Carcinogenesis predictions using ILP

A. Srinivasan [1], R.D. King [2], S.H. Muggleton [1]
M.J.E. Sternberg [3]

[1] Oxford University Computing Laboratory, Wolfson Building,
Parks Road, Oxford, UK
[2] Department of Computer Science, The University of Wales Aberystwyth
Penglais, Ceredigion, UK
[3]Biomolecular Modelling Laboratory, Imperial Cancer Research Fund
44 Lincoln's Inn Fields, London, U.K.

Abstract

Obtaining accurate structural alerts for the causes of chemical cancers is a problem of great scientific and humanitarian value. This paper follows up on earlier research that demonstrated the use of Inductive Logic Programming (ILP) for predictions for the related problem of mutagenic activity amongst nitroaromatic molecules. Here we are concerned with predicting carcinogenic activity in rodent bioassays using data from the U.S. National Toxicology Program conducted by the National Institute of Environmental Health Sciences. The 330 chemicals used here are significantly more diverse than the previous study, and form the basis for obtaining Structure-Activity Relationships (SARs) relating molecular structure to cancerous activity in rodents. We describe the use of the ILP system Progol to obtain SARs from this data. The rules obtained from Progol are comparable in accuracy to those from expert chemists, and more accurate than most state-of-the-art toxicity prediction methods. The rules can also be interpreded to give clues about the biological and chemical mechanisms of cancergenesis, and make use of those learnt by Progol for mutagenesis. Finally, we present details of, and predictions for, an ongoing international blind trial aimed specifically at comparing prediction methods. This trial provides ILP algorithms an opportunity to participate at the leading-edge of scientific discovery.

1 Introduction

The task of obtaining the molecular mechanisms for biological toxicity has been a prominent area of application for Inductive Logic Programming (ILP) systems. Recently, this has seen the use of an ILP program to the task of predicting the mutagenic activity of a restricted class of molecules [12, 21]. The results reported, while interesting, were preliminary for the following reasons. Firstly, the data pertain to a relatively homogeneous class of compounds –although, in themselves, they were more diverse than those analysed previously by ILP (see [13]). Secondly, while some comparative studies were performed ([22]), they

were not against state-of-the-art methods designed specifically for toxicity prediction. Finally, a single success is clearly not sufficient grounds for claiming general applicability of a technique. In this paper we remedy each of these shortcomings. In the course of doing so, we present an important new problem where any scientific discoveries made by ILP programs will be measured against international competition in true blind trials.

This paper is organised as follows. Section 2 describes the problem of carcinogenesis prediction of rodent bioassays. These assays are conducted as part of the National Toxicology Program (NTP) by the U.S. National Institute for Environmental Health Sciences (NIEHS). A prominent feature associated with the NTP is the NIEHS Predictive Toxicology Evaluation – or PTE – project ([6]). The PTE project identifies a "test" set of chemicals from those currently undergoing tests for carcinogenicity within the NTP. Predictions on this test set were invited and then compared against the true activity observed in rodents, once such data are available. The description of these blind trials, including details of a trial scheduled for completion in 1998, is described in Section 3. Section 4 describes the use of the ILP program Progol ([17]) to extract molecular descriptions for cancerous activity. These are used to compare against state-of-the-art predictions on an earlier trial in the PTE project (Section 4.4). Predictions of the activity of chemicals in the ongoing PTE trial are also in Section 4.4. Section 5 concludes this paper.

2 The Carcinogenesis problem and the NTP data base

Prevention of environmentally-induced cancers is a health issue of unquestionable importance. Almost every sphere of human activity in an industrialised society faces potential chemical hazards of some form. In [9], it is estimated that nearly 100,000 chemicals are in use in large amounts every day. A further 500–1000 are added every year. Only a small fraction of these chemicals have been evaluated for toxic effects like carcinogenicity. The U.S. National Toxicology Program (NTP) contributes to this enterprise by conducting standardised chemical bioassays – exposure of rodents to a range of chemicals – to help identify substances that may have carcinogenic effects on humans. However, obtaining empirical evidence from such bioassays is expensive and usually too slow to cope with the number of chemicals that can result in adverse effects on human exposure. This has resulted in an urgent need for models that propose molecular mechanisms for carcinogenesis. It is envisaged that such models would (a) generate reliable toxicity predictions for all kinds of chemicals; (b) enable low cost identification of hazardous chemicals; and (c) refine and reduce the reliance on the use of large number of laboratory animals [6]. Pattern-recognition methods can "...help identify, characterise, and understand the various mechanisms or modes of action that determine the type and level of response observed when biological systems are exposed to chemicals" [6].

Tests conducted by the NTP have so far resulted in a data base of more than 300 compounds that have been shown to be carcinigenic or otherwise in

rodents. Amongst other criteria, the chemicals have been selected on the basis of their carcinogenic potential – for example, positive mutagenicity tests – and on evidence of substantial human exposure ([9]). Using rat and mouse strains (of both genders) as predictive surrogates for humans, levels of evidence of carcinogenicity are obtained from the incidence of tumors on long-term (two years) exposure to the chemicals. The NTP assigns the following levels of evidence: CE, clear evidence; SE, some evidence; E, equivocal evidence; and NE, no evidence. Precise definitions for determining these levels can be found in [9], and a complete listing of all chemicals tested is available at the NTP Home Page: *http://ntpserver.niehs.nih.gov/*.

The diversity of these compounds present a general problem to many conventional SAR techniques. Most of these, such as the regression-based techniques under the broad category called Hansch Analysis ([15]), can only be applied to model compounds that have similar mechanisms of action. This "congeneric" assumption does not hold for the chemicals in the NTP data base, thus limiting the applicability of such methods. The Predictive Toxicology Evaluation project undertaken by the NIEHS aims to obtain an unbiased comparison of prediction methods by specifying compounds for blind trials. One such trial, PTE-1, is now complete. Complete results of NTP tests for compounds in the second trial, PTE-2, will be available by mid 1998.

3 The blind trials PTE-1 and PTE-2

The PTE project ([6]) is concerned with predictions of overall cancerous activity of a pre-specified set of compounds. This overall activity is either "POS" if the level of activity is CE or SE, or "NEG". The PTE project identifies a set of compounds either scheduled for, or currently undergoing, NTP tests. Information concerning the bioassays is disseminated with the view of encouraging the use of state-of-the-art toxicity prediction schemes. Once the true results of biological activity are available, the project collects a set of leading predictors and publishes their results. The first of these trials, termed PTE-1 is now complete, and results for 39 chemicals are available in [3]. [1]

A second round of toxicology evaluation – PTE-2 – consisting of 30 compounds (of which 5 are inorganic) is currently in progress. True biological activity for 13 of these have been determined at the time of writing of this paper. A complete description of chemicals in PTE-2, along with a schedule of dates is available in [6]. The remaining activity levels should be determined by 1998. In this paper, we intend to use Progol to obtain structural alerts from chemicals in the NTP data base. In the first instance, predictions from these alerts will be compared against other predictions available for PTE-1. This will be followed by predictions for compounds in PTE-2.

[1] A preliminary effort by the ILP system Progol in presented in [14]. The results in this paper subsume these early results as a number of toxicology indicators were unavailable to us at that time. Further details are in Section 4.

4 Carcinogenesis predictions using Progol

4.1 Aims

The experiment described here has the following aims.

1. Use the ILP system Progol to obtain rules for carcinogenicity from data that does not include compounds in PTE-1 and PTE-2.
2. Predict carcinogenic activity of compounds in PTE-1, and compare against other state-of-the-art toxicity prediction methods.
3. Predict carcinogenic activity of compounds in PTE-2.

4.2 Materials

Data

Figure 1 shows the distribution of compounds in the NTP data base having an overall activity of $POS(+)$ or $NEG(-)$. Appendix A gives a complete listing of the compounds, along with identifiers into the NTP data base and the actual class labels.

Compounds	+	−	Total
PTE-1	20	19	39
PTE-2	$\geq$7	$\geq$6	30
Rest	162	136	298

Figure 1: Class distribution of compounds. Complete details of PTE-2 will be available by 1998.

Background knowledge

The following background knowledge is available for each category of compounds listed in Figure 1. Complete Prolog descriptions of each of the following are available via anonymous ftp to *ftp.comlab.ox.ac.uk*, in the directory *pub/Packages/ILP/Datasets*.

Atom-bond description. These are ground facts representing the atom and bond structures of the compounds. The representation first introduced in [21] is retained. These are Prolog translations of the output of the molecular modelling package QUANTA. Bond information consist of facts of the form *bond(compound,atom1,atom2,bondtype)* stating that *compound* has a bond of *bondtype* between the atoms *atom1* and *atom2*. Atomic structure consists of facts of the form *atm(compound,atom,element,atomtype,charge)*, stating that in *compound*, *atom* has element *element* of *atomtype* and partial charge *charge*.

Generic structural groups. This represents generic structural groups (methyl groups, benzene rings etc.) that can be defined directly using the atom and bond description of the compounds. Here we use definitions for 29 different structural groups, which expands on the 12 definitions used in [22]. We pre-compute these structural groups for efficiency. An example fact that results is in the form *methyl(compound,atom_list)*, which states that the list of atoms *atom_list* in *compound* form a methyl group. Connectivity amongst groups is defined using these lists of atoms.

Genotoxicity. These are results of short-term assays used to detect and characterize chemicals that may pose genetic risks. These assays include the *Salmonella* assay, in-vivo tests for the induction of micro-nuclei in rat and mouse bone marrow etc. A full report available at the NTP Home Page lists the results from such tests in one of 12 types. Results are usually + or − indicating positive or negative response. These results are encoded into Prolog facts of the form *has_property(compound,type,result)*, which states that the *compound* in genetic toxicology *type* returned *result*. Here *result* is one of p (positive or n (negative). In cases where more than 1 set of results are available for a given type, we have adopted the position of returning the majority result. When positive and negative results are returned in equal numbers, then no result is recorded for that test.

Mutagenicity. Progol rules from the earlier experiments on obtaining structural rules for mutagenesis are included ([12, 20]). Mutagenic chemicals have often been found to be carcinogenic ([7]), and we use all the rules found with Progol (see [20] for a complete listing).

Structural indicators. We have been able to encode some of the structural alerts used in [1]. At the time of writing this paper, the NTP proposes to make available nearly 80 additional structural attributes for the chemicals. Unfortunately, this is not yet in place for reuse in experiments here.

Prediction methods

The ILP system used here is P-Progol (Version 2.3). This a Prolog implementation of the Progol algorithm ([17]), and we will refer to this simply as Progol in the rest of this paper. P-Progol is available via anonymous ftp to *ftp.comlab.ox.ac.uk*, in the directory *pub/Packages/ILP*. The other toxicity prediction methods compared against Progol's PTE-1 predictions are: Ashby [23], RASH [10], TIPT [2], Benigni [5], DEREK [19], Bakale [4], TOPKAT [8], CASE [18], and COMPACT [16]. We take the PTE-1 predictions of each these algorithms as reported in [3].

4.3 Method

The task is to obtain a theory for carcinogenesis using the 298 chemicals under the "Rest" category in Figure 1. This theory is then to be used to predict the classes of compounds in PTE-1 and PTE-2. Progol constructs theories within

the language and statistical constraints provided by the user. In domains such as the one considered here, it is difficult to know beforehand any reasonable set of constraints to provide. Further, it is not evident that the theory returned by default settings within the program is the best possible. Consequently, we adopt the following three-stage procedure.

Stage 1: Parameter identification. Identify 1 or more critical parameters for Progol. Changing these should result in significant changes in the theory returned by Progol.

Stage 2: Model selection. This proceeds as follows.

1. Randomly select a small subset of the 298 chemicals to act as a "validation" set. The remaining chemicals form the "training" set for Progol.
2. Systematically vary the the critical parameters. For each setting obtain a theory from the training set, and record its accuracy on the validation set.
3. Return the theory with the highest accuracy on the validation set.

Stage 3: Model evaluation. The predictions for PTE-1 and PTE-2 by the theory returned from Stage 2 are recorded. For other toxicity prediction methods, the probability that Progol classifies PTE-1 compounds in the same proportion as that method is obtained using McNemar's Test (see below).

For a given set of background predicate definitions, theories returned by Progol are usually affected by the following parameters: (1) c, bounding the number of literals in any hypothesised clause; (2) *noise*, bounding the minimum acceptable training set accuracy for a clause; and (3) *nodes*, bounding the number of clauses searched. Initial experimentation (Stage 1) suggested that the most sensitive parameter for Progol was *noise*. The experiments here consider theories arising from 4 settings corresponding *noise* values 0.35, 0.30, 0.25, and 0.20. For the data here, the size of the validation set is taken to be 30% of the 298 chemicals – that is, 89 compounds. Of these 49 are labelled + and the remaining 40 are labelled −. This leaves 209 compounds for training. Of these 113 are + and the remaining 96 are −.

We also note one other detail concerning the procedure for obtaining a final theory. The Prolog implementation used here can obtain clauses using two different search strategies. The first is as in [17], and results in redundant examples being removed after an acceptable clause is found. A second strategy retains these examples, which gives correct estimates for the accuracy of the clause found. Clauses obtained in this fashion can have significant overlap in the examples they make redundant. The preferred final theory is then the subset of these clauses that has maximal compression (within acceptable resource limits). [2]

[2]This subset is currently obtained by a companion program to P-Progol called T-Reduce (Version 1.0). Compression of a set of clauses is defined analogous to the measure in [17],

		Predicted (A_1) Correct	Predicted (A_1) Incorrect	
Predicted (A_2)	Correct	n_1	n_2	n_a
	Incorrect	n_3	n_4	n_b
		n_c	n_d	N

Figure 2: Cross-comparison of the predictions of a pair of algorithms $A_{1,2}$ n_1 is the number of compounds whose class is correctly predicted by both algorithms. Similarly for the entries $n_{2,3,4}$.

McNemar's Test

McNemar's test for changes is used to compare algorithms For a pair of algorithms, this is done by a cross-comparison of the compounds correctly and incorrectly classified as shown in Figure 2.

The null hypothesis is that the proportion of examples correctly classified by both algorithms is the same. If there is no significant difference in the performance of the two algorithms, half of the n_2+n_3 cases whose classifications disagree should be classified correctly by A_1 and A_2 respectively. Because of small numbers, we directly estimate the probability of a chance classification using the binomial distribution, with probability of success at 0.5. In effect, this is likened to probability of obtaining at least n_2 (or n_3, if greater) heads in a sequence of $n_2 + n_3$ tosses of a fair coin.

It is evident that repeated cross-comparisons will yield occasions when Progol's performance will apparently seem better than its adversary. For repeated comparisons of a given pair of algorithms on different random samples of data, it is possible to apply a correction (known as the Bonferroni adjustment) for this problem. The situation of repeated comparisons of different pairs of algorithms on a given set of data (as is here) does not, on the surface, appear to be amenable to the same correction. However, adopting the spirit of the correction, we advocate caution in quantitative interpretations of the binomial probabilities obtained.

4.4 Results and discussion

Figure 3 tabulates the accuracies on the validation set for each of the parameter settings explored. These results lead to the choice of 0.30 as the preferred setting for minimum noise for acceptable clauses.

Figure 4 shows an English translation of the theory with highest validation accuracy in Figure 3. Each disjunct in Figure 4 represents a rule followed by

namely, $P - N - L$ where P is the positive examples covered by the theory, N is the negative examples covered by the theory, and L is the number of clauses in the theory. T-Reduce is available on request from the first author.

Noise	Validation accuracy
0.35	0.63
0.30	0.70
0.25	0.63
0.20	0.65

Figure 3: Validation set accuracies at the model selection stage. "Noise" values provide a lower bound on the training set accuracy for a clause hypothesised by Progol. "Validation accuracy" is the corresponding accuracy on the validation set of the theory obtained from Progol at that noise level.

Progol. Rules 1-3 are based on biological tests. Additional comments on the rules follow.

Rule 1. The result of the Ames biological test for mutagenicity. The effectivness of the Ames test is widely recognised, but it is gratifying that Progol identifies it as the most important.

Rule 2. This rule is a test based on using whole (not cell culture) Drosopha. Like the Ames test it tests for mutagenicity.

Rule 3. This rule is puzzling as it would be expected that a positive test for chromosome aberration would be a test for carcinogenesis, not a negative test. More specialised variants of this rule were obtained in other theories obtained in Stage 1 of the experimental methodology, suggesting absence of chromosal aberrations does have some role to play, reasons for which requires investigation.

Rule 4. Aromatic compounds are often carcinogens and the low partial charge indicates relative reactivity. The use of a precise number for partial charge is an artifact of using the information from QUANTA, resulting from a particular molecular substructure around the aromatic carbon.

Rule 5. Amine groups are recognised by Ashby ([23]) as indicators of cancergenesis. This rule is a more accurate specification of this rule.

Rule 6. Aromatic hydrogen with a very high partial charge (often chlorinated aromatics). Such aromatics are relatively unreactive (perhaps giving time to diffuse to DNA).

Rule 7. The high partial charge on the hydroxyl oxygen suggests that the group is relatively unreactive. The significance of the aromatic (or resonant) hydrogen is unclear.

Rule 8. Compounds with bromine have been widely recognised as carcinogens ([23]).

Compound A is carcinogenic if:
(1) it tests positive in the Salmonella assay; or
(2) it tests positive for sex-linked recessive lethal mutation in Drosphila;
(3) it tests negative for chromosome aberration (an in-vivo cytogenetic assay); or
(4) it has a carbon in a six-membered aromatic ring with a partial charge of -0.13; or
(5) it has a primary amine group and no secondary or tertiary amines; or
(6) it has an aromatic (or resonant) hydrogen with partial charge $\geq$ 0.168; or
(7) it has an hydroxy oxygen with a partial charge $\geq$ -0.616 and an aromatic (or resonant) hydrogen; or
(8) it has a bromine; or
(9) it has a tetrahedral carbon with a partial charge $\leq$ -0.144 and tests positive on Progol's mutagenicity rules.

Figure 4: Progol's theory for carcinogenesis.

Rule 9. A tetrahedral carbon with low partial charge. The Progol rules for mutateginicity are shown to have utilty outside of their original application domain. This is interesting as it displays perhaps the first reuse of ILP-constructed knowledge between different scientific problems.

Predicting PTE-1

Figure 5 tabulates the accuracies of the different toxicity prediction methods on the compounds in PTE-1. This shows Progol to be comparable to the top 3 state-of-the-art toxicity predictors.

This result should be seen in the following perspective. The only method apparently more accurate than Progol is that of Ashby, which involves the participation of human experts and a large degree of subjective evaluation. All the methods with accuracy close to Progol (Ashby, RASH, and TIPT) have access to biological data that was not available to Progol (information form short-term - 13 week - rodent tests). It should also be noted that all the methods compared with Progol were specifically devloped for chemical structure activity and toxicity prediction. Some recent information available to us suggest that results are also results are comparable to those obtained by a mixture of ILP and regression with additional biological information. [3]

Predicting PTE-2

Figure 6 tabulates the predictions made by the theory in Figure 4 for compounds in PTE-2. The results to date show that Progol has currently predicted $8/13 \approx 62\%$ of the compounds correctly. Progol is currently ranked equal first for accuracy. The accuracy of Progol is again comparable to Ashby (7/13) and RASH (8/13) (no predictions are available as for TIPT). The lower accuracy of Progol (and the other participating methods) in PTE-2 compared with PTE-1 probably reflects the different distribution of compounds in PTE-2 compared

[3] Personal communication from S. Kramer to the second author.

Method	Type	Accuracy	P
Ashby†	Chemist	0.77	0.29
Progol	ILP	0.72	1.00
RASH†	Biological potency analysis	0.72	0.39
TIPT†	Propositional ML	0.67	0.11
Bakale	Chemical reactivity analysis	0.63	0.09
Benigni	Expert-guided regression	0.62	0.02
DEREK	Expert system	0.57	0.02
TOPKAT	Statistical discrimination	0.54	0.03
CASE	Statistical correlation analysis	0.54	< 0.01
COMPACT	Molecular modelling	0.54	0.01
Default	Majority class	0.51	0.01

Figure 5: Comparative accuracies on PTE-1. Here P represents the binomial probability that Progol and the corresponding toxicity prediction method classify the same proportion of examples correctly. The "Default" method predicts all compounds to be carcinogenic. Methods marked with a † have access to short-term in-vivo rodent tests that were unavailable to other methods. Ashby and RASH also involve some subjective evaluation to decide on structural alerts.

to PTE-1 and training data. For example: the percentage of compounds with positive a Ames test in PTE-2 is only 16% compared to an average 42% for PTE-1 and the training data. The changing distribution has been previously noted in [23] and probably reflects a different testing strategy by the NIEHS.

5 Conclusions

The carcinogenesis prediction trials conducted by the NIEHS offer ILP systems a unique opportunity to participate in true scientific discovery. The prediction of chemical cancerogensis is both an important medical problem and a fascinating research area. This paper provides initial performance benchmarks that we hope will act as an incentive for participation by other ILP systems in the field. Progol has achieved accuracy as good or bettter that current state-of-the-art methods of toxicity prediction. Results from other studies ([20]) suggest that addition of further relevant background knowledge should improve the Progol's prediction accuracy even further. In addition, Progol has produced nine rules that can be biologically and chemically interpreted and may help to provide a better understanding of the mechanisms of cancerogenesis.

The results for the predicion of carcinogenesis, taken together with the previous applications of predicting mutagenicity in nitro-atomatic compounds, and the inhibition of angiogenesis by suramin analogues [11], show that ILP can play an important role in understanding cancer related compounds.

Compound Id.	Name	Actual	Progol
6533-68-2	Scopolamine hydrobroamide	-	-
76-57-3	Codeine	-	-
147-47-7	1,2-Dihydro-2,2,4-trimethyquinoline	+	-
75-52-8	Nitromethane	-	-
109-99-9	Tetrahydrofuran	+	+
1948-33-0	t-Butylhydroquinone	-	+
100-41-4	Ethylbenzene	+	-
126-99-8	Chloroprene	+	+
8003-22-3	D&C Yellow No. 11	+	-
78-84-2	Isobutyraldehyde	-	-
127-00-4	1-Chloro-2-Propanol	T.B.A.	+
11-42-2	Diethanolamine	T.B.A.	-
77-09-8	Phenolphthalein	+	-
110-86-1	Pyridine	T.B.A.	+
1300-72-7	Xylenesulfonic acid, Na	-	-
98-00-0	Furfuryl alcohol	T.B.A.	+
125-33-7	Primaclone	+	+
111-76-2	Ethylene glycol monobutyl ether	T.B.A.	-
115-11-7	Isobutene	T.B.A.	-
93-15-2	Methyleugenol	T.B.A.	-
434-07-1	Oxymetholone	T.B.A.	-
84-65-1	Anthraquinone	T.B.A.	+
518-82-1	Emodin	T.B.A.	+
5392-40-5	Citral	T.B.A.	-
104-55-2	Cinnamaldehyde	T.B.A.	-
10026-24-1 †	Cobalt sulfate heptahydrate	T.B.A.	+
1313-27-5 †	Molybdenum trioxide	T.B.A.	-
1303-00-0 †	Gallium arsenide	T.B.A.	-
7632-00-0 †	Sodium nitrite	T.B.A.	+
1314-62-1 †	Vanadium pentozide	T.B.A.	-

Figure 6: Progol predictions for PTE-2. The first column are the compound identifiers in the NTP database. The column headed "Actual" are tentative classifications from the NTP. Here the entry T.B.A. means "to be announced" – confirmed classifications will be available by July, 1998. The 5 compounds marked with a † are inorganic compounds.

Acknowledgements

This research was supported partly by the Esprit Basic Research Action Project ILP II, the SERC project project 'Experimental Application and Development of ILP' and an SERC Advanced Research Fellowship held by Stephen Muggleton. Stephen Muggleton is a Research Fellow of Wolfson College Oxford. R.D. King was at Imperial Cancer Research Fund during the course of much of the early work on this problem. We would also like to thank Professor Donald Michie and David Page for interesting and useful discussions concerning the use of ILP for predicting biological activity.

References

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

[2] D. Bahler and D. Bristol. The induction of rules for predicting chemical carcinogenesis. In *Proceedings of the 26th Hawaii International Conference on System Sciences*, Los Alamitos, 1993. IEEE Computer Society Press.

[3] D. Bahler and D. Bristol. The induction of rules for predicting chemical carcinogenesis in rodents. In L. Hunter, D. Searls, and J. Shavlick, editors, *Intelligent Systems for Molecular Biology-93*, pages 29–37. MA:AAI/MIT, Cambridge, MA, 1993.

[4] G. Bakale and R.D. McCreary. Prospective ke screening of potential carcinogens being tested in rodent bioassays by the US National Toxicology Program. *Mutagenesis*, 7:91–94, 1992.

[5] R. Benigni. Predicting chemical carcinogenesis in rodents: the state of the art in the light of a comparative exercise. *Mutation Research*, 334:103–113, 1995.

[6] D.W. Bristol, J.T. Wachsman, and A. Greenwell. The NIEHS Predictive-Toxicology Evaluation Project. *Environmental Health Perspectives*, pages 1001–1010, 1996. Supplement 3.

[7] A.K. Debnath, R.L Lopez de Compadre, G. Debnath, A.J. Schusterman, and C. Hansch. Structure-Activity Relationship of Mutagenic Aromatic and Heteroaromatic Nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34(2):786 – 797, 1991.

[8] K. Enslein, B.W. Blake, and H.H. Borgstedt. Predicition of probability of carcinogenecity for a set of ntp bioassays. *Mutagenesis*, 5:305–306, 1990.

[9] J.E. Huff, J.K. Haseman, and D.P. Rall. Scientific concepts, value and significance of chemical carcinogenesis studies. *Ann Rev Pharmacol Toxicol*, 31:621–652, 1991.

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

[11] R.D. King, S. Muggleton, A.Srinivasan, C. Feng, R.A. Lewis, and M.J.E. Sternberg. Drug design using inductive logic programming. In *Proceedings of the 26th Hawaii International Conference on System Sciences*, Los Alamitos, 1993. IEEE Computer Society Press.

[12] R.D. King, S.H. Muggleton, A. Srinivasan, and M.J.E. Sternberg. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proc. of the National Academy of Sciences*, 93:438–442, 1996.

[13] R.D. King, S.H. Muggleton, and M.J.E. Sternberg. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. of the National Academy of Sciences*, 89(23):11322–11326, 1992.

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

[15] H. Kubini. *QSAR: Hansch Analysis and Related Approaches.* VCH, New York, 1993.

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

[17] S. Muggleton. Inverse Entailment and Progol. *New Gen. Comput.*, 13:245–286, 1995.

[18] H.S. Rosenkranz and G. Klopman. Predicition of the carcinogenecity in rodents of chemicals currently being tested by the US National Toxicology Program. *Mutagenesis*, 5:425–432, 1990.

[19] D.M. Sanderson and C.G. Earnshaw. Computer prediction of possible toxic action from chemical structure. *Human Exp Toxicol*, 10:261–273, 1991.

[20] A. Srinivasan, Ross D. King, and Stephen Muggleton. The role of background knowledge: using a problem from chemistry to examine the performance of an ILP program. In N. Lavrač, E. Keravnou, and B. Zupan, editors, *Under review for Intelligent Data Analysis in Medicine and Pharmacology.* Kluwer Academic Press, 1996.

[21] A. Srinivasan, S.H. Muggleton, R.D. King, and M.J.E. Sternberg. Mutagenesis: ILP experiments in a non-determinate biological domain. In S. Wrobel, editor, *Proceedings of the Fourth International Inductive Logic Programming Workshop*. Gesellschaft fur Mathematik und Datenverarbeitung MBH, 1994. GMD-Studien Nr 237.

[22] A. Srinivasan, S.H. Muggleton, R.D. King, and M.J.E. Sternberg. Theories for mutagenicity: a study of first-order and feature based induction. *Artificial Intelligence*, 85:277–299, 1996.

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

A NTP compounds used in this study

The following tabulation lists all the compounds used in this study.

Compound Id.	Name	Class
117-79-3	2-Aminoanthraquinone	+
6109-97-3	3-Amino-9-ethylcarbazole HCl	+
82-28-0	1-Amino-2-methylanthraquinone	+
134-29-2	o-Anisidine HCl	+
5131-60-2	4-Chloro-m-phenylenediamine	+
95-83-0	4-Chloro-o-phenylenediamine	+
569-61-9	Cl Basic Red 9 HCl	+
2832-40-8	Cl Disperse Yellow 3	+
120-71-8	p-Cresidine	+
135-20-6	Cupferron	+
39156-41-7	2,4-Diaminoanisole Sulfate	+
95-80-7	2,4-Diaminotoluene	+
2784-94-3	HC Blue 1	+
22-66-71	Hydrazobenzine	+
13552-44-8	4,4'-Methylenedianiline 2HCl	+
129-15-7	2-Methyl-1-nitroanthraquinone	+
2243-62-1	1,5-Naphthalenediamine	+
139-94-6	Nithiazide	+
602-87-9	5-Nitroacenaphthene	+
99-59-2	5-Nitro-o-anisidine	+
1836-75-5	Nitrofen	+
156-10-5	p-Nitrosodiphenenylamine	+
101-80-4	4-4' Oxydianiline	+
136-40-3	Phenazopyridine HCl	+
139-61-1	4,4' Thiodianiline	+
636-21-5	o-Toluidine HCl	+
137-17-7	2,4,5, -Trimethylaniline	+
67-20-9	Nitrofurantoin	+
59-87-0	Nitrofurazone	+
26471-62-5	2,4-/2,6-Toluene Diisocyanate	+
20265-96-7	p-Chloroaniline HCl	+
20325-40-0	3,3'-Dimethoxybenzidine 2HCl	+
612-82-8	3,3'-Dimethylbenzidine 2HCl	+
142-04-1	Aniline HCl	+
103-33-3	Azobenzene	+
95-79-4	5-Chloro-o-toluidine	+
5160-02-1	D and C Red 9	+
91-93-0	3,3'-Dimethoxybenzidine-4-4'-diisocyanate	+
121-14-2	2,4-Dinotrotoluene	+
99-55-8	5-Nitro-o-toluidine	+
80-08-0	4,4'-Sulfonyldianiline	+
1582-09-8	Trifuralin	+
3165-93-3	4-Chloro-o-toluidine HCl	+
2475-45-8	Cl Disperse Blue 1	+
102-50-1	m-Cresidne	+
609-20-1	2,6-Dichloro-p-phenylenediamine	+
94-52-0	5(6)-Nitrobenzimadazole	+
842-07-9	C.I. Solvent yellow 14	+
17026-81-2	3-Amino-4-ethoxyacetanilide	+
119-34-6	4-Amino-2-nitrophenol	+
121-66-4	2-Amino-5-nitrothiazole	+
105-11-3	p-Benzoquinone Dioxime	+
2185-92-4	2-Biphenylamine HCl	+

Compound Id.	Name	Class
133-90-4	Chloramben	+
1777-84-0	3-Nitro-p-acetophenetide	+
5307-14-2	2-Nitro-p-phenylenediamine	+
99-57-0	2-Amino-4-nitrophenol	+
121-88-0	2-Amino-5-nitrophenol	+
6373-74-6	C.I. Acid Orange 3	+
20265-97-8	p-Anisidine HCl	-
106-47-8	p-Chloroaniline	-
56-38-2	Parathion	-
952-23-8	Proflavin HCl	-
2871-01-4	H.C. Red 3	-
135-88-6	N-Phenyl-2-Napthylamine	-
121-19-7	Roxarsone	-
989-38-8	Rhodamine 6G HCl	-
140-49-8	4-(Chloroacetyl)acetanilide	-
61702-44-1	2-Chloro-p-phenylenediamine Sulfate	-
95-74-9	3-Chloro-p-toluidine	-
54150-69-5	2-4-Dimethoxyaniline HCl	-
298-00-0	Methyl Parathion	-
619-17-0	4-Nitroanthanilic Acid	-
99-56-9	4-Nitro-o-phenylenediamine	-
101-54-2	N-Phenyl-p-pheneyllenediamine	-
15481-70-6	2,6-Toluenediamine 2HCl	-
1936-15-8	C. I. Acid Orange 10	-
6358-85-6	Diarylanilide Yellow	-
33229-34-4	HC Blue 2	-
1465-25-4	N-(1-Napthyl)ethylenediamine 2HCL	-
86-57-7	1-Nitroaphthalene	-
624-18-0	p-Phenylenediamine 2HCl	-
127-69-5	Sulfisoxazole	-
6369-59-1	2,5-Toluenediamine Sulfate	-
63449-39-8	Chloronated paraffins C12: 60% Cl)	+
57653-85-7	Hexachlorodibenzodioxin 1	+
57635-85-7	Hexachlorodibenzodioxin 2	+
67774-32-7	Polybrominated Biphenyl	+
1746-01-6	2,3,7,8-Tetrachlorodibenzo-p-dioxin	+
86-06-2	2,4,6-Trichlorophenol	+
115-28-6	Chlorendic Acid	+
106-46-7	1,4-Dichlorobenzene	+
127-18-4	Tetrachloroethylene	+
67-72-1	Hexachloroethane	+
87-86-5	Pentachlorophenol	+
79-00-5	1,1,2-Trichloroethane	+
150-68-5	Monuron	+
12789-03-6	Chlordane	+
510-15-6	Chlorobenzilate	+
1897-45-6	Chlorothalonil	+
1163-19-5	Decabromodiphenyl Oxide	+
72-55-9	Dichlorodiphenyldichloroethylene	+
76-44-8	Heptachlor	+
76-01-7	Pentachloroethane	+
630-20-6	1,1,1,2-Tetrachloroethane	+
79-34-5	1,1,2,2-Tetrachloroethane	+
79-01-6	Trichloroethylene	+
309-00-2	Aldrin	+

Compound Id.	Name	Class
63449-39-8	Chlorinated paraffins (C23:43% Cl)	+
115-32-2	Dicofol	+
54-31-9	Furosemide	+
108-90-7	Chlorobenzene	-
33857-26-0	2,7-Dichlorodibenzo-p-dioxin	-
60-57-1	Dieldrin	-
72-56-0	Di(p-ethylphenyl)dichloroethane	-
1918-02-1	Picloram	-
72-54-8	Tetrachlorodiphenylethane	-
58-93-5	Hydrochlorothiazide	-
101-05-3	Anilazine	-
999-81-5	2-Chloroethyltrimethylammonium Chloride	-
95-50-1	1,2-Dichlorobenzene	-
72-20-8	Endrin	-
72-43-5	Methyoxychlor	-
77-65-6	Carbromal	-
94-20-2	Chlorpropamide	-
50-29-3	DDT	-
58-89-9	Lindane	-
82-68-8	Pentachloronitrobenzene	-
13366-73-9	Photodieldrin	-
75-35-4	Vinylidene Chloride	-
2698-41-1	o-Chlorobenzalmelanotrile	-
2438-88-2	2,3,5,6-Tetrachloro-4-nitroanisole	-
113-92-8	Chloropheniramine Maleate	-
120-83-2	2,4-Dichlorophenol	-
71-43-2	Benzene	+
117-81-7	Di(2-ethylhexyl)phthalate	+
139-13-9	Nitrilotriacetic Acid	+
50-55-5	Reserpine	+
123-31-9	Hydroquinone	+
2432-99-7	11-Aminoundeconic Acid	+
17924-92-4	Zearalenone	+
140-11-4	Benzyl Acetate	+
149-30-4	2-Mercaptobenzothiazole	+
389-08-2	Nalidixic Acid	+
103-23-1	Di(2-ethylhexyl)adipate	+
85-68-7	Butyl Benzyl Phthalate	+
120-62-7	Piperonyl Sulfoxide	+
78-42-2	Tris(2-ethylhexyl)phosphate	+
98-85-1	a-Methylbenzyl Alchohol	+
80-05-7	Bisphenol A	-
120-61-6	Dimethyl Terephthalate	-
121-79-9	Propyl Gallate	-
7177-48-2	Ampicillin Trihydrate	-
136-77-6	4-Hexylresorcinol	-
41372-08-1	Methyldopa Sesquihydrate	-

Compound Id.	Name	Class
2058-46-0	Oxytetracycline Hydrochloride	-
83-79-4	Rotenone	-
147-24-0	Diphenhydramine HCl	-
968-81-0	Acetohexamide	-
50-81-7	L-Ascorbic Acid	-
128-37-0	Butylated Hydroxytoluene	-
262-12-4	Dibenzo-p-dioxin	-
150-38-9	EDTA (tri-Na salt)	-
9002-18-0	Agar	-
119-53-9	Benzoin	-
105-60-2	Caprolactam	-
134-72-5	Ephedrine Sulfate	-
15356-70-4	di-Menthol	-
108-95-2	Phenol	-
85-44-9	Phthalic Anhydride	-
1156-19-0	Tolazamide	-
76-87-9	Triphenyltin Hydroxide	-
434-13-9	Lithocholic Acid	-
69-65-8	D-Mannitol	-
114-86-3	Phenformin	-
88-96-0	Phthalamide	-
51-03-6	Piperonyl Butoxide	-
64-77-7	Tolbutamide	-
73-22-3	L-Tryptophan	-
100-51-6	Benzyl Alchohol	-
132-98-9	Penicilin VK	-
64-75-5	Tetracycline Hydrochloride	-
108-30-5	Succinic Anhydride	-
643-22-1	Erithromycin Stearate	-
61-76-7	Phenylephrine Hydrochloride	-
1330-20-7	Xylenes Commercial Mixture	-
55-31-2	L-Epinephrine Hydrochloride	-
108-88-3	Toluene	-
2835-39-4	Allyl Isovalerate	+
87-29-6	Cinnamyl Anthranilate	+
123-91-1	1,4-Dioxane	+
271-89-6	Benzofuran	+
98-01-1	Furfural	+
50-33-9	Phenylbutazone	+
105-55-5	N,N'-Diethylthiourea	+
86-30-6	N-Nitrosodiphenylamine	+
100-52-7	Benzaldehyde	+
128-66-5	Cl Vat Yellow 4	+
78-59-1	Isophorone	+
108-78-1	Melamine	+
2489-77-2	Trimethylthiourea	+
137-30-4	Ziram	+
5989-27-5	a-Limonene	+
131-17-9	Diallyl Phthalate	-
142-46-1	2,5-Dithiobiurea	-
20941-65-5	Ethyl Tellurac	-
97-53-0	Eugenol	-
2164-17-1	Fluometuron	-
116-06-3	Aldicarb	-
3567-69-9	Cl Acid Red 14	-
118-92-3	o-Anthranilic Acid	-

Compound Id.	Name	Class
1212-29-9	N,N'-Dicyclohexylthiourea	-
536-33-4	Ethionamide	-
19010-66-3	Lead Dimethyldithiocarbamate	-
89-25-8	1-Phenyl-3-methyl-5-pyrazolone	-
148-18-5	Sodium Diethyldithiocarbamate	-
97-77-8	Tetraethylthiuram Disulfate	-
–	Vinyl Toluenes (meta/papa 70:30)	-
2783-94-0	FD and C Yellow 6	-
315-18-4	Mexacarbate	-
105-85-5	1-Phenyl-2-thiourea	-
77-79-2	3-Sulfolene	-
105-87-3	Geranyl Acetate	-
6959-48-4	3-Chloromethylpyridine HCl	+
96-12-8	1,2-Dibromo-3-chloropropane	+
106-93-4	1,2-Dibromoethane	+
107-06-2	1,2-Dichloroethane	+
542-75-6	1,3-Dichloropropene	+
3546-10-9	Phenestrin	+
75-56-9	1,2-Propylene Oxide	+
961-11-5	Tetrachlorovinphos	+
512-56-1	Trimethylphosphate	+
126-72-7	Tris(2,3-dibromopropyl)phosphate	+
563-47-3	3,Chloro-2-Methylpropene	+
62-73-7	Dichlorovos	+
101-90-6	Diglycidyl Resorcinol Ether (DRGE)	+
74-96-4	Bromoethane	+
556-52-5	Glycidol	+
5634-39-9	Iodinated Glycerol	+
106-87-6	4-Vinyl-1-cyclohexene Diepoxide	+
108-60-1	Bis(2-chloro-1-methyethyl)ether	+
868-85-9	Dimethyl Hydrogen Phosphite	+
106-88-7	1,2-Epoxybutane	+
22966-79-6	Estradiol Mustard	+
597-25-1	Dimethyl Morpholinophosphoramidate	+
1955-45-9	Pivalolactone	+
8001-35-2	Toxaphene	+
78-87-5	1,2-Dichloropropane	+
115-96-8	Tris(2-chloroethyl)phosphate	+
57-06-7	Allyl Isothiocyanate	+
756-79-6	Dimethyl Methylphosphonate	+
106-92-3	Allyl Glcidyl Ether	+
75-00-3	Chloroethane	+
86-50-0	Azinphosmethyl	-
55-38-9	Fenthion	-
13171-21-6	Phosphamidon	-
532-27-4	2-Chloroacetophenone	-
78-11-5	Pentaerythritol Tetranitrate	-
109-69-3	n-Butyl Chloride	-
107-07-3	2-Chloroethanol	-
56-72-4	Coumaphos	-
60-51-5	Dimethoate	-
1634-78-2	Malaoxon	-
124-64-1	Terakis (Hydroxymethyl) Phosphonium Chloride/Sulfate	-
6959-47-3	2-Chloromethylpyridine HCl	-

Compound Id.	Name	Class
333-41-5	Diazinon	-
78-34-2	Dioxathion	-
121-75-5	Malathion	-
75-09-2	Dichloromethane	+
75-27-4	Bromodichloromethane	+
75-25-2	Trbromomethane (bromoform)	+
124-48-1	Chlorodibromomethane	+
75-47-8	Iodoform	-
101-61-1	4,4'-Methylenebis(N,N'-dimethyl)benzenamine	+
90-94-8	Michler's Ketone	+
121-69-7	N,N-Dimethylaniline	+
140-56-7	Fenaminosulf	-
509-14-8	Tetranitromethane	+
504-88-1	3-Nitropropionic Acid	-
140-88-5	Ethyl Acrylate	+
924-42-5	N-Methylolacrylamide	+
80-62-6	Methyl Methacrylate	-
24382-04-5	Malonaldehyde Sodium Salt	+
828-00-2	Dimethoxane	-
95-06-7	Sulfallate	+
513-37-1	Dimethylvinyl Chloride (DMVC)	+
133-06-2	Captan	+
598-55-0	Methyl Carbamate	+
1596-84-5	Succinic Acid 2,2-dimethylhydrazide	+
95-14-7	1,2,3-Benzotriazole	-
148-24-3	8-Hydroxyquinoline	-
115-07-1	Propylene	-
60-13-9	Amphetamine sulfate	-
91-20-3	Napthalene	+
9005-65-6	Polysorbate 80 (Tween 80)	-
58-33-3	Promethazine hydrochloride	-
108-46-3	Resorcinol	-
96-48-0	g-Butyrolactone	-
79-11-8	Monochloroacetic acid	-
100-02-7	p-Nitrophenol	-
1330-78-5	Tricresyl phosphate	+
120-32-1	o-Benzyl-p-chlorophenol	+
3296-90-0	2,2-Bis (bromomethyl)-1,3,-propanediol	+
75-65-0	t-Butyl alchol	+
119-84-6	3,4-Dihydrococoumarin	+
107-21-1	Ethylene glcol	-
298-59-9	Methylphenidate hydrochloride	+
96-69-5	4,4'-Thiobis(6-t-butyl-m-cresol)	-
396-01-0	Triamterene	+
57-41-0	Diphenylhydantoin	+
1825-21-4	Pentachloroanisole	+
10599-90-3	Chloramine	-
81-11-8	4,4'-Diamino-2,2'stilbenedisulfonic acid	-
74-83-9	Methyl Bromide	-
62-23-7	p-Nitrobenzoic acid	+
28407-37-6	Cl Direct blue 218	+
2425-85-6	Cl Pigment red 3	+
6471-49-4	Cl Pigment red 23	-
137-09-7	2,4-Diaminophenol dihydrochloride	+
103-90-2	4-Hydroxyacetanilide	-

Compound Id.	Name	Class
1271-19-8	Salicylazosulfapyridine	-
6459-94-5	Cl Acid red 114	+
2429-74-5	Cl Direct blue 15	+
91-64-5	Coumarin	+
96-13-9	2,3-Dibromo-1-propanol	+
119-93-7	3,3'-Dimethylbenzidine	+
52551-67-4	HC Yellow 4	-
100-01-6	p-Nitroaniline	-
91-23-6	o-Nitroanisole	+
96-18-4	1,2,3-Trichloropropane	+