The Induction of Rules for Predicting Chemical Carcinogenesis in Rodents*

Dennis Bahler
Dept. of Computer Science
North Carolina State University
Box 8206, Raleigh NC 27695-8206
bahler@ncsu.edu
(919) 515-3369, Fax 515-7896

Douglas W. Bristol
National Institute of
Environmental Health Sciences
Box 12233, B3-04
Research Triangle Park, NC 27709
bristol@vaxe.niehs.nih.gov
(919) 541-2756

Abstract

This paper presents results from an ongoing effort in applying a variety of induction-based methods to the problem of predicting the biological activity of noncongeneric (structurally dissimilar) chemicals. It describes initial experiments, the long-term goal of which is to assist toxicologists, cancer researchers, regulators, and others to predict the toxic effects of chemical compounds. We describe a series of experiments in tree and rule induction from a set of example chemicals whose carcinogenicity has been determined from long-term animal studies, and compare the resulting classification accuracy with eight published human and computer predictions for a common set of 44 test chemicals. The accuracy of our system is comparable to the most accurate human expert prediction yet published, and exceeds that of any of the computer-based predictions in the literature. The induced rules provide confirmation of current expert heuristic knowledge in this domain. These early results show that an inductive approach has excellent potential in predictive toxicology.

Introduction

Cancer is an obvious public health problem, and there is considerable evidence of carcinogenicity in chemical compounds found commonly in the environment. It follows that determining which chemicals are carcinogenic is of clear public benefit. At the same time, however, experiments such as those currently being performed by the U. S. National Toxicology Program (NTP) (Huff & Haseman 1991) to determine which chemicals cause cancer in rodents, while clearly important, are 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 project to test our hypothesis, and this paper presents some of our early results. 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.

Current Approaches to Predicting Carcinogenesis

Toxicity and carcinogenicity studies currently conducted by the NTP are almost exclusively empirical, consisting of short- and long-term animal studies using rats and mice. In a typical NTP carcinogenicity study, the test chemical is administered for 104 weeks to each sex of either one or two species of rodent (usually Fisher Rat and/or $B_6C_3F_1$ Mouse). At the end of the two-year exposure phase, animal tissues are examined histopathologically for the presence of lesions at any of 59 organ sites. The results are interpreted and, after peer review, the level of evidence for carcinogenic activity of the chemical in each sex/species experiment is classified as either Clear Evidence, Some Evidence, No Evidence, Equivocal Evidence, or Inadequate Study. A variety of schemes have been used to classify the overall carcinogenicity of a chemical (Tennant 1993), but predictive toxicology researchers usually classify a chemical as either positive, if it produces any Clear- or Some-Evidence calls, or negative, if it produces only No-Evidence calls or a combination of Equivocal- and No-Evidence calls in any of the NTP bioassay experiments. Some predictors consider the combination of Equivocal- and No-Evidence calls to represent a third, equivocal, classification. The sex/species results are

*This work is supported in part by NIEHS/NIH.

published as an NTP Technical Report and are often used by an appropriate regulatory agency as an important part of an overall assessment of risk. The time from design to reporting a long-term study is five years or more. For more information on NTP experimental protocol, see (Huff & Haseman 1991).

About 450 Technical Reports have been published, but the number of chemicals in the environment that have never been subjected to carcinogenicity testing has been estimated as high as 100,000. At the moment there appear to be few good alternatives to long-term empirical carcinogen bioassays of the kind done by the NTP.

The Training Set

This paper describes a series of experiments in supervised tree and rule induction from a training set of example chemicals, whose carcinogenicity has been determined by long-term rodent studies. It then compares the resulting classification accuracy with a set of published human and computer predictions for a common set of 44 test chemicals. For the training set, we had ready access to information on 301 chemicals, both organic and inorganic, for which both long-term bioassays and short-term mutagenesis assays had been completed (Ashby & Tennant 1991). The information was organized into 301 training examples, with each example containing data on 189 attributes. These attributes encode values for:

***Salmonella* mutagenesis.** The Ames test for mutagenesis had been performed on *Salmonella typhimurium* bacteria on most of the chemicals in the test set (Ashby et al. 1989). This is a short-term (typically 14-day) *in vitro* test, whose results were negative if and only if there was no mutation, with and without "activation" by S9 liver microsomal fraction.

Alerting chemical structures. Human expertise has identified certain functional-group substructures of organic molecules that may predispose the parent molecule towards 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 (Ashby et al. 1989). An attribute for each of the following 21 structural alerts for DNA reactivity was included in our training set: alkyl esters of either phosphonic or sulphonic acids; aromatic nitro groups; aromatic azo groups; aromatic ring *N*-oxides; aromatic mono- and di-alkylamino 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.

Route of administration and MTD doses. The route of administration for each NTP carcinogenesis bioassay is usually set according to the predominant mode of human exposure in the environment at large. Alternative routes of administration are feed, drinking water, gavage (intubation), skin paint, and inhalation. A preliminary subchronic (typically 90-day) toxicity study is conducted, primarily to determine the minimally-toxic dose (MTD) for the long-term study and to observe any subchronic pathology. The MTD is the highest dose for a chronic study that will not shorten the longevity of treated animals from effects attributed to the test chemical other than the induction of neoplasms. For the MTD attributes in our training set, concentration units were normalized to mg per kg per day for each sex/species bioassay experiment, regardless of the route of administration.

Subchronic organ pathology. After the exposure phase for a subchronic study is completed, organs of all test animals are examined for gross lesions at necropsy, and each of up to 59 organs may, as warranted, be subjected to microscopic examination. Toxicologic pathologists have standard codes for an average of 40 morphological lesions per organ, but on average fewer than 4 are observed at any one topological site. The study results to which we had access contained organ pathology data for 35 organs. Those organ sites are: adrenal, brain, bile duct, bone, clitoral, circulatory, heart, harderian, hematopoietic, intestine, integumentary, kidney, liver, lung, mammary, mesothelial, nose, nervous system, ovary, oral cavity, osteosarcoma, pancreas, parathyroid, preputial, pituitary, stomach, subcutaneous tissue, skin, spleen, thymus, thyroid, tunica vaginalis, uterus, urinary bladder, and Zymbal's gland.

Miscellaneous short-term tests. In addition to the *Salmonella* mutagenicity test, data were available for 65 of the 301 chemicals on some combination of other short-term tests that have been studied as surrogate predictors of toxicity by the NTP. Among these are tests of chromosomal aberration, sister chromatid exchange, and mutagenesis tests on mouse lymphoma cells and *Drosophila*.

The distribution of the 301 chemicals for which we had data is summarized in Table 1.

The Problem of Missing Data

The specific organ toxicity observed during a subchronic study is potentially one of the most reliable indicators of long-term effect. However, of the original training set of 301 chemicals, no subchronic organ

Subchron data	Carcinogenicity Classification			Total
	Positive	Negative	Equivocal	
Some	83 (56%) (52%)	39 (27%) (39%)	25 (17%) (58%)	147 (49%)
None	76 (49%) (48%)	60 (39%) (61%)	18 (12%) (42%)	154 (51%)
TOTAL	159 (53%)	99 (33%)	43 (14%)	301

Table 1: Distribution of the 301-Chemical Training Set

lesion data were available for 154 (51%). To investigate the effect of this missing data, we performed two groups of experiments: four experiments that trained on all 301 chemicals regardless of missing data, and four that trained only on those 147 for which at least some organ toxicity data was available. (It should be noted that even some of the 147 have missing organ toxicity data for between one and three of the four sex-species combinations.)

The Problem of Equivocal Classification

As mentioned above, a rodent carcinogenesis bioassay sometimes yields results that are equivocal overall, and in fact are so certified by the NTP peer-review process (Ashby et al. 1989; Huff & Haseman 1991). There is considerable debate in the predictive toxicology research community about the most appropriate way in which to handle such equivocal classifications. In particular, treating equivocal as simply a third classification does not appropriately capture all the information of interest. To gain some insight into this issue, in each group of experiments we performed four sets of train-and-test, using different treatments of the equivocal classification.

The Test Data

(Tennant et al. 1990) published predictions of the potential carcinogenicity of 44 chemicals still under study by the NTP at that time. NTP results for 36 of these 44 chemicals are now known. 18 (50%) of the 36 have been officially classified positive, 9 (25%) negative, and 9 (25%) equivocal. We took these 36 chemicals as our test set. Throughout our experiments, all aspects of the training and test set were mutually exclusive.

The Experiments

Following is a brief description of eight experiments, indicating the classification distribution of the training set used for each experiment.

Group 1. All 301 chemicals used.

Experiment 1.1.
Equivocal classifications included as such. Training set size 301 (53% +, 33% —, 14% E).

Experiment 1.2. Equivocal classifications reclassified as positive. Training set size 301 (67% +, 33% —).

Experiment 1.3. Equivocal classifications reclassified as negative. Training set size 301 (53% +, 47% —).

Experiment 1.4. Equivocal classifications eliminated from training set. Training set size 258 (62% +, 38% —).

Group 2. Chemicals eliminated for which organ toxicity data were entirely unavailable.

Experiment 2.1.
Equivocal classifications included as such. Training set size 147 (56% +, 27% —, 17% E).

Experiment 2.2. Equivocal classifications reclassified as positive. Training set size 147 (73% +, 27% —).

Experiment 2.3. Equivocal classifications reclassified as negative. Training set size 147 (56% +, 44% —).

Experiment 2.4. Equivocal classifications eliminated from training set. Training set size 122 (68% +, 32% —).

The Conduct of Each Experiment

Tree Induction

TIPT (Tree Induction for Predictive Toxicology) is a similarity-based classification system based on the tree and rule induction system C4.5 (Quinlan 1993). TIPT uses supervised learning over a training set of biological-activity attributes of chemicals to devise concept descriptions capable of successfully classifying unseen chemicals. Each chemical in the training set was preclassified according to the NTP assessment of its carcinogenesis in rodents.

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 (Quinlan 1986), which is computed as follows. Assume we have a set T of examples, each example belonging to one of k classes $C_1, \ldots, C_k$. 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 (Shannon 1949)

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

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

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

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

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

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

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

represents the information generated by the mere act of partitioning T into n subsets. Then $\mathcal{GR}(A) = \mathcal{G}(A)/\mathcal{P}(A)$ 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 $\mathcal{GR}$ over all untested attributes is used. This is the gain-ratio criterion. (The gain criterion, an older alternative, maximizes $\mathcal{G}$ rather than $\mathcal{GR}$ (Quinlan 1986).)

In the case of missing data (Experiments 1.1 through 1.4), $\mathcal{I}(T)$ and $\mathcal{I}_A(T)$ were computed using only known values, $\mathcal{G}(A)$ was computed as $K \times (\mathcal{I}(T) - \mathcal{I}_A(T))$, where K is the proportion of T with known values of A, and $\mathcal{P}(A)$ 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.

Rather than considering attribute values only singly, trees were generated in which the values of discrete attributes were grouped for testing information gain, but no significant improvement in accuracy resulted. Finally, the older information-gain criterion was tried for each of these combinations, as an alternative to the gain-ratio criterion, but generally yielded lower accuracy. For more details on decision tree induction, see (Quinlan 1986; Quinlan 1987).

Windowing

For each of the four Group 1 and four Group 2 experiments outlined above, approximately 20% of the training examples were selected using a stratified random procedure and assembled into a so-called "window." Selection for the initial window was stratified so that the distribution of positive and negative examples (and equivocal examples, depending on the experiment) was as uniform as possible. A decision tree was developed from this window, and this tree was then used to classify the training examples not included in the window. Half of the misclassified training examples were then added to the window and a second tree was generated. This cycle was repeated until the tree correctly classified all the training examples outside the window, or until classification accuracy ceased to improve.

Ten trees with maximum asymptotic accuracy were grown in this manner, each beginning with an initial window of the same size but different composition.

Converting Trees to Rules

Decision tree structures can be large, difficult for humans to understand, and can contain redundant subgraphs which hide the underlying structure of information. Production rules can avoid these difficulties. Therefore, after the tree induction phase produced ten trees from each experiment, a single set of production rules was generated from these ten 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 minimise false positives, and the class which contained the most training cases not covered by any rule was chosen as the default class.

RIPT (Rule Induction for Predictive Toxicology), also developed from C4.5, is a system which converts sets of trees into sets of production rules.

Results from the Experiments

Error rate was measured by the ratio of misclassified training examples to total training examples. The best-of-10 error rates for all the experiments are summarized in Table 2.

The tree with the lowest error rate on training examples of the 80 generated by the asymptotic windowing procedure is given in Figure 1. It was generated in Experiment 2.4, the experiment which showed the best combined tree and rule accuracy, i.e., lowest error rate, on the training set. It should be emphasised that *training set accuracy alone* was used as a selection criterion to facilitate fair comparison with other published predictions on the 36 test chemicals, which were prospective rather than retrospective. Experiment 2.4 eliminated from the training set all chemicals for which organ toxicity data were unavailable and all those with equivocal carcinogenicity classification, yielding a training set of 122 chemicals. Of these, the

Exp. No.	% +	% —	% E	Best-of-10 Err. % Tree	Rule
1.1	53	33	14	16.9	35.2
1.2	67	33	—	10.3	29.2
1.3	53	47	—	10.0	25.2
1.4	62	38	—	8.1	24.4
2.1	56	27	17	12.9	32.0
2.2	73	27	—	6.1	18.4
2.3	56	44	—	5.4	33.3
2.4	68	32	—	6.6	17.2

Table 2: Best-of-10 Training Error Rates

Most Accurate Tree			Corresponding Rules			
+	—	E	+	—	E	←classed as
17	1	0	14	4	0	NTP +
4	5	0	4	5	0	NTP —
6	3	0	6	3	0	NTP E
Sensitivity:		94.4%		77.8%		
Specificity:		55.6%		55.6%		
+ Predictivity:		81.0%		77.8%		
— Predictivity:		83.3%		55.6%		
Accuracy:		81.5%		70.4%		

Table 3: Confusion Matrices (Equivocal Classifications Omitted for Comparison Purposes)

windowing process made use of 93 to achieve convergence.

The rule set generated from the set of ten trees generated in Experiment 2.4 is given in Figure 2. The confusion matrices for the predictions generated by the most accurate tree and the rules generated by the corresponding experiment are shown in Table 3.

Table 4 shows the chemical-by-chemical outcome of TIPT/RIPT classification on the test set. Classification is accompanied by a confidence estimate, which is a subset of the unit interval in the case of trees, and a single number for rules. For completeness, predictions for all 44 chemicals are included, although classification is known for only 36 at this time.

Other Published Predictions

Although a classification accuracy of 70-80% is not especially high on many machine learning problems, it is nevertheless equal to or better than all but one published approach to the identical problem of predicting carcinogenesis for some or all of the original 44 chemicals. Methods that have been described in the literature include both human expert predictions, experimental measurements, and computer-based systems of various kinds. A brief description of these other published prediction methods follows; see the references for more details.

1. The initial prospective predictions on the 44 chemicals were based on human expert evaluation of chemical structural alerts, short-term *in vitro* toxicity, subchronic *in vivo* toxicity, and MTD dose level, for each chemical in isolation (Tennant et al. 1990). Aside from TIPT/RIPT, this was the only group to attempt predictions on all 44 chemicals from the original set.

2. (Bakale & McCreary 1992) experimentally measured electrophilic reactivity (k_e) values for individual chemicals and used these to predict carcinogenicity. This approach provided predictions on 31 of the 44 chemicals.

3. Deductive Estimation of Risk from Existing Knowledge (DEREK) (Sanderson & Earnshaw 1991) is a rule-based expert system derived from the LHASA chemical synthesis system. DEREK identifies chemical substructures in a molecule and relates these to likely types of toxicity. DEREK made predictions on 41 of the 44 chemicals.

4. Computer-Optimized Molecular Parametric Analysis of Chemical Toxicity (COMPACT) (Lewis, Ioannides, & Parke 1990) is a system that computes the shape and molecular orbital energy levels of a chemical structure and evaluates whether it can interact with the active site of cytochrome P450 I or to the binding site of the Ah receptor, and thereby induce cancer. COMPACT made predictions on 40 of the 44 chemicals.

5. MultiCASE (Rosenkrans & Klopman 1990) is a program that automatically selects chemical substructures that are statistically associated with biological activity in a training set of known active and inactive chemicals, and uses the presence or absence of these substructures in test chemicals to predict their effects. MultiCASE made predictions on 39 of the 44 chemicals.

6. Toxicity Prediction by Komputer-Assisted Technology (TOPKAT) (Enslein, Blake, & Borgstedt 1990) predicts effects by means of a linear Quantitative Structure-Activity Relationship (QSAR) regression model, developed by discriminant analysis of quantitative descriptors of molecular structure attributes and substructural fragments known to be associated with biological activity. TOPKAT made predictions on 28 of the 44 chemicals.

7. (Benigni 1991) used a QSAR model based on a combination of two activity descriptors, computed electrophilic reactivity (electrophilicity) and Ashby's structural alerts. This model made predictions for 39 of the 44 chemicals.

8. (Jones & Easterly 1991) used a method of relative potency analysis and the dose levels of the NTP bioassays to rank the potential strength of the 44 chemicals. The relative predicted outcome results

Chemical	NTP	TIPT	RIPT	T	BM	SE	LIP	RK	EBB	B	JE
CI Acid red 114	+	+ [.84-1.0]	+ (.9)	+	NP	+	+	—	NP	+	NP
1,2,3-Trichloropropane	+	+ [.84-1.0]	+ (.9)	+	+	+	—	+	+	+	PE
2,3-Dibromo-1-propanol	+	+ [.78-1.0]	+ (.9)	+	+	+	—	+	NP	+	+
3,3'-Dimethylbenzidine 2HCl	+	+ [.84-1.0]	+ (.9)	+	—	+	+	+	+	W/U	+
Pentachloroanisole	+	+ [.84-1.0]	+ (.9)	+	+	+	+	+	—	W/U	—
CI Pigment red 3	+	+ [.84-1.0]	+ (.85)	+	+	+	+	+	+	+	NP
o-Nitroanisole	+	+ [.84-1.0]	+ (.9)	+	+	+	+	+	—	+	+
CI Direct blue 218	+	+ [.84-1.0]	+ (.85)	+	NP	+	+	+	NP	W/U	—
Coumarin	+	+ [.84-1.0]	+ (.9)	+	+	—	+	—	NP	W/U	+
3,4-Dihydrocoumarin	+	+ [.54-1.0]	— (.64)	+	+	—	+	—	—	—	PE
CI Direct blue 15	+	+ [.84-1.0]	+ (.85)	+	NP	NP	+	—	+	W/U	NP
2,4-Diaminophenol 2HCl	+	+ [.84-1.0]	+ (.9)	+	—	—	+	+	+	W/U	+
Tris(2-chloroethyl)phosphate	+	+ [.76-1.0]	+ (.76)	—	+	+	—	+	+	W/U	—
Triamterene	+	+ [.84-1.0]	+ (.85)	+	NP	+	+	+	—	—	—
o-Benzyl-p-chlorophenol	+	+ [.61-1.0]	— (.64)	+	+	—	—	—	—	W/U	+
Mercuric chloride	+	+ [.84-1.0]	+ (.85)	+	—	NP	NP	NP	NP	NP	PE
Diphenylhydantoin	+	+ [.54-1.0]	— (.64)	+	—	—	—	—	NP	W/U	+
Naphthalene	+	— [.57-1.0]	— (.64)	—	—	—	+	—	NP	—	NP
Amphetamine sulfate	—	— [.37-.73]	+ (def)	—	NP	—	—	—	NP	—	—
Ethylene glycol	—	+ [.84-1.0]	+ (.85)	+	—	—	—	—	—	—	—
Promethazine HCl	—	— [.57-1.0]	— (.64)	—	—	—	+	—	NP	—	—
Resorcinol	—	+ [.61-1.0]	— (.64)	—	—	—	+	—	—	—	—
Monochloroacetic acid	—	— [.57-1.0]	— (.64)	—	—	+	—	—	—	+	—
4,4'-Diamino-2,2'-stilbenedisulfonic	—	— [.71-1.0]	— (.64)	—	NP	+	+	+	+	W/U	NP
Methyl bromide	—	+ [.78-1.0]	+ (def)	—	+	+	—	+	NP	W/U	NP
Sodium azide	—	+ [.78-1.0]	+ (.9)	—	NP	+	—	NP	NP	NP	—
p-Nitrophenol	—	— [.57-1.0]	— (.64)	—	+	+	+	—	—	+	—
Chloramine	E	— [.37-.73]	+ (def)	—	NP	—	+	NP	NP	W/U	NP
gamma-Butyrolactone	E	+ [.61-1.0]	— (.64)	—	—	—	—	+	—	—	—
Manganese sulfate monohydrate	E	— [.71-1.0]	— (.64)	—	NP	—	NP	NP	NP	NP	+
4-Hydroxyacetanilide	E	+ [.84-1.0]	+ (.85)	+	—	+	+	—	—	W/U	—
Titanocene dichloride	E	+ [.78-1.0]	+ (.9)	+	+	—	NP	NP	NP	NP	—
HC Yellow 4	E	+ [.78-1.0]	+ (.9)	+	NP	+	+	+[1]	—[2]	W/U	NP
Polysorbate 80	E	— [.71-1.0]	— (.64)	—	—	NP	NP	+	NP	NP	PE
p-Nitroaniline	E	+ [.78-1.0]	+ (def)	+	+	+	+	+	+	+	PE
CI Pigment red 23	E	+ [.78-1.0]	+ (def)	+	NP	+	+	+	+	+	NP
Tricresyl phosphate	?	+ [.84-1.0]	+ (.85)	—	+	—	+	+	NP	—	—
4,4'-Thiobis(6-t-butyl-m-cresol)	?	+ [.84-1.0]	+ (.85)	+	—	—	PE	+	—	—	—
p-Nitrobenzoic acid	?	+ [.84-1.0]	+ (.9)	—	+	+	+	—	—	+	—
Methylphenidate HCl	?	— [.57-1.0]	— (.64)	+	NP	+	—	—	—	—	+
t-Butyl alcohol	?	+ [.54-1.0]	+ (def)	+	—	—	—	—	—	—	—
2,2-Bis(bromomethyl)-1,3-propanediol	?	+ [.84-1.0]	+ (.85)	+	+	—	+	+	—	W/U	NP
Salicylazosulfapyridine	?	— [.71-1.0]	— (.64)	+	NP	+	+	—	+	+	NP
Theophylline	?	— [.57-1.0]	— (.64)	+	—	+	+	—	—	—	+
RESULTS											
(NTP equivocals treated as —)		25/36	22/36	28/36	16/25	18/33	18/32	17/31	13/21	19/31	19/27
Percent		69	61	78	64	55	56	55	62	61	70
(NTP equivocals omitted)		22/27	19/27	24/27	13/20	14/25	17/26	16/25	10/16	18/25	14/18
Percent		81	70	89	65	56	65	64	63	72	78

KEY:
NTP: Actual determination by National Toxicology Program long-term rodent bioassay
T: (Tennant et al. 1990) BM: (Bakale & McCreary 1992)
SE: (Sanderson & Earnshaw 1991) LIP: (Lewis, Ioannides, & Parke 1990)
RK: (Rosenkranz & Klopman 1990) EBB: (Enslein, Blake, & Borgstedt 1990)
B: (Benigni 1991) JE: (Jones & Easterly 1991)
+: prediction of carcinogenic effect —: prediction of no carcinogenic effect
E: equivocal classification NP: no prediction made
PE: prediction of equivocal effect W/U: weak positive or uncertain prediction
def: classified by default only ?: unknown result (bioassay result pending, June 1993)
1,2: Personal communication to D. Bristol, 1993.

Table 4: Comparison of TIPT/RIPT Results with 8 Published Prediction Methods/Systems

were applied as modulators of the original predictions (Tennant et al. 1990) to give Rapid Screening of Hazard (RASH) predictions on 27 of the 44 chemicals.

Aside from a common test set, these methods share very few characteristics with each other or with us; in particular, the groups were under no restriction to make use of the same set of chemicals or attributes that we did.

Preliminary results are available for three sets of human expert predictions; these average no more than 65-66% accurate. The most accurate human predictions (Tennant et al. 1990) are based on heuristics derived over time by the inductive analysis of large collections of data (Ashby & Tennant 1991; Ashby et al. 1989). Consequently, these heuristics reflect the biological nature of the characteristics employed, such as short-term *in vitro* mutagenesis assays, subchronic *in vivo* organ toxicity, and chemical substructure alerts for carcinogenicity. The remaining two sets of human predictions, significantly less accurate than those in (Tennant et al. 1990), are unpublished and are omitted from Table 4.

In contrast to these fundamentally inductive approaches, most other prediction methods (Bakale & McCreary 1992; Benigni 1991; Jones & Easterly 1991; Sanderson & Earnshaw 1991; Lewis, Ioannides, & Parke 1990) involve various deductive approaches to prediction that are based on the electrophilic somatic mutation model of carcinogenesis and employ one or more physico-chemical descriptors that presumably relate chemical structure to carcinogenic activity; their performance in the 44-chemical prediction experiment appears less accurate than that of the most accurate human experts. Two additional methods employ statistical analysis of exploratory data in an effort to identify chemical substructures or physico-chemical parameters that can account for the biological activity (Enslein, Blake, & Borgstedt 1990; Rosenkranz & Klopman 1990), but their performance was no better than those of the deductive methods.

Table 4 compares the results of earlier published predictions, chemical by chemical, with our TIPT/RIPT systems.

The Nature of Predictive Toxicology

The broad overall goal of predictive toxicology research is to develop methods that provide accurate predictions for as many chemicals as possible in the universe of structurally diverse, noncongeneric chemicals. Accordingly, the design of methods that can address this noncongeneric prediction problem must ideally avoid *a priori* restrictions of the concept space used to describe the biological activity being predicted so as to maximize the extent to which the method covers the universe of chemicals.

Our general approach to developing predictive toxicology methods is designed to use inductive approaches for recognizing patterns and relationships as an effective way to address the noncongeneric chemical prediction problem. We believe induction is the most appropriate approach that can be applied to the development of noncongeneric prediction methods, because it enables the discovery of relationships in knowledge domains that lack formal models; i. e., induction requires no specific knowledge of the multiple biological processes or mechanism(s) that determine relationships between the noncongeneric universe of chemicals and a complex biological endpoint such as carcinogenesis. Deductive approaches, on the other hand, require hypotheses that limit the set of chemicals to which they apply. Generally, the most successful applications of the deductive approach involve biological endpoints governed by a single mechanism that is known in detail. A good example is the design of drugs that inhibit the active site of an enzyme, the electronic and steric requirements of which have been fully characterized by x-ray or other physical methods.

By analogy to the inductive approach used by the most successful human-expert predictors, first we have adapted a supervised machine learning technique to address this problem. Second, we chose to use phenomenological biological descriptors to represent the essential attributes of chemicals for the induction analysis; attribute values were compiled from extensive NTP data generated through the exploratory testing of noncongeneric chemicals for toxicity endpoints such as mutagenicity, subchronic toxicity, and carcinogenicity. Various physico-chemical parameters that are often used to represent chemical structure in deductive analyses are readily available; we anticipate incorporating them into future experiments to evaluate their contribution to prediction accuracy. However, they were not utilized in this set of experiments, the better to assess the potential contribution of the more global, biological attributes of chemicals in a computer-based inductive analysis.

What Has Been Learned

1. TIPT/RIPT yielded a tree with overall classification accuracy (concordance) of 81% when the 9 chemicals classified by NTP as equivocal are omitted from the test set. The corresponding rule set showed accuracy of 70%. This compares quite favorably with other published predictions of the potential carcinogenicity of the test set chemicals.
2. (Tennant et al. 1990) and TIPT/RIPT both utilized primarily biological-activity parameters of the same type. TIPT predictions were the same as those made by the most accurate human predictors for 36 of the 44 chemicals (82%). RIPT generated many rule sets which "rediscovered" existing expert heuristics, such as the utility of microbial assays for certain chemicals or the importance of subchronic toxicity in some organs but not others. We consider the content of

these rules to provide useful confirmation that our approach is promising.

3. The seven published prediction methods that are based primarily on chemical-structure parameters were too specialised to handle some of the chemicals. Only the original set of human predictors and TIPT/RIPT, the system described in this paper, made predictions for all 44. Unlike most of the predictive methods in this area, rule induction using phenomenological data from biological tests can handle inorganic molecules as well as organic, noncongeneric organic molecules as well as congeneric, and mixtures.
4. Rule induction can be employed on data involving a wide variety of information: genotoxicity, short-term microbial studies, minimally-toxic dosages, subchronic lesions specific to an organ or organ system, and molecular structure and alerting substructures. In other words, TIPT/RIPT can utilise any parameter, enabling it to exploit a learning set containing biological as well as chemical data. In terms of its ability to use whatever information may be appropriate to its task, TIPT/RIPT resembles the human heuristic approach more than it does other published computer systems in this domain.
5. Precisely because they are uninformed of domain knowledge, pure inductive approaches are not biased toward one or another hypothesis about carcinogenesis.
6. Some of the rules that we discovered have never been enunciated by humans. To the extent that they have no apparent connection to hypothesised causal mechanisms of carcinogenesis, they are unlikely to be readily accepted *per se* by cancer researchers; however, they may stimulate the formation of new mechanistic hypotheses and further research.
7. Mechanisms of carcinogenesis almost surely involve a variety of factors, ranging from molecular structure, to metabolic factors, to the genotoxic effects of electrophilic chemicals on DNA, to the modulation of hormones or various receptors that regulate gene expression. For this reason, even inaccurate classification by induced rules can be useful to experts by illustrating what additional information would be necessary to repair the rule.
8. Rule induction output is readily understood by human experts in toxicology, since the rules use largely the same terms and concepts.
9. Rule induction is much faster and allows a broader variety of experimentation than many other forms of computer-based analysis.
10. Rule induction is likely to be suitable for pattern analysis and prediction of other biological endpoints, such as genotoxicity, immunotoxicity, teratogenicity, and organ-specific toxicity.

What Remains To Be Done

We consider TIPT/RIPT to be no more than a successful proof-of-concept that rule induction can play a role in understanding chemical carcinogenesis, and thereby to help point the way to future mechanism research as well as to more efficient testing of chemicals in rodents. Considerable experimentation with the content of examples and attributes in the training set, however, is need to understand the fullest potential of this approach.

The attributes in this domain, particularly those pertaining to organ-specific toxicity, possess considerable underlying structure, and as part of our ongoing efforts we are investigating modified tree-induction methods which can accommodate, for example, set-valued attributes without incurring inordinate computational cost. Another current limitation of this approach is its lack of a natural way to handle continuous or probabilistic classifications.

Purely inductive techniques are unlikely to be the entire solution to the larger problem of predictive toxicology. At the same time, however, it is unlikely that any single domain theory can be constructed that will constitute an adequate explanation for all the data on the universe of chemicals, let alone warrant confidence in its predictive accuracy. Little is known about causal mechanisms in toxicology, so the codification of domain knowledge is a hard problem, and such theories will necessarily be partial and uncertain. Detailed mechanisms that explain the causality of cancer remain largely a mystery even to expert researchers. We are convinced that a combination of rule induction and knowledge-based methods promise ultimately to make the most significant contribution in this field (Bahler 1992; Bahler & Craycroft 1990).

Besides rule induction, among the other techniques of long-term interest in this project are:

- unsupervised learning and/or clustering;
- explanation-based methods with incomplete or uncertain domain theories;
- various ways of hybridising artificial neural networks and intelligent systems.

Acknowledgements

Thanks are due to Dr. Ray Tennant and Stanley Stasiewicz of NIEHS for making the data available, and to J. Ross Quinlan for a prepublication version of the C4.5 system.

References

Ashby, J. and R. W. Tennant 1991. Definitive Relationships among chemical structure, carcinogenicity, and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutation Research* 257: 229-306.

Ashby, J., R. W. Tennant, E. Zeiger, and S. Stasiewicz 1989. 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. *Mutation Research* 223: 73-103.

Bahler, D. 1992. Methods of Decision Tree Induction. *4th North Carolina Symposium on Art. Intell. and Advanced Computing Tech.*, Raleigh.

Bahler, D. and R. Craycroft 1990. Learning from Examples in a Production System Language. *3rd North Carolina Symposium on Artificial Intelligence*, Research Triangle Park, NC.

Bakale, G. and R. D. McCreary 1992. Prospective k_e screening of potential carcinogens being tested in rodent bioassay by the U. S. National Toxicology Program. *Mutagenesis* 7(2): 91-94.

Benigni, R. 1991. QSAR prediction of rodent carcinogenicity for a set of chemicals currently bioassayed by the US National Toxicology Program. *Mutagenesis* 6(5): 423-425.

Enslein, K., B. W. Blake, and H. H. Borgstedt 1990. Prediction of probability of carcinogenicity for a set of ongoing NTP bioassays. *Mutagenesis* 5(4): 305-306.

Huff, J. and J. Haseman 1991. Long-term chemical carcinogenesis experiments for identifying potential human cancer hazards: Collective data base of the National Cancer Institute and National Toxicology Program (1976-1991). *Environmental Health Perspectives* 96: 23-31.

Jones, T.D. and C.E. Easterly 1991. On the rodent bioassays currently being conducted on 44 chemicals: a RASH analysis to predict test results from the National Toxicology Program. *Mutagenesis* 6(6): 507-514.

Lewis, D. F. V., C. Ioannides and D. V. Parke 1990. A prospective toxicity evaluation (COMPACT) on 40 chemicals currently being tested by the National Toxicology Program. *Mutagenesis* 5(5): 433-436.

Quinlan, J. R. 1986. Induction of Decision Trees. *Machine Learning* 1: 81-106.

Quinlan, J. R. 1987. Simplifying Decision Trees. *Int. J. Man-Machine Studies* 27: 221-234.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.

Rosenkrans, H. S. and G. Klopman 1990. Prediction of the carcinogenicity in rodents of chemicals currently being tested by the US National Toxicology Program: structure–activity correlations. *Mutagenesis* 5(5): 425-432.

Sanderson, D. M. and C. G. Earnshaw 1991. Computer prediction of possible toxic action from chemical structure: the DEREK system. *Human and Experimental Toxicology* 10: 261-273.

Shannon, C. E. 1949. *The Mathematical Theory of Communication.* Urbana, IL: University of Illinois Press.

Tennant, R. W. 1993. Stratification of rodent carcinogenicity bioassay results to reflect relative human hazard. *Mutation Research* 286: 111-118.

Tennant, R. W., J. Spalding, S. Stasiewicz, and J. Ashby 1990. Prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 44 chemicals by the National Toxicology Program. *Mutagenesis* 5(1): 3-14.

```
IF chemical has an alkyl ester
     structural alert
  THEN class is positive
ELSE IF there is subchronic pathology in
     female rat kidney
  THEN class is positive
ELSE IF chemical mutates Salmonella
  THEN class is positive
ELSE IF there is subchronic pathology in
     male rat brain
  THEN class is negative
ELSE IF there is subchronic pathology in
     female rat liver
  THEN class is positive
ELSE IF 187.5 =< adjusted max dose
     (mg/kg/day) < 900
  THEN class is positive
ELSE class is negative
```

Figure 1: The Most Accurate Tree Generated

```
IF chemical mutates Salmonella
 AND adjusted maximum rat dose =< 750
     (mg/kg/day)
  THEN class is positive CONFIDENCE 89.5%
IF there is subchronic pathology in
     female rat kidney
  THEN class is positive CONFIDENCE 84.6%
IF chemical has an alkyl ester structural
     alert
  THEN class is positive CONFIDENCE 75.8%
IF chemical has an aliphatic epoxide/
     aromatic oxide structural alert
  THEN class is positive CONFIDENCE 70.7%
IF chemical does not mutate Salmonella
 AND there is no subchronic pathology in
     male rat pituitary or spleen or
     urinary/bladder
 AND there is no subchronic pathology in
     female rat kidney
 AND chemical has no alkyl ester structural
     alert
  THEN class is negative CONFIDENCE 63.5%
class is positive BY_DEFAULT
```

Figure 2: The Rules from Experiment 2.4