Machine Learning, 30, 217–240 (1998)

Knowledge-Based Learning in Exploratory Science: Learning Rules to Predict Rodent Carcinogenicity

YONGWON LEE ylee@ict.atc.lmco.com
Lockheed Martin Missiles and Space, 3251 Hanover Street, H1-43, B255, Palo Alto, CA 94304

BRUCE G. BUCHANAN buchanan@cs.pitt.edu
Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260

JOHN M. ARONIS aronis@cs.pitt.edu
Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260

Editors: Ron Kohavi and Foster Provost

Abstract. In this paper, we report on a multi-year collaboration among computer scientists, toxicologists, chemists, and a statistician, in which the RL induction program was used to assist toxicologists in analyzing relationships among various features of chemical compounds and their carcinogenicity in rodents. Our investigation demonstrated the utility of knowledge-based rule induction in the problem of predicting rodent carcinogenicity and the place of rule induction in the overall process of discovery. Flexibility of the program in accepting different definitions of background knowledge and preferences was considered essential in this exploratory effort. This investigation has made significant contributions not only to predicting carcinogenicity and non-carcinogenicity in rodents, but to understanding how to extend a rule induction program into an exploratory data analysis tool.

Keywords: Rule induction, chemical carcinogenicity, semantic constraints, scientific discovery

1. Introduction

Knowledge discovery in real-world scientific data demands more software support than classification of instances or association of variables. To be useful, in addition to discovering relationships among variables, an inductive learning program must support the iterative process of scientific experimentation and allow investigators to compare results as they run the program under different assumptions and perspectives (Fayyad, Haussler & Stolorz, 1996a; 1996b). It must also support experimentation with different biases, including different criteria of what is interesting, and be tolerant of the incompleteness of real-world data. In this paper, we report on a multi-year collaboration among computer scientists, toxicologists, chemists, and a statistician, in which the RL rule induction program (Buchanan, 1994; Clearwater & Provost, 1990; Provost, 1992) was used to assist toxicologists in analyzing relationships among various features of chemical compounds and their carcinogenicity (i.e., propensity to cause cancer) in rodents (mice and rats). This investigation has made significant contributions not only to predicting carcinogenicity and non-carcinogenicity in rodents, but to understanding how to extend a rule induction program into a knowledge discovery tool.

Over 1,000 chemicals are released into the environment each year, with little or no knowledge about their carcinogenic effects on humans. Extensive efforts in this area have not yielded a satisfactory set of criteria for distinguishing carcinogens from non-carcinogens.

In fact, this is not surprising considering the complexity of carcinogenic events, i.e., multiple stages and multiple causes, the lack of a well-established theory and data, the diversity of chemical structures, and the interactions and transformations among chemicals.

The cancer hazard identification problem provides an interesting opportunity for knowledge discovery. It illustrates the need for a data analysis tool to explore relationships between a large number of chemical features and carcinogenic activity. In addition, just like other exploratory scientific problems, it requires tools that are capable of dealing with various questions and experimental assumptions that are difficult to formulate. Such a tool must also allow analysts and scientists to explore the consequences of many different sets of assumptions, their intuitions and prior knowledge. While most learning programs offer a variety of heuristic search methods to find interesting patterns, not all of them are flexible enough to be effective data analysis tools suitable to scientific problems.

In the work reported here, the toxicologist working with us, Professor Herbert Rosenkranz, wished to look for nonlinear relationships using both symbolic and numeric variables. While the toxicologist's ultimate goal of the investigation was to find more predictive criteria for identifying carcinogens and non-carcinogens in rodents, he also wished to investigate several hypotheses about the value of incorporating particular features of chemicals into a predictive model. Thus, rather than simply applying the RL induction program to a database of chemicals and inducing a set of classification rules, RL was used as a data analysis tool which was applied iteratively to assist scientists in testing their intuitions and assumptions.

Our investigation consisted of three studies conducted over three years. In the first two studies, our efforts were concentrated on developing more predictive and semantically plausible criteria for identifying rodent carcinogens and non-carcinogens. Different databases of chemicals and different features of chemicals were used in each study. In the third study, the RL program was used to analyze the predictions of other induction programs to find strategies for interpreting and combining predictions of multiple induction programs. Major steps of our investigation included: (1) preparation of data according to the problem to be solved, (2) application of the RL program, (3) evaluation of the results of experiments with the experts to assess the semantics of rules, as well as their predictive strength, and (4) revision of semantic assumptions, constraints of the problem, and syntactic induction biases. Our approach is similar to published steps in the knowledge discovery process (Fayyad, Piatetsky-Shapiro & Smyth, 1996a; 1996b).

One of the important steps in knowledge discovery is the interpretation and evaluation of patterns produced by an inductive learning component. The evaluation of discovered patterns must be done not only by data coverage metrics such as accuracy, but also by their concordance to domain semantics. In our investigation, rule sets learned in each experiment were evaluated by collaborating experts, and the outcome of evaluations was then used to either revise previously used semantic assumptions and syntactic search biases, or create new constraints in subsequent experiments. While a complete semantic model may be difficult to incorporate, some degree of semantics can be included to guide rule formation during rule search. Towards this end, we extended the RL program to use some semantics about attributes and their values during rule formation. We believe these modifications to RL are general and are applicable both to other learning programs and in other application domains.

The results of our studies are very encouraging. The rule sets learned have demonstrated their predictive strength by identifying rodent carcinogenicity and non-carcinogenicity of new sets of chemicals more accurately than other methods (Parry, 1994). In addition, through testing various semantic assumptions and comparing results among experiments, several interesting discoveries were made about relationships between features of chemicals and carcinogenicity.

In the following section, a brief description of the problem of identifying chemical carcinogenicity is provided. Our investigation consisted of three studies which are described in Section 3 along with the results and evaluations. Further details of our studies can be found in Buchanan and Lee (1995), Lee, Buchanan, Klopman, Dimayuga, and Rosenkranz (1996), Lee, Buchanan, Mattison, Klopman, and Rosenkranz (1995), and Lee, Buchanan, and Rosenkranz (1996). The RL induction program is described in Section 4. Section 5 provides a summary and conclusions.

2. Background

2.1. Chemical carcinogenesis in rodents

Over 80,000 chemicals are available commercially, most with little or no data about their long-term health effects. Despite considerable research, the health effects of only about 15% of those chemicals are known, and definitive studies of the propensity to cause cancer in humans have been done for only about 50 chemicals. The small number of definitive studies is a result of the time and cost involved in large-scale epidemiological studies, which are necessitated by the ethical principle of not experimenting directly on humans. Also, carcinogenic chemicals may show no immediate effects, further necessitating long-term epidemiological studies that could definitively demonstrate effects in humans.

The U.S. regulatory decisions of chemical carcinogenicity are primarily based on long-term animal bioassays conducted by the National Toxicology Program (NTP), in which live animals (mostly rodents) are exposed to a chemical and the effects on various organs are observed. However, the long-term animal bioassays themselves are time-consuming (at least two years) and very costly (at least $2 million per chemical). Furthermore, not all chemicals are routinely subjected to long-term animal bioassays. Thus, the carcinogenic effects in rodents are known only for a relatively small number of chemicals, which may not constitute a representative set of all chemicals available. Up to this date, the NTP database contains only about 340 chemicals with a panel's assessment of their carcinogenicity based on results of two-year (long-term) rodent studies. About 65% of the 340 chemicals are judged to be likely carcinogens by the panel.

Many scientists have been investigating alternative ways to identify chemical carcinogenesis using various features of chemical compounds. These include 2- or 3-dimensional molecular structures, physical properties, results of various kinds of *in vitro* and *in vivo* short-term assays, and observations of organ-specific toxicity from subchronic (13–week) animal studies. However, no single type of information has yet been proved effective and predictive of chemical carcinogenicity, and relationships among various chemical features themselves have yet to be explored thoroughly. For example, the extent to which a battery of short-term assays and other cost-effective features serves as a useful predictor for carcino-

genicity has not been fully explored. Structural features alone have not proved sufficient for predicting carcinogenic activity.

Also, due to the bias of currently available data, as more chemicals are tested in long-term bioassays and their results made available, there is increasing chance that previously discovered and accepted criteria for identifying carcinogens and non-carcinogens will be contradicted and their predictive strength decreased. For example, in the early 1970's, the responses in *Salmonella* short-term assays, which test whether a chemical damages the DNA of the bacterium *Salmonella*, were more than 90% accurate for predicting carcinogenicity in rodents (Ames, Durston, Yamasaki & Lee, 1973). That is, more than 90% of chemicals causing DNA-damaging effects were rodent carcinogens, and more than 90% of chemicals causing no damaging effects on DNA were rodent non-carcinogens. However, currently the concordance between the responses in *Salmonella* assays and rodent carcinogenicity is at best 65% due to broader testing over more classes of chemicals. In fact, the value of using short-term assays has been questioned in print: "If current *in vitro* short-term assays are expected to replace long-term rodent studies for the identification of chemical carcinogens, then that expectation should be abandoned" (Tennant *et al.*, 1987).

2.2. Prior work

The cancer hazard identification problem has not been investigated by the AI community until recently. The CASE (Klopman, 1984) and MultiCASE (Klopman, 1992) structure-activity analysis programs, which were developed by chemists, are probably the first programs which used an automated inductive approach to the cancer hazard identification problem. Both programs explore a very large space of chemical substructures (typically more than 20,000 structural fragments) derived from training data and find two sets of structural fragments, one containing fragments common in carcinogens, and the other containing fragments common in non-carcinogens. The discovered structural fragments are used in two different ways to make carcinogenicity predictions. First, using Bayesian decision analysis, the predictive strengths of individual fragments are measured and combined. Second, CASE and MultiCASE generate quantitative models from the discovered structural fragments and physical-chemical properties.

Bayesian decision analysis also forms the heart of the CPBS (Carcinogenicity Prediction and Battery Selection) approach (Pet-Edwards, Haimes, Chankong, Rosenkranz & Ennever 1989), which helps decision makers to select the best set of bioassays and to interpret the results of a battery of tests. The CPBS approach also utilizes other methods such as cluster analysis, dynamic programming, and multi-objective decision making.

Probably the first application of a general purpose induction program to the cancer hazard identification problem is described by Bahler and Bristol (1993), who used C4.5 (Quinlan, 1993) to build a decision tree to classify rodent carcinogens and non-carcinogens using a variety of chemical information (189 attributes). More recently, Busey and Henry (1995) reported the use of a neural network to predict rodent carcinogenicity using organ-specific toxicity data. King and Srinivasan (1996) applied an inductive logic programming approach to make predictions of rodent carcinogenicity using molecular structures of chemicals.

While not an inductive approach, DEREK (Sanderson & Earnshaw, 1991) is a rule-based expert system designed to cover the broad field of mammalian toxicity including

carcinogenicity. DEREK makes predictions based on chemical substructures responsible for toxicological effects.

3. Study overview

3.1. The objective of the study

The objective of our study was to determine whether a rule induction program can provide meaningful assistance in exploratory science. In particular, scientists asked whether the RL induction program could analyze relationships among various features of chemical compounds to find more accurate criteria for *directly* and *indirectly* predicting rodent carcinogenicity. Three separate investigations were done, each with different datasets and purposes. In the first and second investigations, the RL induction program was used to find accurate rules that can *directly* make predictions of rodent carcinogenicity, using features cheaper than the 2-year long-term assays. The two investigations were different because different databases and chemical features were used. Further details of these two investigations can be found in Lee, Buchanan, Mattison, Klopman, and Rosenkranz (1995), and Lee, Buchanan, Klopman, Dimayuga, and Rosenkranz (1996).

In the third study, RL was applied to a dataset generated by two structural analysis programs, CASE and MultiCASE, to aid the CASE and MultiCASE programs in making more accurate predictions. That is, RL was used to learn rules to *indirectly* improve the predictive strength by aiding the CASE and MultiCASE programs. The CASE and MultiCASE programs have demonstrated their abilities to identify structural fragments associated with carcinogenic and non-carcinogenic events (Rosenkranz, Frierson & Klopman, 1986; Rosenkranz & Klopman, 1990b; 1995). However, predicting the carcinogenicity of new chemicals using the identified structural fragments has been problematic because neither program makes concise predictions. A prediction of the carcinogenicity of a chemical by CASE and MultiCASE usually results in twelve predicted quantities, called probabilities and potency values. An expert must arrive at a final conclusion in an implicit and sometimes inconsistent manner due to the lack of policies for combining these multiple predictions. That is, the expert makes a prediction of carcinogenicity, not only using probabilities and potency values predicted by CASE and MultiCASE according to his or her intuition and knowledge, but also relying on structural fragments and other chemical features such as results in short-term assays. Our goal was to find policies and strategies for interpreting and combining the predicted quantities generated by the CASE and MultiCASE programs.

3.2. The RL program

The RL learning program (Buchanan 1994; Clearwater & Provost, 1990; Provost, 1992) is a descendant of Meta-DENDRAL (Buchanan & Feigenbaum, 1978). RL generalizes the chemistry-specific aspects of Meta-DENDRAL and is a general purpose rule induction system. During our study, RL has been extended to accommodate our need to explore the databases in several ways, as discussed in the later sections.

RL uses heuristic search to generate IF-THEN rules and evaluates each of them against a set of data. Rules can be used individually, but the entire set of rules forms a disjunctive

class description. RL performs a general-to-specific search of the space of rules defined by conjunctions of attribute-value pairs (features). The goal of RL's search is to find rules that satisfy user-defined performance criteria. Each rule has a set of conditions and a predicted class, which RL evaluates statistically. The space of possible rules includes all possible combinations of conditions, so the size of the search space grows exponentially with the allowable number of conditions in a rule. However, RL uses a beam search to ensure that the time complexity of the search is linear in the number of conditions. The evaluation function and beam width are defined by the user. Each rule is tested against the entire set of data to calculate performance statistics (cf., RISE (Domingos, 1994)), and does not recursively partition training data (cf., C4.5), nor eliminate training data as learning proceeds (cf., CN2 (Clark & Niblett, 1989)).

A unique feature of the RL program is the Partial Domain Model (PDM) which provides flexibility in rule searching and evaluation. The PDM can be considered as a set of constraints that are used to guide the search through the space of rules. The information in the PDM includes the definitions of attributes (types, values), a list of classes, constraints on rule size and content, and other search biases. The constraints and domain knowledge usually take the form of preference criteria of desirable properties of rules to be induced.

The RL program has been applied to several real world problems including identification of human developmental toxicity (Gomez, Lee & Mattison, 1993; 1994), trigger design in high energy physics experiments (Lee & Clearwater, 1992), sensitivity analysis of rule performance in high energy physics (Clearwater & Lee, 1993), detecting and locating faults in a telecommunication network (Danyluk & Provost, 1993), analyzing large quantities of data on infant mortality (Provost & Aronis, 1996), inducing rules for biological macro-molecule crystallization (Hennessy, Gopalakrishnan, Buchanan, Rosenberg & Subramanian, 1994), and predicting pneumonia outcome (Cooper *et al.*, 1997), and fraud detection (Fawcett & Provost, 1997).

3.3. Materials

In addition to the RL induction program, our study included three databases of chemical compounds, and two chemical structural analysis programs. The three databases were (1) an NTP (National Toxicology Program) database of 301 chemical compounds, (2) an NIEHS (National Institute of Environmental Health Sciences) database of 108 compounds, and (3) a CPDB (Carcinogenicity Potency Data Base) database consisting of 1300 chemical compounds.

Based on organ pathologies observed at the end of long-term (two-year) rodent studies, the chemical compounds in each database were classified by a panel of experts into three classes: rodent carcinogens, rodent non-carcinogens, and equivocal. We decided not to include chemicals in the equivocal class because a chemical may be classified as equivocal not only because its biological evidence was not convincing, but also because the panel of experts did not agree on its carcinogenicity.

Both the NTP and CPDB databases contained chemical features such as *in vitro* or *in vivo* short-term effects, physical properties, and structural features. The NIEHS database contained much more detailed information on biological effects, with data on organ-specific toxicity. Because the chemicals in the NIEHS database already had been tested for rodent

carcinogenicity in long-term bioassays under the aegis of the NTP, the NIEHS database presents a novel opportunity to explore the relationship between organ-specific toxicity and carcinogenicity in rodents.

In the third study, RL used a database of predictions generated by two Structure–Activity Relationship (SAR) programs, CASE and MultiCASE, which used the NTP and CPDB databases to train their models, as shown in Figure 1.

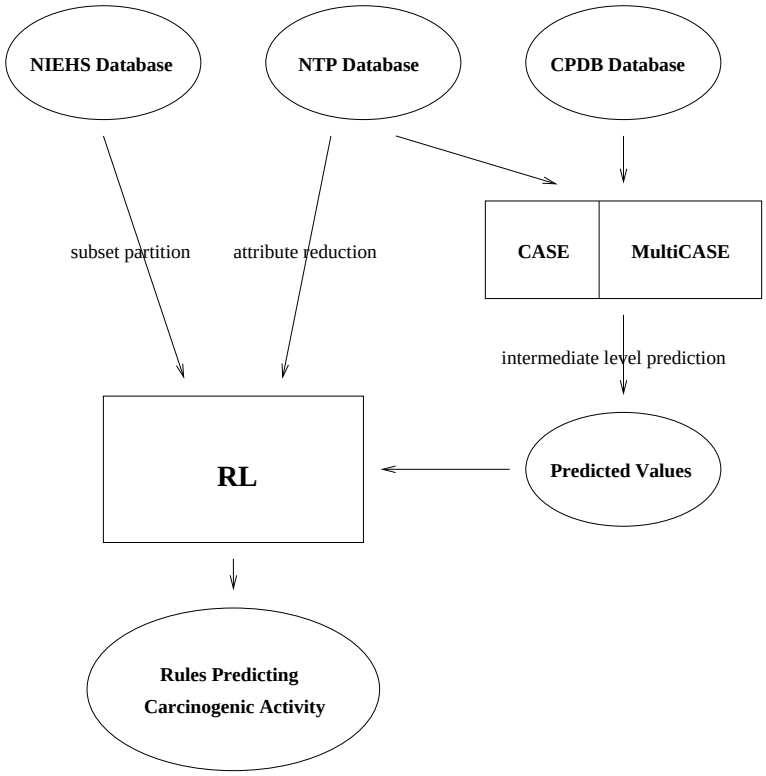

Figure 1. Databases and programs used in our study. The arrows indicate the flow of data along with data preprocessing steps: (1) examining a subset of the NTP data, (2) reducing a large feature set associated with the NIEHS data, and (3) using values predicted by other programs as input to RL.

3.4. Data preprocessing

In each of the three investigations, the original data were preprocessed before being given to RL in order to obtain a better analysis based on the semantics pertaining to chemical carcinogenicity. Thus, preprocessing was not simply a matter of reformatting and syntactic translations. The data preprocessing in our study included (1) data partitioning, (2) attribute reduction, and (3) intermediate level prediction.

3.4.1. Data partitioning The NTP database contained over 70 attributes, many of which were predicted rather than observed or experimentally measured. After excluding the predicted attributes, a total of 13 attributes were selected consisting of 7 short-term assays (including the *Salmonella* mutagenicity assay), 4 physical-chemical properties, and 2 dose measurement attributes.

The NTP database was partitioned into two databases, one containing chemicals with positive responses, and the other containing chemicals with negative responses. As more chemicals have been tested and have been included in the standard databases, the *Salmonella* assay's apparent reliability as a predictor of carcinogenicity has decreased from over 90% acuracy to less than 65% accuracy. For the NTP database, the *Salmonella* assay's accuracy is now only 64%. Further analysis has shown that the decrease is due mainly to the inclusion of more chemicals for which the *Salmonella* assay was negative (i.e., they had no DNA-damaging effects), but which turned out to be carcinogens. Specifically, while 80% of the chemicals that are positive in the *Salmonella* are rodent carcinogens, only 50% of the chemicals that are negative in *Salmonella* are non-carcinogens. So, rather than searching for rules that are more predictive over the entire database, we chose to explore the problematic subclass containing chemicals that are negative in the *Salmonella* assay. For these, no reliable indicator of carcinogenicity exists.

3.4.2. Attribute reduction The NIEHS database contains data on the organ-specific toxicity of chemicals observed at the end of 13-week subchronic animal studies. The organ-specific toxicity of a chemical is described by the presence or absence of a total of 124 lesions, each of which is a morphological effect in an organ. Overall, the 124 lesions specify 43 morphological effects on 32 organs. Because the lesions were observed in both sexes of two species, mice and rats, there were 496 Boolean attributes for each chemical. Overfitting is obviously a danger because there are only 108 compounds in the NIEHS database. Thus, reorganizing the attributes was necessary to reduce the number of attributes for an effective analysis of relationships between organ toxicity data and rodent carcinogenicity.

After discussions with experts, the 496 attributes were aggregated into 75 Boolean attributes. The organ toxicity data were pooled irrespective of species and gender, thus reducing the number of attributes to 124, because we were interested in analyzing relationships of organ toxicity to carcinogenicity in rodents rather than to carcinogenicity in a specific gender of a specific species.

We further generalized the toxicity data by separating morphological effects from organ-specificity, thus further reducing the number of attributes to 75 (43 morphological effects and 32 organs). For example, the presence or absence of degeneration in the kidneys (a lesion attribute) was generalized to the presence or absence of degeneration in any organ (a morphological-effect attribute) and the presence or absence of any type of morphological effect in the kidney (an organ-specificity attribute). While there was a loss of information by generalizing lesion attributes, the generalized attributes were more suitable to our study, because our purpose was not to address the correlation of organ-specific toxicity with organ-specific carcinogenicity. This has been studied by a number of other investigators. Also, there appears to be no correlation between the site specificity of carcinogenicity in rodents

and humans, since a carcinogenic effect anywhere in rodents might indicate a carcinogenic risk to humans.

3.4.3. Intermediate prediction In the third investigation RL used as features the probabilities and potency values predicted by CASE and MultiCASE. Three datasets were required: (1) a training dataset of chemicals on which CASE and MultiCASE were trained, (2) a test dataset of chemicals on which CASE and MultiCASE made predictions, and (3) a set of probabilities and potency values generated by CASE and MultiCASE, which RL used to induce rules during cross-validation experiments.

The NTP and CPDB databases were chosen as training datasets for CASE and MultiCASE because the two programs had previously used those two databases for training. For the test dataset, about 60 chemicals that were common to the NTP and CPDB databases were selected. To make predictions on each test chemical and generate a dataset of probabilities and potency values, CASE and MultiCASE were trained on the rest of the chemicals in the NTP and CPDB databases separately. For each test chemical, a total of 24 quantities were generated by CASE and MultiCASE. Thus, the dataset which RL used consisted of 60 chemicals, each with 24 numeric measurements.

While the number of test chemicals that RL used was small, it was unavoidable because not many chemicals were common to both databases. Furthermore, training CASE and MultiCASE takes a great deal of time and human effort, since they explore a very large number of structural fragments (usually more than 20,000 structures). A very large number of cross validation experiments were required to make predictions for all chemicals in the test set, as the CASE and MultiCASE programs were trained on each database separately to predict carcinogenicity in rats, mice, and rodents. The details about generating the intermediate predicted values can be found in Zhang (1995).

3.5. Semantic constraints

One of the key features of our study was to incorporate domain semantics during rule search to generate rules which were more biologically plausible and that satisfied experts' intuitions. For example, toxicologists prefer rules that predict chemicals are carcinogenic based on positive responses in short-term assays, rather than on ranges of molecular weights. The former is more plausible and more easily justified than the latter. Similarly, general intuition is that positive responses in short-term assays are plausible bases for predicting carcinogenicity (because positive responses indicate damage to DNA), while negative responses in short-term assays should be bases for non-carcinogenicity (because negative responses indicate no DNA-damaging effect). However, without this semantic constraint, RL would find rules based only on their example coverages, resulting in rules contradicting the background knowledge. For example, a rule may use positive responses in short-term assays (thus, DNA-damaging effects) to classify a chemical as non-carcinogenic. Of course, as mentioned previously, there are chemicals whose carcinogenic activities contradict the general background knowledge. Thus, for such chemicals, a preferred rule would use attributes other than short-term assays to explain them.

The complete semantics of carcinogenic activity are difficult to incorporate due to the complexity of the domain and the lack of biological knowledge. However, some semantics can be incorporated so that rules look more plausible and are easier to justify. In particular, we incorporated the following semantics of the usage of attributes and their values in rules:

- Some values of an attribute can only be used in rules predicting a certain class. For example, positive response in the *Salmonella* assays can only be used in rules predicting rodent carcinogenicity.
- Every rule must use one or more attributes from a specified list. For example, rules using one or more short-term assays (in addition to other features) are easier to understand than rules using physical-chemical features only.
- Types of ranges of certain numeric attributes are constrained. For example, when probability attributes are used in rules to predict rodent carcinogenicity, they should be of the form ($A > x$), where A is a probability attribute, $>$ means "greater than," and x is a cut-off threshold. Higher predicted probabilities by CASE and MultiCASE mean higher likelihood of carcinogenic activity. (However, it was noticed that without this constraint, RL would find rules using the probability attributes in the opposite manner.)

These semantic constraints were not hard-coded into the RL program. Rather, we developed several predicates in RL's Partial Domain Model (PDM) by which semantic constraints can be specified. These are described in Section 4. Note that in our study the semantic constraints took priority over other constraints, including performance requirements.

3.6. Experimental methods

Our investigation was exploratory, by which we mean that various assumptions and different values of RL's bias parameters were explored. In each of the three studies, a number of cross-validation experiments were done. Every cross-validation experiment was automated, but the comparisons of results among multiple cross-validation experiments were done manually. It should be noted that the main purpose of multiple cross-validation experiments was not to tune RL's bias parameters, but to explore the data with different assumptions and different sets of attributes. We were aware that given relatively small datasets, even cross-validation experiments, if repeated, could lead to overfitting the data. The biases that were varied among experiments are listed below in the order of the most varied to the least varied:

- **Attributes**: In many experiments, subsets of attributes were used. For example, in the first study, each of the possible subsets of available short-term assay attributes was tested in each experiment to measure and compare the predictive strength of batteries of short-term assays.
- **Performance requirements**: Rule performance requirements such as minimum rule coverage were also varied.
- **Semantic constraints**: Constraints on attributes and their value usages in rules and on feature combinations were varied.

- **Evaluation functions**: Evaluation functions for a rule and for a set of rules were varied.
- **Miscellaneous**: Miscellaneous parameters included flags such as whether to find an exception range for a numeric feature, and the number of splits.

In the first two studies, the available data were divided randomly into ten mutually exclusive sets of approximately equal size. A set of rules was learned from nine sets and tested on the remaining set, yielding ten trials of learning and testing. In the third study, due to the relatively smaller database available, we used a leave-one-out test, in which a set of rules was learned from all but one of the training examples and tested on the hold-out example. This procedure yielded as many trials as there were instances in the dataset.

In each trial, the following four performance statistics were measured to capture the predictive strength of a set of rules learned:

- **Sensitivity**: The number of carcinogens that are predicted correctly, divided by the total number of carcinogens for which predictions are made.
- **Specificity**: The number of non-carcinogens that are predicted correctly, divided by the total number of non-carcinogens for which predictions are made.
- **Accuracy**: The number of correct predictions divided by the total number of predictions.
- **Generality or Prediction Rate**: The number of chemicals for which predictions are made divided by the total number of chemicals in the test set.

Each of the above four quantities was measured in two ways depending on whether to use default predictions or not. When the learned rules failed to make any predictions for a chemical, its class was predicted according to the chemical's response in the *Salmonella* assay. That is, a chemical is predicted to be carcinogenic if it is positive in the *Salmonella* assay, and non-carcinogenic if it is negative in the *Salmonella* assay. However, the experts were reluctant to make any default predictions. This reluctance is due, in part, to the high cost of missing potential carcinogens. Thus, in all of the experiments we classified unpredicted chemicals as "unknown" rather than using a default prediction. We measured the performance with default predictions only for reference.

Making default predictions when a learned model fails has become general practice in Machine Learning research. However, default predictions must be used with care especially when the effect of default predictions is detrimental to the predictive strength of learned models. For example, Schaffer (1994) acknowledged that increasing rule generality does not always avoid overfitting. However, a number of experiments in different domains shows that increasing the rule generality requirement does not decrease the predictive accuracy of a learned rule set (Lee, 1995). On the other hand, increasing the rule generality requirements does decrease the generality of a rule set (i.e., the learned rule set makes fewer predictions). The predictive accuracy decreases substantially only when default predictions are used in addition to the learned rule set.

The rule sets were also evaluated semantically. While true biological evaluations would require extensive laboratory experiments, sets of rules were evaluated by our collaborating scientists. The semantic evaluations mostly concerned the "appropriate" use of attributes and their values in rules. For example, they would ask, "Which attributes and combination

of attributes are meaningful in predicting carcinogenicity? Which attribute values are meaningful in this problem?"

3.7. Results

The results of our experiments were encouraging, and the experts were enthusiastic about the performance of the learned rule sets. In the experiments using short-term assays together with dose measurement values and physical-chemical properties, the learned rule sets were 70% accurate for predicting carcinogenicity of chemicals whose responses in the *Salmonella* assays were negative. This is a marked improvement over current predictions using only the results of short-term assays (50% accuracy) or using only the physical-chemical properties (50% accuracy). It also exceeds the accuracy of human experts who used the whole spectrum of acute and subchronic toxicity results and structural properties as well as human knowledge and intuition.

Note that in the whole NTP database, 80% of chemicals that are positive in the *Salmonella* assays are carcinogenic in rodents. When this rule is combined with rule sets learned to predict rodent carcinogenicity of chemicals which are negative in the *Salmonella* assays, over 76% accuracy in predicting rodent carcinogenicity is obtained. This also was a significant improvement (at greater than .99 significance level) over the 71% accuracy obtained in our initial study (Ambrosino, Lee, Rosenkranz, Mattison, Buchanan, Provost & Gomez, 1993). Also, the learned rules were more plausible than those learned earlier, although some rules still remained to be explained.

We also observed interesting patterns of performance of rule sets while testing various batteries of one or more short-term assays and comparing the average performance measures (accuracy, sensitivity, specificity, prediction rate) of rule sets across multiple cross-validation experiments. In particular, we discovered that rule sets learned in the experiments using two particular short-term assays (SCE and ChrA) exhibited 17% to 19% higher sensitivity (more accurate prediction of rodent carcinogens) than did rule sets learned in the experiments which excluded the two short-term assay attributes. Also, we discovered another pair of short-term assays which were good for predicting rodent non-carcinogenicity. The rule sets learned with the pair showed 10% higher specificity than rule sets learned without the pair.

In the second study, rule sets learned were 81% accurate on average, with 83% sensitivity and 82% specificity. The 81% accuracy was significantly higher than the 74% accuracy obtained using the standard battery of liver and kidney tests. Also, the responses in the *Salmonella* assays were only 60% accurate for the chemicals used in the NIEHS database. Although our study included only 88 chemicals, the results provided at least a partial answer to several questions. We found that the majority of lesions had no or very little relationship to rodent carcinogenicity, although this may be in part due to the lack of sufficient data. We also found that there was little relationship between the responses in *Salmonella* assays and toxic effects in livers or kidneys. A more detailed description of results from the second study can be found in Lee, Buchanan, Klopman, Dimayuga, and Rosenkranz (1996).

In the third study, the primary focus was on the attributes used in rules (and the data used to generate rules) rather than on the performance of rule sets. That is, experts were interested in which predicted quantities were used in rules, which cut-off points were used

in rules, whether CASE or MultiCASE generated more predictive quantities, and which of the NTP and CPDB databases provided better training data. The results showed that using all predicted quantities based on both databases at the same time was less predictive of rodent carcinogenicity than using quantities based on a single database. Presumably this result is due to different conventions and standards used in collecting the two datasets, but, whatever the reason, it raises a cautionary flag for combining datasets in data mining.

Our experimental results also showed that the CPDB database provided better predictive models than the NTP database. In addition, the results showed that using both CASE and MultiCASE programs does not provide more predictive power than using one of them. In fact, rules learned using the probabilities and potency values generated by the CASE program were more accurate than those learned using values from the MultiCASE program.

When probabilities and potency values are generated using the same database, and given to RL as new features, the probabilities were more predictive of carcinogenicity than the potency values. For example, using the probabilities predicted based on the models learned from the NTP database, RL's rule sets were 73% accurate in predicting rodent carcinogenicity. However, when both probabilities and potency values based on the NTP database were used, learned rule sets were only 59% accurate. On the other hand, when RL used the probabilities and potency values generated by CASE or MultiCASE, but not both, rules learned with the potency values were more accurate than those learned with the probabilities. Complete results of the study can be found in Lee, Zhang, Sussman, and Rosenkranz (1995).

In addition to RL's rule induction method, we also applied a k-NN (k-nearest-neighbor) instance-based algorithm (Aha, 1989) and the C4.5 decision tree learning program (Quinlan, 1993). We selected the k-NN because the similarity of physical features (e.g., molecular weight, water solubility, etc.) would indicate structural similarity, which in turn would indicate similar carcinogenic activity. However, the results were discouraging. The k-NN method yielded at best 60% accuracy, even with the short-term assays. We also applied the C4.5 program, varying its parameters, pruning significance, etc. In cross-validation experiments, learned trees on average yielded 63% accuracy. This performance is similar to the concordance of the responses in the *Salmonella* assays. Note that we did not use the kinds of semantic constraints given to RL in either k-NN or C4.5, because it is hard to impose such semantic constraints on attributes or feature combinations.

3.8. Evaluation

To evaluate the predictive strength of the rule sets learned with the short-term assay attributes, we selected one final rule set and tested it on a new set of 24 chemicals. These 24 chemicals were among the 44 chemicals (Tennant, Spalding, Stasiewicz & Ashby, 1990) which formed the subject of the International Workshop of Predicting Chemical Carcinogenesis in Rodents (Parry, 1994; Ashby & Tennant, 1994), a comparative study similar to other comparative "bake-offs" in AI. Based on the results of 2-year rodent cancer bioassays, a scientific panel assigned 8 of the 24 chemicals as carcinogens, 7 as non-carcinogens, and 3 as equivocal. For 6 chemicals, carcinogenicity was listed as unknown even after the long-term assays. Since the rules were learned to discriminate carcinogenic from non-carcinogenic compounds, we omitted the 9 chemicals for which carcinogenicity is either equivocal or unknown. We

Table 1. Summary of predictions of 15 (non-genotoxic) chemicals, made by a rule set learned by RL and other methods (Study 1). The 15 chemicals were negative in the *Salmonella* assay and formed the subject of the International Workshop of Predicting Chemical Carcinogenesis in Rodents, held at NIEHS in 1993. [1] Made predictions for 10 of 15 chemicals. [2] Made predictions for 14 of 15 chemicals.

Method of prediction	**Information used**	**Accuracy**
RL rule set	8 attributes (2 short-term assays, 4 physical features, 2 dose measurements)	0.80 (12/15)
Human experts (Tennant *et al.*, 1990)	various information such as subchronic pathology, structural alerts, short-term assays, and structural features	0.80 (12/15)
MultiCASE (Rosenkranz & Klopman, 1990a)	substructures–biophores and biophobes	0.60 (9/15)
C4.5 decision tree (Bahler & Bristol, 1993)	189 attributes consisting of structural alerts, dose measurements, subchronic organ pathology, short-term assays	0.80 (12/15)
C4.5 rules (Bahler & Bristol, 1993)	189 attributes consisting of structural alerts, dose measurements, subchronic organ pathology, short-term assays	0.60 (9/15)
C4.5 decision tree learned from the same data that RL used	Same attributes as RL used	0.33 (5/15)
TOPKAT (Enslein *et al.*, 1990)	structural activity, physical properties, substructures	0.60 (6/10)[1]
DEREK (Sanderson & Earnshaw, 1991)	chemical substructures	0.50 (7/14)[2]
COMPACT (Lewis *et al.*, 1990)	shape and molecular energy levels of chemical structures	0.53 (8/15)

compared the predictions made by RL's rule set with those made by other computerized methods as well as those made by human experts. We also trained C4.5 on the same data using the same attributes RL used and made predictions with the decision tree for the same 24 chemicals.

Table 1 summarizes the accuracy of predictions submitted to the workshop for the 15 (8 carcinogens and 7 non-carcinogens) chemicals. For the 8 carcinogens and 7 non-carcinogens, the rule set learned by RL made correct predictions for 12 of 15 chemicals. The RL rule set was the only method that correctly predicted naphthalene as a carcinogen. Both the human experts and the decision tree obtained by Bahler and Bristol (1993) using C4.5 also made 12 correct predictions. However, the rule set learned by RL used only eight attributes (two short-term assays, four physical properties, and two dose measurements). The experts used a whole spectrum of information about chemicals including pathology and structural features. Bahler and Bristol used C4.5 with 189 attributes and the induced tree used more expensive features than those in RL's rule set. When a decision tree was learned (by us) using the same data and attributes that RL used, only five chemicals were classified correctly.

For the other 9 chemicals whose carcinogenicity was either unknown or equivocal, predictions made by RL's rule set agreed with the experts' decisions better than any other sub-

Table 2. Concordance of predictions made for 30 chemicals in the NIEHS PTE Project (Study 2) to two sets of predictions made by experts, Tennant and Spalding (1996) and Ashby (1996). Note that the concordance between the two experts' predictions is 0.63.

	Concordance to experts	
Prediction method	**Tennant & Spalding**	**Ashby**
Rule set learned by RL using subchronic organ toxicity (Lee *et al.*, 1996)	0.64	0.55
CASE and MultiCASE (Zhang *et al.*, 1996) using substructures and QSAR model	0.67	0.42
CSWG (Huff *et al.*, 1996)	0.64	0.41
Bootman (1996)	0.58	0.37
RASH (Jones & Easterly, 1996)	0.63	0.43
QSAR (Benigni *et al.*, 1996)	0.59	0.36
SHE (Kerckaert *et al.*, 1996)	0.53	0.35
QSAR (Purdy, 1996)	0.47	0.29
DEREK (Merchant, 1996)	0.42	0.40
FALS (Moriguchi *et al.*, 1996)	0.67	0.60
PROGOL (King & Srinivasan, 1996)	0.60	0.48
COMPACT (Lewis *et al.*, 1996)	0.53	0.48
Huff *et al.* (1996)	0.53	0.47
Tennant and Spalding (1996)	-	0.63
Ashby (1996)	0.63	-

mitted predictions. The detailed comparisons of predictions can be found in Lee, Buchanan, Mattison, Klopman, and Rosenkranz (1995).

The results of the second and third studies are currently being evaluated by making predictions for a new set of 30 chemicals from the NIEHS Predictive-Toxicology Evaluation (PTE) Project (Bristol, Wachsman & Greenwell, 1996). From the second study, two rule sets (one learned with the assumptions about liver and kidney and the other learned with no assumptions) were selected to make predictions (Lee, Buchanan & Rosenkranz, 1996). The rules learned from the third study were not directly applied to make predictions. Rather, rules learned by RL were used to assist the toxicologists in making predictions using the CASE and MultiCASE programs. In particular, the toxicologists made predictions using the discovered knowledge about combining probabilities and potency values predicted by the CASE and MultiCASE programs (Zhang, Sussman, Macina, Rosenkranz & Klopman, 1996). For example, breaking with the past, the experts did not use the potency values that the CASE program predicted, because in our experiments the performance of rule sets learned with the potency values from CASE was inferior to the performance of rule sets learned without the CASE potency values.

A total of 17 sets of predictions were submitted to the workshop by a number of scientists using different methods, including two rule sets learned by RL, one by the CASE and MultiCASE programs, and at least two sets by human experts. The list of chemicals and list of participants along with their predictions can be found in Bristol, Wachsman, and Greenwell (1996). The long-term rodent bioassays for these chemicals have not been completed and the validation of predictions has yet to be performed. At this point, we can only compare how well predictions by RL rule sets and the CASE and MultiCASE programs agree with those of experts, as shown in Table 2. Each set of predictions submitted to the NIEHS PTE Project is compared with two sets of predictions made by experts, one by Tennant and Spalding (1996) and the other by Ashby (1996). Note that the concordance between the two experts' predictions is only 0.63.

The predictions by the FALS program, which uses chemical structures, agreed best with both experts' predictions. The predictions made by CASE and MultiCASE agreed most with the predictions of Tennant and Spalding's, but did not agree well with the predictions by Ashby. The predictions by RL based on subchronic organ toxicity agreed nearly as well as FALS with both experts' predictions.

4. RL in chemical carcinogenesis

During our study, RL has been extended to accommodate our need to explore the databases in various ways. It should be noted that not all of these extensions were of benefit to the study described here, but we didn't know that until we tried them. Also, we tried to design these extensions to be as general as possible so that they can be applied easily to other problem domains.

The semantic constraints discussed previously in Section 3.5 mainly concern the use of attributes and their values in rules. We implemented these semantic assumptions as constraints on rule formation during learning. We extended RL's PDM with a set of constraints which allows a user to specify preference on attributes, their values, and their combinations.

4.1. Attribute definition

The attribute definition in the PDM was extended to specify semantic types ("meanings") of attributes, priorities and importance levels of attributes, and certainties of attributes' values. For example, the attribute SAL, representing responses in the *Salmonella* assays, is defined as:

```
(SAL (type symbolic)
     (values + - m ?)
     (meaning short-term-assay)
     (order 1)
     (certainty 0.8))
```

The `values` field specifies that the values of the attribute are `+` (positive), `-` (negative), `m` (marginal), and `?` (unknown). The `meaning` field was to provide a reference to a set of attributes with similar semantics. The `order` specifies a rank by which RL specializes attributes. If omitted, attributes are ordered according to their information gain (Quinlan,

1993). The `certainty` indicates an estimated probability that the attribute value measurements are correct, with a default of 1.0 to mean that the measurements are 100% certain. In this case, it is believed that the responses of short-term assays are at best 80% reproducible. In some experiments, attribute certainties were combined with a rule's performance to calculate the overall certainty of the rule.

4.2. Semantic constraints

Several semantic constraints were incorporated into the PDM to represent a user's wish to use attributes and their values in a particular way. First, the PDM was extended to require that every rule learned must use certain attributes. We generalized this requirement to create layers of attribute usage and to learn rules according to the layers:

```
(Layers (Num_Layers  n)
        (Layer   1   (A_11, ..., A_1i))
        (Layer   2   (A_21, ..., A_2j))
                 . . .
        (Layer   n   (A_n1, ..., A_nk)))
(Layered_learning   yes)
```

where A_{ij}'s are attributes. The above specifies that rules must be learned in layers. To learn the first layer, RL finds only rules which include at least one of the attributes A_{11}, ..., A_{1i}. If there are still examples left uncovered, RL learns rules for the next layer which must include one or more of the attributes A_{21}, ..., A_{2i}. The process continues through layers until all of the examples are covered. For example, in some experiments of our study, experts preferred rules which used at least one or more short-term assays attributes. This preference was met by the following statements in the PDM:

```
(Layers (Num_Layers  2)
        (Layer  1  (short-term-assays)))
(Layered_learning   yes)
```

If a rule set is learned in layers, then rules are also applied in layers such that rules in layer `x` are only applied to the cases for which rules in layer `x-1` fail to make any predictions.

Second, the PDM was extended to include constraints representing permissible or forbidden combinations of attributes and values. Each constraint has a system-defined predicate followed by a rule template. Three predicates were used: `USE`, `NOT_USE`, and `EXCLUDE`. The predicates `USE` and `NOT_USE` specifiy attributes or combinations which may or may not be used in rules. For example, in several experiments, the following was used to constrain search such that the positive responses in short-term assays be used in rules predicting carcinogenicity and the negative responses be used in rules for non-carcinogenicity:

```
(USE ((IF (short-term-assays +)) (THEN C))
     ((IF (short-term-assays -)) (THEN NC)))
```

where `C` and `NC` are class names for rodent carcinogenicity and non-carcinogenicity, and the form `((IF ...)  (THEN ...))` specifies a rule template. Note that the assumptions

represented by USE and NOT_USE were considered and implemented as exclusive assumptions. For example, the above constraints mean that the positive responses can be used only for C and the negative responses be used only for NC.

In addition, the predicate EXCLUDE was introduced to prevent the formation of rules with certain attributes or attribute-value pairs. For example, in some experiments in the third study, we wanted to test the predictive strength of rule sets learned without using predicted values from the CASE program. This was achieved by:

```
(EXCLUDE (NC C) (CASE-probabilities CASE-potency-values))
```

where CASE-probabilities and CASE-potency-values refer to all attributes whose values are probabilities and potency values, respectively, predicted by the CASE program. The EXCLUDE predicate was initially added to avoid the need to process the data whenever a subset of attributes were to be used, thus improving efficiency in experimentation.

4.3. Search

RL uses a beam search to find rules that meet both performance and semantic requirements. At each step of the search, all rules in the beam are specialized with one or more new features, and each newly created rule is examined. If a rule meets all constraints, it is added into a rule set. If it is too specific (i.e., there are too many features in the left–hand side already, or covers too few examples), or it does not satisfy constraints, the rule is removed from further consideration. If the rule is too general (i.e., too short or covers too many examples incorrectly) but meets all other constraints, then it is saved for further specialization. The saved rules are then sorted and only the top b rules are kept, where b is the beam width.

In our study, we used a two-level beam search with two beams of rules, each with user-defined size. The two-level beam search works just like a regular beam search, except that a rule which meets all constraints is not added to a class definition (rule set) right away. Rather, such rules are saved in the second beam, and when the second beam becomes full the best rule—based on a rule evaluation criterion—is selected and added to the class definition. The two-level beam search is an attempt to lessen the effects of local maxima.

The PDM was also extended to include one or more partial rules from which rule search begins, thus providing a way to test and further specialize prior knowledge or hypotheses. For example, the domain experts wanted to find out if the predictive strength of the presence of toxic effects in kidneys can be improved with toxic effects on other organs. This was achieved by providing the following two rules:

```
(SEED ((Kidney yes) ==> C)
      ((Kidney no) ==> NC))
```

where Kidney is the name of attribute, yes and no are values indicating the observation of toxic effects, and (Kidney yes) ==> C is the partial rule, "if any toxic effect is observed in kidneys, the chemical is a carcinogen." Partial rules are also useful in focusing on a subset of training data without physically modifying the training data.

4.4. Numeric attributes

For numeric attributes, RL selects interval endpoints dynamically as the search proceeds. The number of intervals is dependent on the types of numeric ranges being sought. In our study, the experts wished numeric features to specify open ranges only. That is, ($A < x$) or ($A > x$), but not ($x < A < y$), where A is a numeric attribute, and x and y are thresholds. The experts noted that it was difficult to justify such closed ranges of numeric values. Avoiding closed ranges in numeric features further required another constraint on rule formation. For example, if the values of two numeric attributes, A_1 and A_2, are linearly correlated (either positively or negatively), the conjunction of the two features ($N1 < v1$) and ($N2 > v2$) has an effect similar to specifying a closed range of values of A_1 or A_2. To avoid such cases, when a rule contains two or more numeric features and their attributes are specified to have open ranges only, correlation coefficients of each pair of numeric attributes are used to decide whether to combine two numeric features. If a rule has two numeric features and the values of two attributes are strongly positively correlated (i.e., coefficient > 0.5), both numeric features should contain the same inequality operator ($<$ or $>$), and if the values are strongly negatively correlated (i.e., coefficient $<$ -0.5), then the two numeric features should have different operators.

Using numeric features with open ranges may miss closed intervals representing an exception. To find such exceptions, we extended RL's threshold selection algorithm to search for an exception for each open range. We found that including exceptions did not improve performance. This may be because our study does not include sufficient training data for an exception to be statistically significant.

4.5. Evaluation functions

RL provides several different methods for evaluating individual rules and disjunctive rule sets. For individual rules, these include Quinlan's (1987) certainty factor, MYCIN's certainty factor (Buchanan & Shortliffe, 1984), PROSPECTOR's odds calculation functions (Duda, Hart & Nilsson, 1978), RECON's rule evidence measure (Simoudis, Livezey & Kerber, 1994; Kerber, 1991).

In some of our experiments, the certainty of a rule based on its example coverage was coupled with the certainty of attribute value measurements (included in the attribute definition in the PDM) to give the rule's certainty. While there are several ways to combine these, we experimented with a method in which the certainty of a rule was a product of certainties of those attributes used in the rule multiplied by the certainty of a rule (based on its example coverage):

$$CF(R) = E(R) \prod_{\substack{Attribute\ A \\ in\ rule\ R}} C(A)$$

where R is a rule, $C(A)$ is the certainty of attribute A, and $E(R)$ is a certainty measure based on the rule R's example coverage.

For a disjunctive set of rules, several different evidence-gathering or rule combination methods are provided, including simple and weighted voting of individual rule certainties, MYCIN's certainty factor combining method, PROSPECTOR's odds combination,

AQ15's rule strength combination (Michalski, Mozetic, Hong & Lavrac, 1986), and also no combination of rule certainties.

Many of the cross-validation experiments in our study used layered learning, as described previously. When a rule set is learned in layers depending on the attribute usage, the certainties of rules are also combined according to the layers. That is, only the certainties of rules in the same layer are combined.

5. Discussion and conclusions

The studies reported here explain a puzzle that arose in the first applications of inductive learning programs to scientific discovery over twenty-five years ago: *If scientific discovery involves finding patterns in empirical data, why are inductive learning programs not routinely making interesting discoveries?* In applications of Meta-DENDRAL (Buchanan & Feigenbaum, 1978) and subsequently in other investigators' applications of their inductive learning programs, it was necessary to tune the parameters of the program, change the feature language, and adjust the evaluation function before the program produced interesting rules. We initially took this as a design flaw, although we believed the fundamental paradigm and search strategy in Meta-DENDRAL to be sound.

The flaw, however, was believing an overly simplified model of science as hypothesis testing, which left little room for exploratory work. Automated bias space search (Provost & Buchanan, 1995) addresses part of the problem of understanding the nature of the exploratory work, but for the moment no one has fully automated the imaginative work of creative scientists who are exploring new problems. Thus, the answer to the question posed twenty-five years ago is two-fold:

1. Inductive learning programs, even with bias space search, do not adequately address the creative play involved in the early stages of scientific explorations.

2. Additional tools that support exploration of alternative questions to ask and methods to use have only recently been seen as important for data mining.

Inductive rule formation finds general rules that are understandable and accurate, but these are not sufficient to support discovery. Tools to support discovery in science need to be flexible enough to be used iteratively under different sets of assumptions, with different vocabularies, and with different questions in mind. Rules resulting from inductive learning must also satisfy semantic constraints, where the constraints differ as the focus of attention shifts. We have shown how one inductive rule learning system provides some of the flexibility needed to be a useful tool in an important health-related domain. These studies have helped us understand some of the general kinds of semantic constraints that transform a set of plausible and implausible associations into suggestions that scientists take seriously. We recognize that there is considerably more to be done, but we believe that using knowledge in learning is an essential component.

The main theme of our study was to include some domain semantics in the data mining process so that more plausible rules could be learned. We achieved this with constraints on the usage of attributes and their values in rules, and constraints reflecting users' preferences. We attempted to provide a rather general framework in which a user can easily test assumptions and his or her intuitions on attribute and value usages. It is important to note that the

performance of rule sets learned during cross-validation experiments and the success on predicting new chemicals were largely due to the semantic constraints. Without the constraints, rule sets learned during cross-validation experiments were at most 64% accurate, and made correct predictions for only 9 of 15 chemicals in the International Workshop of Predicting Chemical Carcinogenesis in Rodents.

Our study did not include some of the general issues of knowledge discovery in scientific domains, such as those involving a large amount of raw data, large dimensionality, and data quality (Fayyad, Haussler & Stolorz, 1996). Just as these issues are important, it is also important to provide a way to incorporate prior domain knowledge. With a general scheme to represent prior domain knowledge yet to be developed, our study was an attempt to incorporate certain semantic relationships of attributes and their values to target classes, which may not be found by relying only on coverage statistics.

Our experience in other domains—including medical diagnosis, clinical toxicology, infant mortality, etc.—suggests that the methods and constraints used in our study are not limited to the rodent carcinogenicity problem. It is conceivable that there are scientific domains where the flexibility and semantic constraints that RL offers through its PDM are of little value, because either scientists do not have a very clear idea of what to look for, or simply they do not know where to start. On the other hand, we believe there are many scientific problems for which the kind of flexibility RL offers is useful, especially those problems such as the one we studied where knowledge continues to evolve and the previously found knowledge often becomes contradictory as more data are obtained.

In summary, our investigation demonstrated the utility of knowledge-based rule induction in the problem of predicting rodent carcinogenicity and the place of rule induction in the overall process of discovery. Flexibility of the program in accepting different definitions of background knowledge and preferences was considered essential in this exploratory effort. The rule sets learned by RL demonstrated a marked improvement of predictive performance over current prediction methods. Although true biological evaluation of individual rules is yet to be performed, and such evaluation will take time, the exploratory framework in which we used rule induction revealed that some features of chemicals were more useful for predicting carcinogenicity than some scientists believed (Ramel, 1988; Tennant *et al.*, 1987).

Acknowledgments

We are greatly indebted to Professor Herbert Rosenkranz of the Department of Environmental and Occupational Health, and Professor Donald Mattison of the School of Public Health at the University of Pittsburgh for their collaboration in our investigation. We also thank Ying Ping Zhang for her assistance in preparing databases in all three studies, Dr. Nancy Sussman for her assistance in statistical analysis of data in the third study, Dr. Gilles Klopman and Dr. Mario Dimayuga for help in using the CASE and MultiCASE programs, and the past and present members of the Intelligent Systems Laboratory for helpful discussions. This investigation was supported in part by grants from the International Life Science Institute-Risk Science Institute, the W.M. Keck Foundation (grant 921277), the Department of Defense (grant DAAA21-93-C-0046), the Center of Alternatives to Animal Testing, the National Institutes of Health (grant 2-P41-RR06009-06, administered through

the Pittsburgh Supercomputing Center), and the National Science Foundation (grant IRI-9412549).

References

Aha, D.W., & Kibler, D. (1989). Noise-tolerant instance-based learning algorithms. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 794–799).

Ambrosino, R., Lee, Y., Rosenkranz, H.S., Mattison, D.R., Buchanan, B.G., Provost, F.J., & Gomez, J. (1993). The use of a knowledge-based induction program to predict chemical carcinogenesis in rodents. A report distributed at the International Workshop of Predicting Chemical Carcinogenesis in Rodents, held at National Institute of Environmental Health Sciences in Research Triangle Park, NC, USA.

Ames, B.N., Durston, W.E., Yamasaki, E., & Lee, F.D. (1973). Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection. *Proceedings National Academy of Science*, 70, 2281–2285.

Ashby, J. & Tennant, R.W. (1994). Prediction of rodent carcinogenicity for 44 chemicals: results. *Mutagenesis*, 9, 7-15.

Ashby, J. (1996). Prediction of rodent carcinogenicity for 30 chemicals. *Environmental Health Perspectives*, 104(Supplement 5), 1101-1104.

Bahler, D. & Bristol, D.W. (1993). The induction of rules for predicting chemical carcinogenesis in rodents. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology* (pp. 29–37).

Benigni, R., Andereoli, C., & Zito, R. (1996). Prediction of rodent carcinogenicity of 30 chemicals bioassayed by the US National Toxicology Program. *Environmental Health Perspectives*, 104(Supplement 5), 1041–1044.

Bootman, J. (1996). Speculations on the rodent carcinogenicity of 30 chemicals currently under evaluation in rat and mouse bioassays organized by the US National Toxicology Program. *Environmental and Molecular Mutagenesis*, 27, 237–243.

Bristol, D.W., Wachsman, J.T., & Greenwell, A. (1996). The NIEHS predictive-toxicology evaluation project. *Environmental Health Perspectives*, 104(Supplement 5), 1001–1010.

Buchanan, B.G. (1994). The role of experimentation in artificial intelligence. *Philosophical Transactions*, 349(1689), 143–166.

Buchanan, B.G., & Feigenbaum, E.A. (1978). DENDRAL and Meta-DENDRAL: their applications dimension. *Artificial Intelligence*, 11, 5–24.

Buchanan, B.G., & Lee, Y. (1995). Exploring alternative biases prior to learning in scientific domains. *Working Notes of AAAI 1995 Spring Symposium Series*.

Busey, W.M., & Henry, J.F. (1995). Intelligent toxicology predictions systems. *Toxicologist*, 15, 178–179.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, 3, 261–283.

Clearwater, S.H., & Lee, Y. (1993). Use of a learning program for trigger sensitivity studies. *Proceedings of the Third International Workshop on Software Engineering, Artificial Intelligence and Expert Systems for High Energy and Nuclear Physics* (pp. 207–212).

Clearwater, S.H., & Provost, F.J. (1990). RL4: a tool for knowledge-based induction. *Proceedings of Tools for Artificial Intelligence 1990* (pp. 24–30).

Cooper, G., Aliferis, C.F., Ambrosino, R., Aronis, J.M., Buchanan, B.G., Caruana, R., Fine, M.J., Glymour, C., Gordon, G., Hanusa, B., Janowsky, J.E., Meek, C., Mitchell, T.M., Richardson, T., & Spirtes, P. (1997) An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9, 107–138.

Danyluk, A.P., & Provost, F.J. (1993). Small disjuncts in action: Learning to diagnose errors in the local loop of the telephone network. *Proceedings of the Tenth International Conference on Machine Learning.*

Domingos, P. (1994). The RISE system: conquering without separating. *Proceedings of Tools with Artificial Intelligence 1994*.

Duda, R.O, Hart, P.E., Nilsson, N.J., & Sutherland, G.L. (1978). Semantic network representation in rule based inference system. In Waterman, D.A. and Hayes-Roth, F. (eds), *Pattern Directed Inference Systems*. Academic Press.

Enslein, K., Blake, B.W., & Borgstedt, H.H. (1990). Predictions of probability of carcinogenicity for a set of ongoing NTP bioassays. *Molecular Mutagenesis*, 13, 332–338.

Fawcett, T. & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3), 291–316.

Fayyad, U., Haussler, D., & Stolorz, P. (1996a). KDD for science data analysis: issues and examples. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 50–56).

Fayyad, U., Haussler, D., & Stolorz, P. (1996b). Mining scientific data. *Communications of the ACM*, 39(11), 51–57.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37–54.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). Knowledge discovery and data mining: Towards a unifying framework. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 82–88).

Gomez, J., Lee, Y., & Mattison, D.R. (1993). Identification of developmental toxicants using a rule learning expert system. *Programs and Abstracts: The Fourteenth Annual Meetings of the American College of Toxicology.*

Gomez, J., Lee, Y., & Mattison, D.R. (1994) RL: An innovative tool for predicting developmental toxicity. *Toxicologist*, 14, 295.

Hennessy, D., Gopalakrishnan, V., Buchanan, B.G., Rosenberg, J.M., & Subramanian, D. (1994). Induction of rules for biological macromolecule crystallization. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (pp. 179–187).

Huff, J.E., Weisburger, E., & Fung, V. (1996). Multicomponent criteria for predicting carcinogenicity: 30 NTP chemicals. *Environmental Health Perspectives*, 104(Supplement 5), 1105–1112.

Jones, T.D., & Easterly, C.E. (1996). A RASH analysis of National Toxicity Program data: Predictions for 30 compounds to be tested in rodent carcinogenesis experiments. *Environmental Health Perspectives*, 104(Supplement 5), 1017–1030.

Kerber, R. (1991). Learning classification rules from examples. *Proceedings of 1991 AAAI Workshop on Knowledge Discovery in Databases*.

Kerckaert, G.A., Brauninger, R., LeBoeuf, R.A., & Isfort, R.J. (1996). Use of the Syrin hamster embryo cell transformation assay for carcinogenicity prediction of chemicals currently being tested by the NTP in rodent bioassays. *Environmental Health Perspectives*, 104 (Supplement 5), 1075–1084.

King, R.D., & Srinivasan, A. (1996). Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. *Environmental Health Perspectives*, 104 (Supplement 5), 1031–1040.

Klopman, G. (1984). Artificial intelligence approach to structure-activity studies: Computer automated structure evaluation of biological activity of organic molecules. *Journal of American Chemical Society*, 106, 7315–7320.

Klopman, G. (1992). MultiCASE 1. A hierarchical computer automated structure evaluation program. *Quantitative Structure-Activity Relationships*, 11, 176–184.

Lee, Y. (1995). Learning a robust rule set. Ph.D. Thesis. Computer Science Department, University of Pittsburgh.

Lee, Y., Buchanan, B.G., Klopman, G., Dimayuga, M., & Rosenkranz, H.S. (1996). The potential of organ specific toxicity for predicting rodent carcinogenicity. *Mutation Research*, 358, 37–62.

Lee, Y., Buchanan, B.G., Mattison, D.R., Klopman, G., & Rosenkranz, H.S. (1995). Learning rules to predict rodent carcinogenicity of non-genotoxic chemicals. *Mutation Research*, 328, 127–149.

Lee, Y., Buchanan, B.G., & Rosenkranz, H.S. (1996). Carcinogenicity predictions for a group of 30 chemicals undergoing rodent cancer bioassays based on rules derived from subchronic organ toxicities. *Environmental Health Perspectives*, 104 (Supplement 5), 1059–1063.

Lee, Y., & Clearwater, S.H. (1992). Tools for automating experiment design: A machine learning approach. *Proceedings of the Fourth International Conference on Tools with Artificial Intelligence*.

Lee, Y., Zhang, Y.P., Sussman, N., & Rosenkranz, H.S. (1995). Evaluation of predicted probabilities and potencies generated by the SAR expert systems, MultiCASE and CASE: Application to rodent carcinogenicity databases. Technical Report, Computer Science Department, University of Pittsburgh.

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

Lewis, D.F., Ioannides, C., & Parke, D.V. (1996). Compact and molecular structure in toxicity assessment: a prospective evaluation of 30 chemicals currently being tested for rodent carcinogenicity by the NCI/NTP. *Environmental Health Perspectives*, 104(Supplement 5), 1011–1016.

Merchant, C. (1996). Prediction of rodent carcinogenicity using the DEREK system for thirty chemicals being tested by the National Toxicology Program. *Environmental Health Perspectives*, 104(Supplement 5), 1065–1073.

Michalski, R.S., Mozetic, I., Hong, J., & Lavrac, N. (1986). The multipurpose incremental learning system AQ15 and its testing application to three medical domains. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 1041–1045).

Moriguchi, I., Hirano, H., & Hirono, S. (1996). Prediction of the rodent carcinogenicity of organic compounds from their chemical structures using the FALS method. *Environmental Health Perspectives*, 104(Supplement 5), 1051–1058.

Parry, J.M. (1994). Detecting and predicting the activity of rodent carcinogens. *Mutagenesis*, 9, 3–5.

Pet-Edwards, J., Haimes, Y.Y., Chankong, V., Rosenkranz, H.S., & Ennever, F.K. (1989). *Risk assessment and decision making using test results: the carcinogenicity prediction and battery selection approach.* Plenum Press.

Provost, F.J. (1992). Policies for the selection of bias in inductive machine learning. Doctoral dissertation, Computer Science Department, University of Pittsburgh.

Provost, F.J., & Aronis, J.M. (1996). Scaling up inductive learning with massive parallelism. *Machine Learning*, 23, 33–46.

Provost, F.J., & Buchanan, B.G. (1995). The pragmatics of bias selection. *Machine Learning*, 20, 35–61.

Purdy, R. (1996). A mechanism-mediated model for carcinogenicity, model content and prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 25 organic chemicals. *Environmental Health Perspectives*, 104(Supplement 5), 1085–1094.

Quinlan, J.R. (1987). Generating production rules from decision trees. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (pp. 304–307).

Quinlan, J.R. (1993). *C4.5: Programs for machine learning.* Morgan Kaufmann.

Ramel, C. (1988). Short-term testing—are we looking at wrong endpoints? *Mutation Research*, 205, 13–24.

Rosenkranz, H.S., Frierson, M.R., & Klopman, G. (1986). Use of structure-activity relationships in predicting carcinogenesis. *Long-term and Short-term Assays for Carcinogen: A Critical Appraisal*, 83, 497–517.

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

Rosenkranz, H.S., & Klopman, G. (1990b). Identification of rodent carcinogens by expert system. *Environmental Mutagenesis*, Part B, 23–48.

Rosenkranz, H.S., & Klopman, G. (1995). The application of structural concepts to the prediction of the carcinogenicity of therapeutic agents. In Wolf, M.E. (Ed), *Medicinal Chemistry and Drug Discovery, Volume 1: Principles and Practice*. John Wiley and Sons.

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

Schaffer, C. (1994). A conservation law for generalization performance. *Proceedings of the Eleventh International Conference on Machine Learning*. Morgan Kaufmann.

Simoudis, E., Livezey, B., & Kerver, R. (1994). Integrating inductive and deductive reasoning for database mining. *Proceedings of AAAI-94 Workshop of Knowledge Discovery in Databases*.

Tennant, R.W., Margolin, B.H., Shelby, M.D., Zeiger, E., Haseman, J.K., Spalding, J., Caspary, W., Resnick, M., Stasiewicz, S., Anderson, B., & Minor, R. (1987). Prediction of chemical carcinogenicity in rodents from in vitro genetic toxicity assays. *Science*, 236, 933–941.

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

Tennant, R.W., & Spalding, J. (1996). Predictions for the outcome of rodent carcinogenicity bioassays: identification of trans-species carcinogens and noncarcinogens. *Environmental Health Perspectives*, 104(Supplement 5), 1095–1100.

Zhang, Y.P. (1995). Validation of prediction models of carcinogenicity derived from QSAR analysis by CASE and MultiCASE expert systems. Lab Report, Department of Environmental and Occupational Health, University of Pittsburgh.

Zhang, Y.P., Sussman, N., Macina, O.T., Rosenkranz, H.S., & Klopman, G. (1996). Predictions of the carcinogenicity of a second group of organic chemicals undergoing carcinogenicity testing. *Environmental Health Perspectives*, 104(Supplement 5), 1041–1044.

Received March 4, 1997
Accepted September 18, 1997
Final Manuscript November 15, 1997