.

(19) Huff, J.; Haseman, J. Long-term chemical carcinogenesis experiments for identifying potential human cancer hazards: Collective database of the National Cancer Institute and National Toxicology Program (1976−1991). *Environ. Health Perspect.* **1991**, *96*, 23−31.

(20) Jensen, F. V.; Olesen, K. G.; Andersen, S. K. An Algebra of Bayesian Belief Universes for Knowledge-Bases Systems. *Networks* **1990**, *20*, 637−659.

(21) Lam, W.; Bacchus, F. Learning Bayesian Belief Networks: An Approach Based on the MDL Principle. *Comput. Intell.* **1994**, *10* (3), 269−293.

(22) Lauritzen, S. L.; Spiegelhalter, S. L. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. R. Stat. Soc. B* **1988**, *50*, 157−224.

(23) Lee, Y.; Buchanan, B. G.; Aronis, J. M. Knowledge-Based Learning in Exploratory Science: Learning Rules to Predict Rodent Carcinogenicity. *Mach. Learn.* **1998**, *30*, 217−240.

(24) Leighton, R. R. *The Aspirin/MIGRAINES Neural Network Software Release v.6.0*; Mitre Corp.: Bedford, MA, 1992.

(25) Lewis, D. F. V. Comparison between Rodent Carcinogenicity Test Results of 44 Chemicals and a Number of Predictive Systems. *Regul. Toxicol. Pharmacol.* **1994**, 215−222.

(26) Pearl, J. *Probabilistic Reasoning in Intelligent Systems;* Morgan Kaufmann: San Francisco, 1988.
(27) Peot, M. A.; Shachter, R. D. Fusion and propagation with multiple observations in belief networks. *Artif. Intell.* **1991**, *46*, 299−318.
(28) Quinlan, J. R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81−106.
(29) Quinlan, J. R. Simplifying Decision Trees. *Int. J. Man−Mach. Stud.* **1987**, *27*, 221−234.
(30) Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann: San Mateo, CA, 1993.
(31) Rodericks, J. V. Risk Assessment, the Environment, and Public Health. *Environ. Health Perspect.* **1994**, *102*, 258−264.
(32) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Internal Representations by Error Propagation. *Parallel Distributed Processing*; MIT Press: Cambridge, MA, 1986; pp 318−362.
(33) Shannon, C, E. *The Mathematical Theory of Communication;* University of Illinois Press: Urbana, IL, 1949.
(34) Srinivasan, A.; Muggleton, S. H.; King, R. D.; Sternberg, M. J. E. The Predictive Toxicology Evaluation Challenge. *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, Morgan Kaufmann: San Francisco, 1997; pp 4−9.
(35) Stone, B *Feature Selection and Rule Extraction for Neural Networks in the Domain of Predictive Toxicology*. Technical Report; Department of Computer Science, North Carolina State University: Raleigh, NC, 1999.
(36) Stone, B.; Bahler, D. *Predicting Chemical Carcinogenesis in Rodents with Artificial Neural Networks and Symbolic Rules Extracted from Trained Networks*. *Proceedings of the AAAI Spring Symposium on Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*; AAAI Press: Menlo Park, CA, 1999.
(37) Tennant, R. W.; Spalding, J.; Stasiewicz, S.; Ashby, J. Prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 44 chemicals by the National Toxicology Program. *Mutagenesis* **1990**, *5* (1), 3−14.
(38) Wellington, C.; Bahler, D. Predicting Rodent Carcinogenicity by Learning Bayesian Classifiers. *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*. Papers from the 1999 AAAI Spring Symposium, Technical Report; SS-99-01; AAAI Press: Menlo Park, CA, 1999; 131−134.
(39) Wellington, C.; Bahler, D. Learning to Predict Carcinogenesis of Unstudied Chemicals in Rodents from Completed Rodent Trials. Proceedings of International IMACS Conference on Scientific Computing and Mathematical Modeling, International Association for Mathematics and Computers in Simulation: Piscataway NJ, 8−16.
(40) Wellington, C. Capitalizing on Asymmetric Relationships When Learning the Structure of a Bayesian Classifier. Ph.D. Dissertation, Department of Computer Science, North Carolina State University, Raleigh, NC, 1997.

CI990116I