The Predictive Toxicology Evaluation Challenge

A. Srinivasan
S.H. Muggleton
Oxford University Computing Laboratory
Wolfson Building Parks Road, Oxford
UK

R.D. King*
M.J.E. Sternberg
Biomolecular Modelling Laboratory
Imperial Cancer Research Fund
44 Lincoln's Inn Fields, London
U.K.

Abstract

Can an AI program contribute to scientific discovery? An area where this gauntlet has been thrown is that of understanding the mechanisms of chemical carcinogenesis. One approach is to obtain Structure-Activity Relationships (SARs) relating molecular structure to cancerous activity. Vital to this are the rodent carcinogenicity tests conducted within the US National Toxicology Program by the National Institute of Environmental Health Sciences (NIEHS). This has resulted in a large database of compounds classified as carcinogens or otherwise. The Predictive-Toxicology Evaluation project of the NIEHS provides the opportunity to compare carcinogenicity predictions on previously untested chemicals. This presents a formidable challenge for programs concerned with knowledge discovery. Desirable features of this problem are: (1) involvement in genuine scientific discovery; (2) availability of a large database with expert-certified classifications; (3) strong competition from methods used by chemists; and (4) participation in true blind trials, with results available by next IJCAI. We describe the materials and methods constituting this challenge, and provide some initial benchmarks. These show the Inductive Logic Programming tool Progol to be competitive with current state-of-the-art. The challenge described here is aimed at encouraging AI programs to avail themselves the opportunity of contributing to an enterprise with immediate scientific value.

1 Introduction

Programs developed under the umbrella of Machine Learning are increasingly being used for "knowledge discovery" tasks. Early specialised programs (for example, [Feigenbaum *et al.*, 1971; Langley *et al.*, 1983]) have given way to more general-purpose ones (for example, [Muggleton, 1995; Muggleton and Feng, 1990]) which have been applied with some success in areas of biochemistry ([King *et al.*, 1996; 1992; Muggleton *et al.*, 1992]). While the experimental studies reported are preliminary, they have at least one commendable feature, namely, they constitute examples of AI programs participating in true scientific discovery tasks. By "true" here, we mean problems where existing scientific knowledge is incomplete, the descriptions found automatically were unknown to experts in the field, and have been acknowledged by publication in peer-reviewed journals in the field. Given the promise shown by machine learning programs in biology and chemistry, this paper describes a challenging test-bed with the following desirable features: (1) a widespread scientific interest in any new results; (2) the availability of a large database of chemicals with classifications certified by experts; (3) strong competition from methods developed by expert chemists; and (4) the opportunity to participate in true blind trials.

The problem concerns obtaining a better understanding of the molecular mechanisms of chemical carcinogenesis. This is central to the prevention of many environmentally induced cancers. One approach is to form Structure Activity Relationships (SARs) that empirically relate molecular structure with ability to cause cancer. This work has been greatly advanced by the long term carcinogenicity tests of compounds in rodents (utilising both genders of one rat and mouse strain) by the US National Toxicology Program (NTP) of the National Institute of Environmental Health Sciences (NIEHS: [Huff and Haseman, 1991]). So far, the NTP tests have resulted in a database of more than 300 compounds that have been shown to be carcinogens or otherwise. The NIEHS Predictive Toxicology Evaluation (or PTE) project ([Bristol *et al.*, 1996]) is closely associated with the NTP. The PTE project identifies a group of assays that are scheduled or ongoing in the NTP. These chemicals form the "test" set for researchers. Predictions

*Now at:Department of Computer Science, The University of Wales Aberystwyth

for presence or absence of carcinogenicity activity are compared against true activity as observed in the rodents. The first such blind trial, PTE-1, is now complete. The second, PTE-2, is ongoing, and true activity levels will be available by June 1998. It is the prediction of, and reasons for, carcinogenic activity in chemicals constituting PTE-2 that we commend as a challenge for AI programs concerned with knowledge discovery from databases.

This paper is organised as follows. Section 2 presents the statement of the challenge. Section 3 summarises the data available in the NTP database, and the chemicals in PTE-1 and PTE-2. Section 4 sets out the evaluation criteria. Section 5 describes the results obtained using the Inductive Logic Programming (ILP) system Progol [Muggleton, 1995]. These results are intended to provide initial benchmarks for future entries. Section 6 concludes this paper including information on submitting entries to the challenge.

2 The PTE Challenge

The aim is to obtain a theory for predicting the carcinogenicity of 30 compounds currently undergoing rodent bioassays in the NTP (called PTE-2: see below). The performance of the theory is to be evaluated according to the criteria described in Section 4.

3 Materials

3.1 The NTP database

A compilation of 330 chemicals is available directly from the database of the National Cancer Institute and NTP ([Huff and Haseman, 1991], and via the Internet at *http://ntp-server.niehs.nih.gov/*. These compounds represent all the organic chemicals that have completed NTP reports at the time of writing this paper. Of the 330 compounds, 182 (55%) are classified carcinogenic, and the remaining 148 non-carcinogenic. Carcinogenicity is determined by analysis of long term rodent bioassays. For the purposes of this challenge, compounds classified by the NTP as equivocal are considered non-carcinogenic, as this allows direct comparison with other SAR predictive methods. No analysis is made of differences in incidence between rat and mouse cancer, or the role of sex, or particular organ sites. 39 of the 330 compounds in the NTP database formed the first of the blind trials (PTE-1) conducted by the PTE project. Results from the bioassays for these chemicals are now available, and show 22 (56%) to be carcinogenic, and the remaining 17 to be non-carcinogenic. Further details of these compounds are available in [Bahler and Bristol, 1993]. The 330 chemicals make this database very large and diverse, making it a great challange to learn in.

The PTE-2 compounds

The second round of blind trials (PTE-2) consists of 30 compounds (of which 5 are inorganic). These are fully described in [Bristol *et al.*, 1996], where the schedule of events suggest that all bioassay results will be available by July 1998.

3.2 Other information available

The NTP has recently made available a number of structural attributes (features) describing a large section of their database. These descriptions are available at *http://ntp-server.niehs.nih.gov/Main_Pages*. The other information available is in the form of the atom and bond connectivity of the compounds (including those in PTE-1,2). This is described further in Section 5.

4 Evaluation

In [Bristol *et al.*, 1996], the goal of predicitive toxicology (PT) is summarised as "...the ultimate value and most important goal of PT research may lie in the development of its potential to 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 ...". Given this emphasis on understandability of models, we follow [Muggleton *et al.*, 1996] in using the following definition for comparing the performance of rival theories.

Definition 1 Performance comparison. *If the predictive accuracies of two theories are statistically equivalent then the theory with better explanatory power has better performance. Otherwise the one with higher accuracy has better performance.*

We now elaborate further on the methods for evaluating predictive accuracy and "explanatory power".

4.1 Predictive accuracy

Predictive accuracy is taken to be the proportion of compounds in PTE-2 whose predicted classification (carcinogenic or non-carcinogenic) agrees with that rodent bioassays. Significant differences in predictive accuracy are best assessed by McNemar test for changes [Bland, 1989]. This test exploits the fact that the different prediction methods are applied to the same data and is based on counting the examples where the methods disagree about predictions. We suggest that differences be judged to be significant at least at $P = 0.10$.

4.2 Explanatory power

In the absence of an expert chemist to act as adjudicator, we propose that the simple criterion that theories are judged to have "explanatory power" – a boolean property – if some or all of it can be represented diagrammatically as chemical structures. The intuition underlying this is that such structural alerts form the preferred mode of discourse amongst chemists.

5 An experiment with the ILP system Progol

5.1 Progol

We refer the reader to [Muggleton, 1995] for complete descriptions of the ILP system Progol. In the current con-

text, Progol is provided with a set of carcinogenic ("positive") and non-carcinogenic ("negative") examples from the NTP database together with background knowledge B about these compounds (see Section 5.2). The aim is to generate a theory (expressed as a set of rules) which explains all the carcinogens in terms of the background knowledge whilst remaining consistent with the non-carcinogens. To achieve this Progol 1) randomly selects a positive example e_i; 2) uses inverse entailment to construct the most specific hypothesis $\bot(B, e_i)$ which explains e_i in terms of B; 3) finds a rule D_i which generalises $\bot(B, e_i)$ and which maximally compresses a set of entailed example E_i; and 4) adds D_i to the theory H and repeats from 1) with examples not covered so far until no more compression is possible. Compression is here defined as the difference, in numbers of descriptors, between E_i and D_i.

5.2 Background knowledge

The generic atom/bond representation used in an earlier study [King *et al.*, 1996; Srinivasan *et al.*, 1996] is used. This consists of two basic relations to represent structure: *atom* and *bond*. For example, the fact *atom(127,127_1,c,ar_c_6_ring,-0.133)* states that in compound 127, atom no. 1 is of element carbon, and of type aromatic carbon in a 6 membered ring, and has a partial charge of -0.133. The type of the atom and its partial charge were taken from the molecular modelling package QUANTA, although any similar modelling package would have been suitable. Equivalently, *bond(127,127_1,127_2, ar)* states that in compound 127, atom no. 1 and atom no. 2 are connected by an aromatic bond. In QUANTA, a partial charge assignment is based on a specific molecular neighbourhood. This has the effect that a specific molecular sub- structure can be identified by an atom type and partial charge. This relational representation is completely general for chemical compounds and no special attributes need to be invented. The structural information of these compounds was represented by $\approx$ 18,300 facts of background knowledge.

Information was also given about the results of Salmonella mutagenicity tests for each compound. The mutagenic compounds were represented by the relation Ames, e.g. *ames(127)* states that compound 127 is mutagenic. The Progol algorithm allows for the inclusion of complex background knowledge, either in the form of facts, or in the form of arbitrary Prolog programs. In this study we included the background knowledge of chemical groups defined in [Srinivasan *et al.*, 1996], along with the structural alerts in [Ashby *et al.*, 1989]. All the information used is available in Prolog form at a prescribed Internet site: *http://www.comlab.ox.ac.uk/oucl/groups/machlearn*.

5.3 Results and discussion

The 39 compounds comprising PTE-1 were excluded and rules for carcinogenicity obtained using Progol. The resulting theory consists of 18 rules. Figure 1 tabulates a comparative evaluation on PTE-1. More details on the rules obtained are available in [King and Srinivasan, 1996]. Figure 2 tabulates the predictions made by the Progol theory for compounds in PTE-2. The first three entries in Figure 1 have been marked out for special attention because they had access to additional information in the form of short-term rodent (*in-vivo*) tests. The first two entries also require a degree of expert evaluation. The Ashby structural alerts are based on electrophilic attack on DNA, which makes them statistically dependent on the Ames test. It is also worth noting that the TIPT and Benigni methods rely on structural alerts derived by the Ashby method for their explanatory component. CASE, TIPT and Progol are the only data-driven inductive methods, and Progol is the only automated method capable of identifying new structural alerts. With these comments in place, the results in Figure 1 offer significant encouragement for machine learning programs on the following counts. First, we point out that one PTE-1, the results of Progol, TIPT and CASE demonstrate performance that is competitive with the current state-of-the-art – Progol has marginally the highest accuracy of all methods that do not use rodent tests. Second, the relatively low accuracies of all methods is primarily due to the diversity of compounds involved. It does however leave the door open for significant improvement. Progol, for example, achieves its performance with a very low-level atom/bond representation of compounds. Enriching this background information with the new structural descriptors available from the NTP could significantly improve its accuracy. These comments are reinforced by early results of using the Progol theory to classify compounds in PTE-2. Figure 2 shows that tentative classifications are available from the NTP for 13 chemicals. Progol's theory has correctly classified 7 of these. This should be seen in context of the performance of other theories listed in [Bristol *et al.*, 1996] which shows that most chemists and automated methods have not been able to better this count. This should provide further impetus for participation by other AI programs.

6 Conclusions

The field of toxicology is a rich source of difficult scientific problems, and there is a pressing need for analysis methods that can advance our understanding of the issues involved. We believe that the Predictive Toxicology Evaluation trials being conducted by the US National Institute for Environmental Health Sciences afford AI programs a unique opportunity to participate in obtaining an improved understanding of the molecular mechanisms underlying chemical carcinogenesis. Should they be successful, it would also constitute a noteworthy example of a realistic application of AI techniques.

Entering the PTE Challenge

Entries to the PTE Challenge can be submitted via *http://www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE*.

Method	Type	Accuracy	Explanation	Performance ranking
Ashby [Tennant *et al.*, 1990]	Chemist	0.77	yes	1
RASH [Jones and Easterly, 1991]	Biological potency analysis	0.72	yes	2
TIPT [Bahler and Bristol, 1993]	Propositional machine learning	0.67	yes	2
Progol [Muggleton, 1995]	Inductive logic programming	0.64	yes	2
Benigni [Benigni, 1995]	Expert-guided regression	0.62	yes	2
DEREK [Sanderson and Earnshaw, 1991]	Expert system	0.57	yes	2
Bakale [Bakale and McCreary, 1992]	Chemical reactivity analysis	0.63	no	3
TOPKAT [Enslein *et al.*, 1990]	Statistical discrimination	0.54	yes	4
CASE [Rosenkranz and Klopman, 1990]	Statistical correlation analysis	0.54	yes	4
COMPACT [Lewis *et al.*, 1990]	Molecular modelling	0.54	yes	4

Figure 1: Benchmarks on PTE-1. Methods above the central horizontal line had access to short-term rodent tests, which were unavailable to others. Further, the Ashby and RASH methods require a degree of subjective evaluation, making them semi-automatic. The performance ranking is obtained using the combined accuracy-explanation criterion described earlier.

Entries are accepted here for compounds in PTE-2. A submission requires the following: (a) name of entry; (b) predictions for the compounds in PTE-2; (c) whether or not the theory has explanatory power; and (d) a short description of the technique used for prediction. This is sufficient to compute automatically the accuracy and performance ranking of each entry.

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

[Ashby *et al.*, 1989] J. Ashby, R.W. Tennant, E. Zeiger, and S. Stasiewicz. Classification according to chemical structure, mutagenicity to salmonella and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. *Mutation Research*, 223:73–103, 1989.

[Bahler and Bristol, 1993] 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.

[Bakale and McCreary, 1992] 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.

[Benigni, 1995] 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.

[Bland, 1989] M. Bland. *An Introduction to Medical Statistics*. Oxford University Press, Oxford, 1989.

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

[Enslein *et al.*, 1990] 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.

[Feigenbaum *et al.*, 1971] E.A. Feigenbaum, B.G. Buchanan, and J. Lederberg. On generality and problem solving: a case study using the DENDRAL program. In D. Michie, editor, *Machine Intelligence 6*. Edinburgh University Press, Edinburgh, 1971.

[Huff and Haseman, 1991] J. Huff and J. Haseman. Long-term chemical carcinogenesis experiments for identifying potential human cancer hazards. *Environmental Health Perspectives*, 96(3):23–31, 1991.

[Jones and Easterly, 1991] 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.

[King and Srinivasan, 1996] 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.

[King *et al.*, 1992] 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

Compound Id.	Name	Actual	Progol prediction
6533-68-2	Scopolamine hydrobroamide	-	+
147-47-7	1,2-Dihydro-2,2,4-trimethyquinoline	+	+
8003-22-3	D&C Yellow No. 11	+	+
78-84-2	Isobutyraldehyde	-	+
125-33-7	Primaclone	+	+
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.	+
76-57-3	Codeine	-	-
75-52-8	Nitromethane	-	-
109-99-9	Tetrahydrofuran	+	-
1948-33-0	t-Butylhydroquinone	-	-
100-41-4	Ethylbenzene	+	-
126-99-8	Chloroprene	+	-
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.	-
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.	-
10026-24-1	Cobalt sulfate heptahydrate	T.B.A.	not predicted
1313-27-5	Molybdenum trioxide	T.B.A.	not predicted
1303-00-0	Gallium arsenide	T.B.A.	not predicted
7632-00-0	Sodium nitrite	T.B.A.	not predicted
1314-62-1	Vanadium pentozide	T.B.A.	not predicted

Figure 2: 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. An entry "+" means carcinogenic, and "-" means non-carcinogenic. The 5 compounds not predicted are inorganic compounds – Progol's rules are applicable to organic compounds only.

structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. of the National Academy of Sciences*, 89(23):11322–11326, 1992.

[King *et al.*, 1996] 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.

[Langley *et al.*, 1983] P. Langley, G.L Bradshaw, and H. Simon. Rediscovering chemistry with the Bacon system. In R. Michalski, J. Carbonnel, and T. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach*, pages 307–330. Tioga, Palo Alto, CA, 1983.

[Lewis *et al.*, 1990] 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.

[Muggleton and Feng, 1990] S.H. Muggleton and C. Feng. Efficient induction of logic programs. In *Proceedings of the First Conference on Algorithmic Learning Theory*, Tokyo, 1990. Ohmsha.

[Muggleton *et al.*, 1992] S. Muggleton, R. King, and M. Sternberg. Predicting protein secondary structure using inductive logic programming. *Protein Engineering*, 5:647–657, 1992.

[Muggleton *et al.*, 1996] S.H. Muggleton, A. Srinivasan, R.D. King, and M.J.E. Sternberg. Biochemical knowledge discovery using Inductive Logic Programming. In R. Michalski, M. Kubat, and I. Bratko, editors, *Methods and Applications of Machine Learning, Data Mining and Knowledge Discovery*. John Wiley, 1996.

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

[Rosenkranz and Klopman, 1990] 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.

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

[Srinivasan *et al.*, 1996] 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.

[Tennant *et al.*, 1990] 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.