Prediction of Rodent Carcinogenicity Bioassays from Molecular Structure Using Inductive Logic Programming

Ross D. King[1] and Ashwin Srinivasan[2]

[1]Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, London, United Kingdom; [2]Computing Laboratory, University of Oxford, Oxford, United Kingdom

The machine learning program *Progol* was applied to the problem of forming the structure–activity relationship (SAR) for a set of compounds tested for carcinogenicity in rodent bioassays by the U.S. National Toxicology Program (NTP). *Progol* is the first inductive logic programming (ILP) algorithm to use a fully relational method for describing chemical structure in SARs, based on using atoms and their bond connectivities. *Progol* is well suited to forming SARs for carcinogenicity as it is designed to produce easily understandable rules (structural alerts) for sets of noncongeneric compounds. The *Progol* SAR method was tested by prediction of a set of compounds that have been widely predicted by other SAR methods (the compounds used in the NTP's first round of carcinogenesis predictions). For these compounds no method (human or machine) was significantly more accurate than *Progol*. *Progol* was the most accurate method that did not use data from biological tests on rodents (however, the difference in accuracy is not significant). The *Progol* predictions were based solely on chemical structure and the results of tests for *Salmonella* mutagenicity. Using the full NTP database, the prediction accuracy of *Progol* was estimated to be 63% (±3%) using 5-fold cross validation. A set of structural alerts for carcinogenesis was automatically generated and the chemical rationale for them investigated—these structural alerts are statistically independent of the *Salmonella* mutagenicity. Carcinogenicity is predicted for the compounds used in the NTP's second round of carcinogenesis predictions. The results for prediction of carcinogenesis, taken together with the previous successful applications of predicting mutagenicity in nitroaromatic compounds, and inhibition of angiogenesis by suramin analogues, show that *Progol* has a role to play in understanding the SARs of cancer-related compounds. — Environ Health Perspect 104(Suppl 5):1031–1040 (1996)

Key words: machine learning, artificial intelligence, SAR, National Toxicology Program

Introduction

An understanding of the molecular mechanisms of chemical carcinogenesis 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 by the National Toxicology Program (NTP) of the National Institute of Environmental Health Sciences (*1*). These tests have resulted in a database of more than 300 compounds that have been shown to be carcinogens or noncarcinogens. The database of compounds can be used to form general SARs relating molecular structure to formation of cancer.

The compounds in the NTP database present a problem for many conventional SAR techniques because the compounds in the NTP databases are structurally very diverse, and many different molecular mechanisms are involved. Most conventional SAR methods are designed to deal with compounds having a common molecular template and presumed similar molecular mechanisms of action—congeneric compounds. Numerous approaches have been taken to forming SARs for carcinogenesis. Ashby and co-workers (*2–4*) developed a successful semiobjective method of predicting carcinogenesis based on the identification of chemical substructures (alerts) that are associated with carcinogenesis. A similar but more objective approach was taken by Sanderson and Earnshaw (*5*), who developed an expert system based on rules obtained from expert chemists. An inductive approach, not directly based on expert chemical knowledge, is the computer-automated structure evaluation (CASE) system (*6,7*). This system empirically identifies structural alerts that are statistically related to a particular activity. A number of other approaches have been applied based on a variety of sources of information and SAR learning methods (*8–13*). The effectiveness of these different SAR methods was evaluated on a test set of compounds for which predictions were made before the trials were completed (round 1 of the NTP's tests for carcinogenesis prediction) (*8,14,15*) There is currently a second round of tests.

The machine-learning methodology Inductive Logic Programming (ILP) has been applied to a number of SAR problems. Initial work was done using the program *Golem* to form SARs for the inhibition of dihydrofolate reductase by pyrimidines (*16–18*). This work was extended by the development of the program *Progol* (*19*) and its adaptation for application to noncongeneric SAR problems (*20*). *Progol* has been successfully applied to predicting the mutagenicity of a series of structurally diverse nitroaromatic compounds (*21*), and the inhibition of angiogenesis by suramin analogues (*20*). The *Progol* SAR method is designed to produce easily understandable rules (structural alerts). For the nitroaromatic and suramin compounds the rules generated provided insight into the chemical basis of action.

Most existing SAR methods describe chemical structure using attributes—general properties of objects. Such descriptions can be displayed in tabular form, with the compounds along one dimension and the attributes along the other dimension. This type of description is very inefficient at representing structural information. A more general method of describing chemical structure is to use logical statements, or relations. This method is also clearer, as chemists are used to relating chemical properties

This paper is part of the NIEHS Predictive-Toxicology Evaluation Project. Manuscript received May 8, 1996; manuscript accepted 9 August 1996.

We thank M.J.E. Sternberg and S.H. Muggleton. This work was supported by ESPRIT (6020), the SERC project, experimental application and development of ILP, and the Imperial Cancer Research Fund.

Address correspondence to R.D. King, Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, Lincoln's Inn Fields, P.O. Box 123, London, WC2A 3PX, U.K. Telephone: +44 171 269 3565. Fax: +44 171 269 3417. E-mail:rd_king@icrf.ac.uk

Abbreviations used: CASE, computer-augmented structure evaluation; DEREK, deductive estimation of risk from existing knowledge; ILP, inductive logic programming; MTD, minimally toxic dose; NTP, National Toxicology Program; QSAR, quantitative structure–activity relationship; SAR, structure–activity relationship.

and functions for groups of atoms. The *Progol* method is the first to use a general relational method for describing chemical structure in SARs. The method is based on using atoms and their bond connectivities and is simple, powerful, and generally applicable to any SAR. The method also appears robust and suited to SAR problems difficult to model conventionally (*21*).

The most similar approaches to *Progol* are those of CASE (*6*), MULTICASE (*7*), and the symbolic machine learning approaches of Bahler and Bristol (*8*) and Lee (*22*). However the *Progol* methodology is more general, as the other approaches are based on attributes and therefore have built-in limitations in representing structural relationships.

This article describes application of the *Progol* SAR method to predicting chemical carcinogenesis. *Progol* was first benchmarked on the test data of round 1 and then applied to produce predictions for round 2. The predictions for round 2 are completely blind trials. Such tests are very important because they ensure that the predictions are free from any conscious or unconscious bias.

Materials and Methods

Data

The compilation of 330 chemicals used in this study was taken from the literature (*2,3,8*) as well as directly from the collective database of the National Cancer Institute (NCI) and the NTP (*1*). The compounds used were all the organic compounds for which there were completed NTP reports at the time of this work. A listing of the compounds and their activities is given in Table 1. Inorganic compounds were not included because it was considered that there are too few of them to allow meaningful generalizations. Of the 330 compounds, 182 (55%) are classified as carcinogenic, and the remaining 148 as noncarcinogenic. Carcinogenicity is determined by analysis of long-term rodent bioassays. Compounds classified by the NTP as equivocal are considered noncarcinogenic, this allows direct comparison with other predictive methods. No analysis was made of differences in incidence between rat and mouse cancer, or the role of sex, or particular organ sites.

The *Progol* SAR method was first tested using the test data considered in the first round of the NTP trial (*3*). This allowed direct comparison with the results of many other SAR techniques (*8*). The training set consisted of 291 compounds, 161 (55%) carcinogens and 130 noncarcinogens. In addition to this train/test split, a 5-fold cross-validation split of the 330 compounds was tested for a more accurate estimate of the efficacy of *Progol*. The compounds were randomly split into five sets, and *Progol* was successively trained on four of the splits and tested on the remaining split.

Progol

In inductive logic programming (ILP) all the inputs and outputs are logical rules (*23*) in the computer language PROLOG. Such rules are easily understandable because they closely resemble natural language. For any application the input to *Progol* consists of a set of positive examples (i.e., for SAR, the active compounds), negative examples (i.e., nonactive compounds), and background knowledge about the problem (e.g., the atom/bond structure of the compounds) (Figure 1). *Progol* outputs consist of a hypothesis expressed as a set of rules that explain the positive and negative examples using the background knowledge. The rule found for each example is optimal in terms of simplicity (information compression) and the language used to describe the examples. This guarantee of optimality does not extend to sets of rules constructed by *Progol*, as it does not follow that a set of rules consisting of individually optimal rules is itself optimal for information compression. Information compression is defined as the difference in the amount of information needed to explain the examples with and without using the rule. It is statistically highly improbable that a rule with high compression does not represent a real pattern in the data (*24*). The use of compression balances accuracy (number of correct predictions/number of total predictions) and coverage (number of examples predicted by the rule/number of examples), i.e., it is a compromise between sensitivity and specificity. The validity of the compression measure was empirically shown by the results of the 5-fold cross-validation trial. *Progol* generates rules in a stepwise manner until all the examples are covered or no more compressive (statistically significant) rules can be found. A simple example of use of the *Progol* algorithm is given in the Appendix.

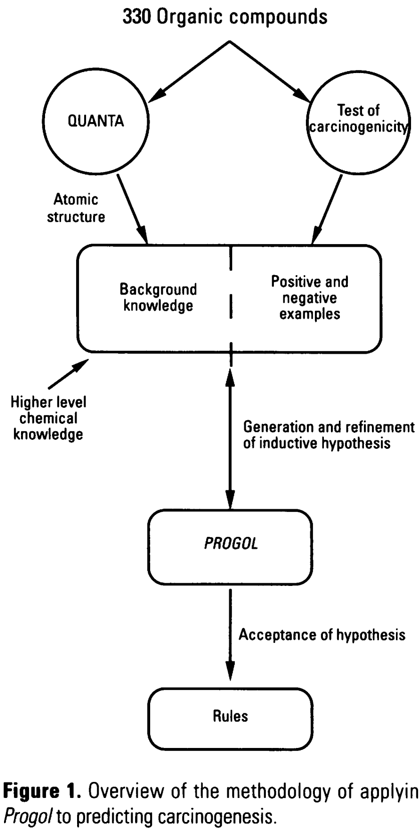

Figure 1. Overview of the methodology of applying *Progol* to predicting carcinogenesis.

Compound Representation for *Progol*

The generic atom/bond representation that we previously applied to mutagenesis was used (*21*). Two basic relations were utilized to represent structure: atom and bond. For example, for compound 1 (CAS no. 117-79-3),

atom(127, 127_1, carbon,
aromatic_carbon_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 modeling package QUANTA^{TM}; any similar modeling package would also have been suitable. Equivalently,

bond(127, 127_1, 127_2, aromatic)

states that in compound 127, atom no. 1 and atom no. 2 are connected by an aromatic bond. In QUANTA^{TM} partial charges assignment is based on a specific molecular neighborhood; this has the effect that a specific molecular substructure 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

Table 1. Compounds used in this trial and their carcinogenic status.

CAS no.	Name	Act.	CAS no.	Name	Act.	CAS no.	Name	Act.
117-79-3	2-Aminoanthraquinone	+	140-49-8	4-(Chloroacetyl)acetanilide	–	120-83-2	2,4-Dichlorophenol	–
6109-97-3	3-Amino-9-ethylcarbazole HCl	+	61702-44-1	2-Chloro-*p*-phenylenediamine sulfate	–	71-43-2	Benzene	+
82-28-0	1-Amino-2-methylanthraquinone	+				117-81-7	Di(2-ethylhexyl)phthalate	+
134-29-2	*o*-Anisidine HCl	+	95-74-9	3-Chloro-*p*-toluidine	–	139-13-9	Nitrilotriacetic acid	+
5131-60-2	4-Chloro-*m*-phenylenediamine	+	54150-69-5	2-4-Dimethoxyaniline HCl	–	50-55-5	Reserpine	+
95-83-0	4-Chloro-*o*-phenylenediamine	+	298-00-0	Methyl parathion	–	123-31-9	Hydroquinone	+
569-61-9	CI Basic Red 9 HCl	+	619-17-0	4-Nitroanthanilic Acid	–	2432-99-7	11-Aminoundeconic acid	+
2832-40-8	CI Disperse Yellow 3	+	99-56-9	4-Nitro-*o*-phenylenediamine	–	17924-92-4	Zearalenone	+
120-71-8	*p*-Cresidine	+	101-54-2	*N*-Phenyl-*p*-pheneylenediamine	–	140-11-4	Benzyl acetate	+
135-20-6	Cupferron	+	15481-70-6	2,6-Toluenediamine 2HCl	–	149-30-4	2-Mercaptobenzothiazole	+
39156-41-7	2,4-Diaminoanisole sulfate	+	1936-15-8	C.I. Acid Orange 10	–	389-08-2	Nalidixic acid	+
95-80-7	2,4-Diaminotoluene	+	6358-85-6	Diarylanilide Yellow	–	103-23-1	Di(2-ethylhexyl)adipate	+
2784-94-3	HC Blue 1	+	33229-34-4	HC Blue 2	–	85-68-7	Butyl benzyl phthalate	+
22-66-71	Hydrazobenzine	+	1465-25-4	*N*-(1-Napthyl)ethylenediamine 2HCL	–	120-62-7	Piperonyl sulfoxide	+
13552-44-8	4,4′-Methylenedianiline 2HCl	+	86-57-7	1-Nitronaphthalene	–	78-42-2	Tris(2-ethylhexyl)phosphate	+
129-15-7	2-Methyl-1-nitroanthraquinone	+	624-18-0	*p*-Phenylenediamine 2HCl	–	98-85-1	α-Methylbenzyl alchohol	+
2243-62-1	1,5-Naphthalenediamine	+	127-69-5	Sulfisoxazole	–	80-05-7	Bisphenol A	=
139-94-6	Nithiazide	+	6369-59-1	2,5-Toluenediamine sulfate	–	120-61-6	Dimethyl terephthalate	=
602-87-9	5-Nitroacenaphthene	+	63449-39-8	Chlorinated paraffins (C12: 60% Cl)	+	121-79-9	Propyl gallate	=
99-59-2	5-Nitro-*o*-anisidine	+	57653-85-7	Hexachlorodibenzodioxin 1	+	7177-48-2	Ampicillin trihydrate	=
1836-75-5	Nitrofen	+	57635-85-7	Hexachlorodibenzodioxin 2	+	136-77-6	4-Hexylresorcinol	=
156-10-5	*p*-Nitrosodiphenenylamine	+	67774-32-7	Polybrominated biphenyl	+	41372-08-1	Methyldopa sesquihydrate	=
101-80-4	4-4′-Oxydianiline	+	1746-01-6	2,3,7,8-Tetrachlorodibenzo-*p*-dioxin	+	2058-46-0	Oxytetracycline hydrochloride	=
136-40-3	Phenazopyridine HCl	+	86-06-2	2,4,6-Trichlorophenol	+	83-79-4	Rotenone	=
139-61-1	4,4′-Thiodianiline	+	115-28-6	Chlorendic acid	+	147-24-0	Diphenhydramine HCl	=
636-21-5	*o*-Toluidine HCl	+	106-46-7	1,4-Dichlorobenzene	+	968-81-0	Acetohexamide	–
137-17-7	2,4,5-Trimethylaniline	+	127-18-4	Tetrachloroethylene	+	50-81-7	L-Ascorbic acid	–
67-20-9	Nitrofurantoin	+	67-72-1	Hexachloroethane	+	128-37-0	Butylated hydroxytoluene	–
59-87-0	Nitrofurazone	+	87-86-5	Pentachlorophenol	+	262-12-4	Dibenzo-*p*-dioxin	–
26471-62-5	2,4-/2,6-Toluene diisocyanate	+	79-00-5	1,1,2-Trichloroethane	+	150-38-9	EDTA (tri-Na salt)	–
20265-96-7	*p*-Chloroaniline HCl	+	150-68-5	Monuron	+	9002-18-0	Agar	–
20325-40-0	3,3′-Dimethoxybenzidine 2HCl	+	12789-03-6	Chlordane	+	119-53-9	Benzoin	–
612-82-8	3,3′-Dimethylbenzidine 2HCl	+	510-15-6	Chlorobenzilate	+	105-60-2	Caprolactam	–
142-04-1	Aniline HCl	+	1897-45-6	Chlorothalonil	+	134-72-5	Ephedrine sulfate	–
103-33-3	Azobenzene	+	1163-19-5	Decabromodiphenyl oxide	+	15356-70-4	*d*-Menthol	–
95-79-4	5-Chloro-*o*-toluidine	+	72-55-9	Dichlorodiphenyldichloroethylene	+	108-95-2	Phenol	–
5160-02-1	D and C Red 9	+	76-44-8	Heptachlor	+	85-44-9	Phthalic anhydride	–
91-93-0	3,3′-Dimethoxybenzidine-4-4′-diisocyanate	+	76-01-7	Pentachloroethane	+	1156-19-0	Tolazamide	–
			630-20-6	1,1,1,2-Tetrachloroethane	+	76-87-9	Triphenyltin hydroxide	–
121-14-2	2,4-Dinotrotoluene	+	79-34-5	1,1,2,2-Tetrachloroethane	+	434-13-9	Lithocholic acid	–
99-55-8	5-Nitro-*o*-toluidine	+	79-01-6	Trichloroethylene	+	69-65-8	D-Mannitol	–
80-08-0	4,4′-Sulfonyldianiline	+	309-00-2	Aldrin	+	114-86-3	Phenformin	–
1582-09-8	Trifuralin	+	63449-39-8	Chlorinated paraffins (C23:43% Cl)	+	88-96-0	Phthalamide	–
3165-93-3	4-Chloro-*o*-toluidine HCl	+	115-32-2	Dicofol	+	51-03-6	Piperonyl butoxide	–
2475-45-8	CI Disperse Blue 1	+	54-31-9	Furosemide	+	64-77-7	Tolbutamide	–
102-50-1	*m*-Cresidne	+	108-90-7	Chlorobenzene	=	73-22-3	L-Tryptophan	–
609-20-1	2,6-Dichloro-*p*-phenylenediamine	+	33857-26-0	2,7-Dichlorodibenzo-*p*-dioxin	=	100-51-6	Benzyl alchohol	–
94-52-0	5(6)-Nitrobenzimadazole	+	60-57-1	Dieldrin	=	132-98-9	Penicilin VK	–
842-07-9	C.I. Solvent Yellow 14	+	72-56-0	Di(*p*-ethylphenyl)dichloroethane	=	64-75-5	Tetracycline hydrochloride	–
17026-81-2	3-Amino-4-ethoxyacetanilide	+	1918-02-1	Picloram	=	108-30-5	Succinic anhydride	–
119-34-6	4-Amino-2-nitrophenol	+	72-54-8	Tetrachlorodiphenylethane	=	643-22-1	Erithromycin stearate	–
121-66-4	2-Amino-5-nitrothiazole	+	58-93-5	Hydrochlorothiazide	=	61-76-7	Phenylephrine hydrochloride	–
105-11-3	*p*-Benzoquinone dioxime	+	101-05-3	Anilazine	–	1330-20-7	Xylenes commercial mixture	–
2185-92-4	2-Biphenylamine HCl	+	999-81-5	2-Chloroethyltrimethylammonium chloride	–	55-31-2	L-Epinephrine hydrochloride	–
133-90-4	Chloramben	+				108-88-3	Toluene	–
1777-84-0	3-Nitro-*p*-acetophenetide	+	95-50-1	1,2-Dichlorobenzene	–	2835-39-4	Allyl isovalerate	+
5307-14-2	2-Nitro-*p*-phenylenediamine	+	72-20-8	Endrin	–	87-29-6	Cinnamyl anthranilate	+
99-57-0	2-Amino-4-nitrophenol	+	72-43-5	Methyoxychlor	–	123-91-1	1,4-Dioxane	+
121-88-0	2-Amino-5-nitrophenol	+	77-65-6	Carbromal	–	271-89-6	Benzofuran	+
6373-74-6	C.I. Acid Orange 3	+	94-20-2	Chlorpropamide	–	98-01-1	Furfural	+
20265-97-8	*p*-Anisidine HCl	=	50-29-3	Dichlorodiphenyltrichloroethane	–	50-33-9	Phenylbutazone	+
106-47-8	*p*-Chloroaniline	=	58-89-9	Lindane	–	105-55-5	*N,N′*-Diethylthiourea	+
56-38-2	Parathion	=	82-68-8	Pentachloronitrobenzene	–	86-30-6	*N*-Nitrosodiphenylamine	+
952-23-8	Proflavin HCl	=	13366-73-9	Photodieldrin	–	100-52-7	Benzaldehyde	+
2871-01-4	HC Red 3	=	75-35-4	Vinylidene chloride	–	128-66-5	C.I. Vat Yellow 4	+
135-88-6	*N*-Phenyl-2-naphthylamine	=	2698-41-1	*o*-Chlorobenzalmelanotrile	–	78-59-1	Isophorone	+
121-19-7	Roxarsone	=	2438-88-2	2,3,5,6-Tetrachloro-4-nitroanisole	–	108-78-1	Melamine	+
989-38-8	Rhodamine 6G HCl	=	113-92-8	Chlorpheniramine maleate	–	2489-77-2	Trimethylthiourea	+

Table 1. *(Continued).*

CAS no.	Name	Act.	CAS no.	Name	Act.
137-30-4	Ziram	+	78-34-2	Dioxathion	–
5989-27-5	α-Limonene	+	121-75-5	Malathion	–
131-17-9	Diallyl phthalate	=	75-09-2	Dichloromethane	+
142-46-1	2,5-Dithiobiurea	=	75-27-4	Bromodichloromethane	+
20941-65-5	Ethyl tellurac	=	75-25-2	Tribromomethane (bromoform)	+
97-53-0	Eugenol	=	124-48-1	Chlorodibromomethane	+
2164-17-1	Fluometuron	=	75-47-8	Iodoform	–
116-06-3	Aldicarb	–	101-61-1	4,4′-Methylenebis	+
3567-69-9	C.I. Acid Red 14	–		(*N,N*′-dimethylbenzenamine)	
118-92-3	*o*-Anthranilic acid	–	90-94-8	Michler's ketone	+
1212-29-9	*N,N*′-Dicyclohexylthiourea	–	121-69-7	*N,N*-Dimethylaniline	+
536-33-4	Ethionamide	–	140-56-7	Fenaminosulf	–
19010-66-3	Lead dimethyldithiocarbamate	–	509-14-8	Tetranitromethane	+
89-25-8	1-Phenyl-3-methyl-5-pyrazolone	–	504-88-1	3-Nitropropionic acid	=
148-18-5	Sodium diethyldithiocarbamate	–	140-88-5	Ethyl acrylate	+
97-77-8	Tetraethylthiuram disulfate	–	924-42-5	*N*-Methylolacrylamide	+
———	Vinyl toluenes (*meta/para* 70:30)	–	80-62-6	Methyl methacrylate	–
2783-94-0	FD & C Yellow 6	–	24382-04-5	Malonaldehyde sodium salt	+
315-18-4	Mexacarbate	–	828-00-2	Dimethoxane	=
105-85-5	1-Phenyl-2-thiourea	–	95-06-7	Sulfallate	+
77-79-2	3-Sulfolene	–	513-37-1	Dimethylvinyl chloride	+
105-87-3	Geranyl acetate	–	133-06-2	Captan	+
6959-48-4	3-Chloromethylpyridine HCl	+	598-55-0	Methyl carbamate	+
96-12-8	1,2-Dibromo-3-chloropropane	+	1596-84-5	Succinic acid 2,2-dimethylhydrazide	+
106-93-4	1,2-Dibromoethane	+	95-14-7	1,2,3-Benzotriazole	=
107-06-2	1,2-Dichloroethane	+	148-24-3	8-Hydroxyquinoline	–
542-75-6	1,3-Dichloropropene	+	115-07-1	Propylene	–
3546-10-9	Phenestrin	+	60-13-9	Amphetamine sulfate	–
75-56-9	1,2-Propylene oxide	+	91-20-3	Naphthalene	+
961-11-5	Tetrachlorovinphos	+	9005-65-6	Polysorbate 80 (Tween 80)	=
512-56-1	Trimethylphosphate	+	58-33-3	Promethazine hydrochloride	–
126-72-7	Tris(2,3-dibromopropyl)phosphate	+	108-46-3	Resorcinol	–
563-47-3	3-Chloro-2-methylpropene	+	96-48-0	γ-Butyrolactone	=
62-73-7	Dichlorovos	+	79-11-8	Monochloroacetic acid	–
101-90-6	Diglycidyl resorcinol ether	+	100-02-7	*p*-Nitrophenol	–
74-96-4	Bromoethane	+	1330-78-5	Tricresyl phosphate	+
556-52-5	Glycidol	+	120-32-1	*o*-Benzyl-*p*-chlorophenol	+
5634-39-9	Iodinated glycerol	+	3296-90-0	2,2-Bis(bromomethyl)-1,3-propanediol	+
106-87-6	4-Vinyl-1-cyclohexene diepoxide	+	75-65-0	*t*-Butyl alchol	+
108-60-1	Bis(2-chloro-1-methyethyl)ether	+	119-84-6	3,4-Dihydroc属coumarin	+
868-85-9	Dimethyl hydrogen phosphite	+	107-21-1	Ethylene glycol	–
106-88-7	1,2-Epoxybutane	+	298-59-9	Methylphenidate hydrochloride	+
22966-79-6	Estradiol mustard	+	96-69-5	4,4′-Thiobis(6-*t*-butyl-*m*-cresol)	–
597-25-1	Dimethyl morpholinophosphoramidate	+	396-01-0	Triamterene	+
1955-45-9	Pivalolactone	+	57-41-0	Diphenylhydantoin	+
8001-35-2	Toxaphene	+	1825-21-4	Pentachloroanisole	+
78-87-5	1,2-Dichloropropane	+	10599-90-3	Chloramine	=
115-96-8	Tris(2-chloroethyl)phosphate	+	81-11-8	4,4′-Diamino-	–
57-06-7	Allyl isothiocyanate	+		2,2′-stilbenedisulfonic acid	
756-79-6	Dimethyl methylphosphonate	+	74-83-9	Methyl bromide	–
106-92-3	Allyl glcidyl ether	+	62-23-7	*p*-Nitrobenzoic acid	+
75-00-3	Chloroethane	+	28407-37-6	C.I. Direct Blue 218	+
86-50-0	Azinphosmethyl	=	2425-85-6	C.I. Pigment Red 3	+
55-38-9	Fenthion	=	6471-49-4	C.I. Pigment Red 23	=
13171-21-6	Phosphamidon	=	137-09-7	2,4-Diaminophenol dihydrochloride	+
532-27-4	2-Chloroacetophenone	=	103-90-2	4-Hydroxyacetanilide	=
78-11-5	Pentaerythritol tetranitrate	=	1271-19-8	Salicylazosulfapyridine	=
109-69-3	*n*-Butyl chloride	–	6459-94-5	C.I. Acid Red 114	+
107-07-3	2-Chloroethanol	–	2429-74-5	C.I. Direct Blue 15	+
56-72-4	Coumaphos	–	91-64-5	Coumarin	+
60-51-5	Dimethoate	–	96-13-9	2,3-Dibromo-1-propanol	+
1634-78-2	Malaoxon	–	119-93-7	3,3′-Dimethylbenzidine	+
124-64-1	Terakis (hydroxymethyl)	–	52551-67-4	HC Yellow 4	=
	phosphonium chloride/sulfate		100-01-6	*p*-Nitroaniline	=
6959-47-3	2-Chloromethylpyridine HCl	–	91-23-6	*o*-Nitroanisole	+
333-41-5	Diazinon	–	96-18-4	1,2,3-Trichloropropane	+

Abbreviations: +, carcinogenic for any species and in any organ; =, equivocal classification; –, noncarcinogenic.

approximately 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 in the form of either facts or computer programs. This allows the addition, in a unified way, of any information that is considered relevant to learning the SAR. In general, the more that is known about a problem, the easier it is to solve. The ability to use a variety of background knowldge is perhaps the most powerful feature of *Progol*. In this study we included the background knowledge of chemical groups from our work on predicting mutagenesis (*21*), and the structural alerts identified by Ashby et al. (*4*) were also encoded and tested. It is important to appreciate that encoding PROLOG programs to define these concepts is not the same as including them as simple indicator variables. This is because *Progol* can learn SARs that use structural combinations of these groups, e.g., *Progol* could in theory learn that a structural indicator of activity is diphenylmethane (as a benzene single-bonded to a carbon atom single-bonded to another benzene). In contrast, a normal SAR method would only be able to use the absence or presence of the different groups, not a bonded combination of them. To represent compounds to the equivalent level of detail using a CASE-type representation (*6*) would require several orders of magnitude more descriptors than needed for only the simple atom/bond representation (*21*). In the future the background knowledge used could be extended to include more information, e.g., 3D structure, knowledge about metabolism, subchronic *in vivo* toxicity, route of administration, minimally toxic dose (MTD) levels, etc.

Other SAR Algorithms Compared with *Progol*

The train/test dataset has previously been studied using a number of SAR methods. We use the predictions from these methods and the predictions from two default methods to compare their results with those of *Progol*. The two default methods that we implemented were the following:

- The largest class prediction method is to predict all compounds to be carcinogenic (the largest class).

- The Ames prediction method is to predict a compound to be carcinogenic if it has any form of a positive Ames test.

The previously applied prediction methods that were compared with *Progol* can be placed into two groups. In the first group are the prediction methods that do not directly use data from experiments on rodents. The *Progol* SAR method belongs to this group and can be directly compared with such methods. These methods are as follows:

- The Bakale and McCreary method (*9*) used experimentally measured electrophilic reactivity (K_e) values to discriminate between carcinogenic and noncarcinogenic compounds.
- The DEREK method (deductive estimation of risk from existing knowledge) (*5*) is an expert-system that predicts carcinogenesis based on a set of rules derived from experienced chemists.
- The COMPACT method (computer-optimized molecular parametric analysis of chemical toxicity) (*10*) predicts carcinogenesis based on the predicted interaction of the compound with cytochrome P450 and the Ah receptor.
- The CASE method (*25*) is based on a statistical method of selecting chemical substructures associated with carcinogenesis.
- The TOPKAT system (toxicity prediction by komputer [*sic*] assisted technology) (*11*) uses structural attributes to describe the compounds and applies statistical discrimination and regression to estimate the probability of carcinogenesis; it uses a number of noncarcinogenic pharmaceuticals and food additives to increase the number of negative examples.
- The Benigni method (*12*) forms a Hansch-type quantitative structure–activity relationship (QSAR) using estimated electrophilic reactivity (K_e) and Ashby's structural alerts (below).

The second group of prediction methods that have been previously applied uses information from biological tests on rodents. It is unfair to directly compare these procedures with methods based only on chemical structure and *Salmonella* mutagenicity since they use more information. Rodent biological tests are very expensive both in money and animal welfare terms. The prediction methods that use rodent biological tests are as follows:

- The Ashby prediction method (*3*) is based on the expert judgment of chemists to evaluate evidence from a set of chemical structural indicators *Salmonella* mutagenicity, subchronic *in vivo* toxicity, the route of administration of the compound, and MTD levels (*4*). When experimental carcinogenicity results from previous studies were available in the literature, this evidence was also taken into consideration (*15*).
- The TIPT method (tree induction for predictive toxicology) (*8*) uses the machine learning algorithm C4.5 (a propositional tree-learning method) to combine the same evidence used by the Ashby prediction method. It cannot identify new structural alerts.
- The RASH method (rapid screening of hazards) (*13*) uses relative potency analysis and dose levels to modulate the Ashby method.

Results

Train/Test Results

The predictions of *Progol* on round 1 of the NTP carcinogenicity trial are given in Tables 2 and 3. The data consisted of 291 training compounds and 39 test compounds (*8*). The small size of the test dataset makes it difficult to show statistically significant differences between algorithms; this difficulty is compounded because some algorithms cannot predict all the examples. Comparing the predictions, using a binomial McNemar test for changes (*26*), shows that no algorithm is significantly more accurate than *Progol* ($p < 0.05$). The McNemar test exploits the fact that the different prediction methods are applied to the same data and are based on counting the examples for which the methods disagree about predictions.

Progol is marginally the most accurate prediction method that does not use rodent tests (although this is not statistically significant). The more accurate prediction methods of Ashby, TIPT, and RASH are based on use of short-term rodent *in vivo* tests. This information is much more difficult and expensive to obtain than chemical structural and *Salmonella* mutagenicity data. The Ashby and RASH methods are also based on the subjective application of a set of structural alerts formed by Ashby et al. (*4*); the TIPT method uses an objective application of these expert defined alerts.

A number of the errors in prediction made by *Progol* were repeated by most other methods, suggesting some anomaly with these compounds (*14*)—methylphenidate hydrochloride and methyl bromide. *Progol* correctly identified naphthalene as a carcinogen, while it was missed by all other methods.

Cross-validation Results

Progol has an accuracy of 63% (standard error ± 3%) for all compounds estimated by 5-fold cross validation. This compares with estimated accuracies of 55% using the default rule, and 63% using the Ames rules. There is a significant difference at $p < 0.05$ between the accuracy of *Progol* and the default rule. Although there is no significant difference in accuracy between *Progol* and the Ames rule, there is a large difference in the number of carcinogens identified. *Progol* makes fewer errors of omission than

Table 2. Accuracy of different prediction methods on 39 compounds previously tested by the NTP.

Information	Prediction method	Accuracy, %	Cover (PP, PN, NP, NN)
Default	Ames test	59	39 (10, 5, 11, 13)
	Largest class	54	39 (21, 18, 0, 0)
No rodent tests	PROGOL (*19*)	64	39 (18, 11, 3, 7)
	Bakale and McCreary (*9*)	63	30 (11, 5, 6, 8)
	Benigni (*12*)	62	37 (16, 9, 5, 7)
	DEREK (*5*)	57	37 (12, 8, 8, 9)
	COMPACT (*10*)	54	37 (14, 10, 7, 6)
	TOPKAT (*11*)	54	26 (6, 3, 9, 8)
	CASE (*25*)	49	37 (11, 9, 10, 7)
Rodent tests	Ashby (*3*)	77	39 (19, 7, 2, 11)
	RASH (*13*)	72	29 (8, 0, 8, 13)
	TIPT (*8*)	67	39 (19, 11, 2, 7)

Terms and abbreviations: Accuracy = number correctly predicted/total number predicted. Cover = number of compounds predicted (PP = predicted to be carcinogenic and is carcinogenic, PN = predicted to be carcinogenic and is not carcinogenic; NP = predicted to be not carcinogenic and is carcinogenic, NN = predicted to be not carcinogenic and is not carcinogenic. Default methods are those that use simplistic prediction strategies. The basic methods are those that use information solely from chemical structure and *Salmonella* mutagenicity tests. The complex methods use information from rodent biological tests; the Ashby and RASH methods also exploit expert chemical knowledge and are therefore not automatic.

the Ames rules and more errors of commission, i.e., *Progol* identifies more carcinogens than the Ames rule at the cost of classifying more noncarcinogens as carcinogens.

Rules

The *Progol* SAR method produces prediction rules in the form of easily understood chemical patterns. The prediction rules are given in Figures 2 and 3. There is a direct translation from the rules generated by *Progol* into chemical structure. For example, rule 3 in PROLOG notation is:

```
active(Drug) if
atom(Drug, Atom_1, Element_1,
  ester_carbon, Charge1) and
atom(Drug, Atom_2, Element_2,
  aromatic_hydrogen, Charge2) and
less_than_or_equal(Charge2, 0.041)
  (names with capital letters are
  variables).
```

The particular use of partial charges requires some explanation. They are given to three significant places because of a peculiarity in the method of assigning partial charges in QUANTA^{TM} (above), not because it is considered that these exact values are important to this accuracy.

It is important that rules produced by any automatic SAR procedure are screened to ensure that they make chemical sense. More confidence can be put in a rule if a mechanism of action can be identified (*27–29*). This is a general application of the principle of using prior knowledge to guide decision making. All the rules found by *Progol* were analyzed to try to identify their chemical rational.

It was found that use of the Ames test for *Salmonella* mutagenicity (rule 1) was the most effective rule for predicting carcinogenicity. While learning rule 1, *Progol* automatically searched for structural features that improved rule 1 and no such rule was found that had higher compression than rule 1 (recall that compression is an objective way of balancing sensitivity/specificity of a rule). This does not conflict with the results of Ashby and Tennant (*4*), who showed that the Ames test was correlated with a set of structural alerts.

The remaining rules found by *Progol* are new and they automatically generated structural alerts for carcinogenesis. As *Progol* removes examples covered by previous rules when searching for a new rule, rules found after rule 1 was covered are indicators for carcinogenic compounds not recognized by the Ames test. This means

Table 3. *Progol* predictions for the test set.

CAS No.	Name	Actual	Prediction
6459-94	C.I. Acid Red 114	+	+
96-18-4	1,2,3-Trichloropropane	+	+
96-13-9	2,3-Dibromo-1-propanol	+	+
119-93-7	3,3′-Dimethylbenzidine 2HCl	+	+
1825-21-4	Pentachloranisole	+	+
2425-85-6	CI Pigment Red 3	+	+
91-23-6	*o*-Nitroanisole	+	+
28407-37-6	C.I. Direct Blue 218	+	+
91-64-5	Coumarin	+	+
2429-74-5	C.I. Direct Blue 15	+	+
137-09-7	2,4-Diaminophenol 2HCl	+	+
115-96-8	Tris(2-chloroethyl)phosphate	+	+
396-01-0	Triamterene	+	+
120-32-1	*o*-Benzyl-*p*-chlorophenol	+	+
57-41-0	Diphenylhydantoin	+	+
91-20-3	Naphthalene	+	+
62-23-7	*p*-Nitrobenzoic acid	+	+
3296-90-0	2,2-Bis(bromomethyl)-1,3-propanediol	+(draft)	+
119-84-6	3,4-Dihydrocoumarin	+	–
1330-78-5	Tricresyl phosphate	+	–
298-59-9	Methylphenidate HCl	+	–
75-65-0	*t*-Butyl alcohol	+	–
10599-90-3	Chloramine	=	–
96-48-0	γ-Butyrolacetone	=	–
9005-56-6	Polysorbate 80	=	–
103-90-2	4-Hydroxyacetanalide	=	+
127-19-8	Titanocene dichloride	=	+
52551-67-4	HC Yellow 4	=	+
100-01-6	*p*-Nitroaniline	=	+
6471-49-4	C.I. Pigment Red 23	=	+
60-13-9	Amphetamine sulfate	–	–
107-21-1	Ethylene glycol	–	–
108-46	Resorcinol	–	–
100-02-7	*p*-Nitrophenol	–	–
79-11-8	Monochloroacetic acid	–	+
81-11-8	4,4′-Diamino-2,2′-stilbenedisulfonic acid	–	+
74-83-9	Methyl bromide	–	+
58-33-3	Promethasine HCl	–	+
96-69-5	4,4-Thiobis(6-*t*-butyl-*m*-cresol)	–	+

Actual = the result of the NTP rodent bioassays (+, carcinogen for any species and in any organ; =, an equivocal classification treated as a noncarcinogen). Pred. = the PROGOL predictions (+, predicted to be a carcinogen; –, a noncarcinogen).

A compound is carcinogenic if it has a(n)
(1) positive Ames test [99-40] or
(2) ester oxygen and ≤ 1 methyl group [8-2] or
(3) ester oxygen and an aromatic hydrogen with a partial charge ≤ 0.041 [9-2] or
(4) ether oxygen with a partial charge ≥ –0.182 [5-1] or
(5) ether oxygen with a partial charge of –0.358 [2-0] or
(6) chlorine, an hydroxyl group and a benzyl ring. [9-1] or
(7) bromine with a partial charge ≤ –0.086 [5-0] or
(8) aldehyde oxygen [3-0] or
(9) aromatic amine group and an unsaturated carbon
with a partial charge ≤ –0.181 [5-0] or
(10) unsaturated carbon with a partial charge of ≥ 0.4 [4-0] or
(11) unsaturated carbon and a 6-membered carbon ring [6-1] or
(12) carbon atom in a 6-membered aromatic ring with a partial charge of 0.005 [3-0] or
(13) carbon atom in a 6-membered aromatic ring with a partial charge of 0.211 [4-1] or
(14) carbon atom in a 6-membered aromatic ring with a partial charge of –0.135 [3-1] or
(15) aliphatic carbon with a partial charge ≥ 0.507 [7-1] or
(16) aliphatic carbon with a partial charge of –0.085. [2-0] or
(17) four halide atoms attached to tetrahedral carbons [4-1] or
(18) carbon in a 5-membered aromatic ring with the
same charge as a carbon in a 6-membered aromatic ring [3-0]

Figure 2. The rules (alerts) found by *Progol* for carcinogenesis using all the compounds. The numbers in brackets are the number of times the rule correctly occurs and the number of times it incorrectly occurs.

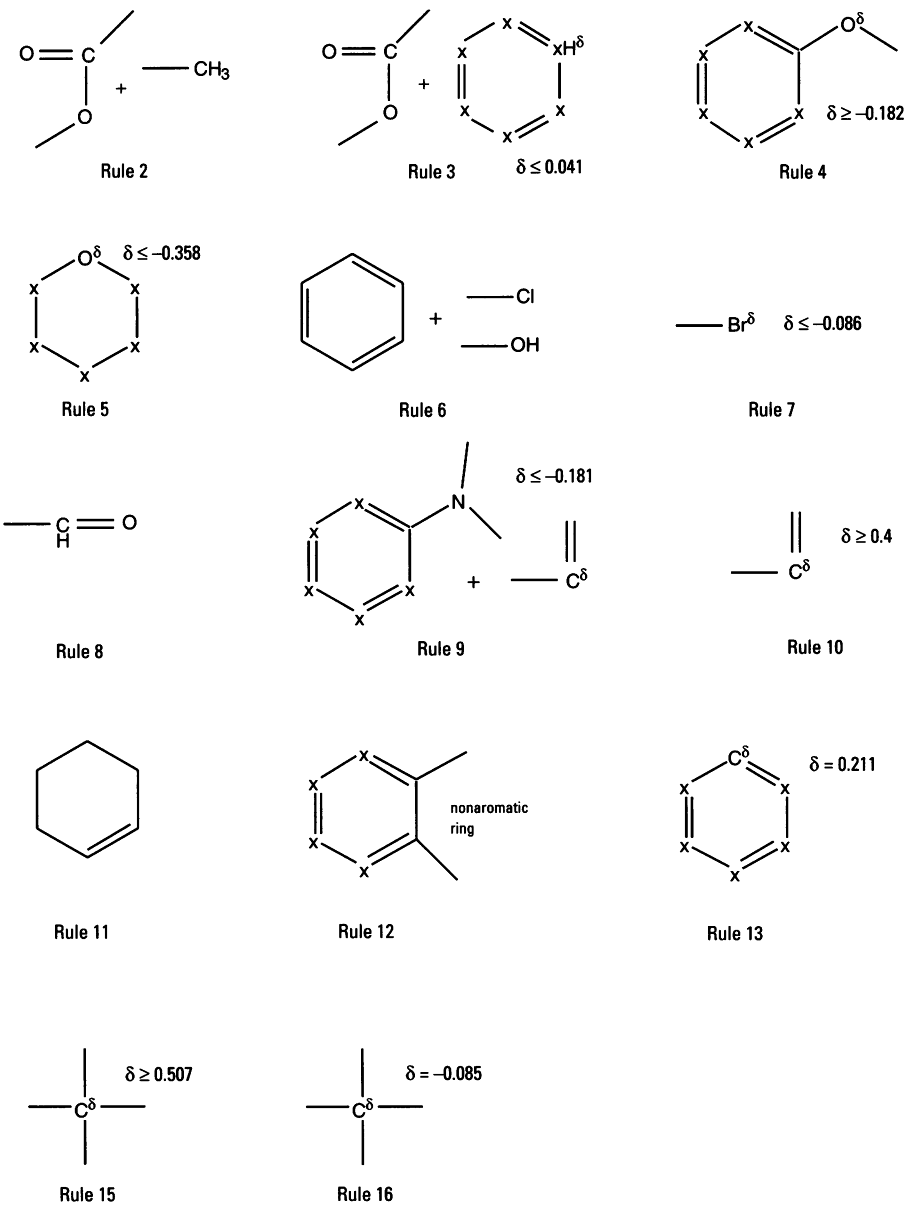

Figure 3. Graphical interpretation of the alerts described in Figure 2.

that they could be either structural alerts for nongenotoxic carcinogenesis (*4*), (i.e., not based on induction of DNA damage by the test agent or its metabolites), or structural alerts for genotoxic carcinogens that are missed by the Ames test. Most of the structural features identified by *Progol* appear to be for highly reactive structures, suggesting that they mainly act by genotoxic carcinogenesis. Chemical interpretations of the rules are given below (arranged by chemical group):

- Rules 2 and 3 identify ester groups as indicators for carcinogenesis. The meaning of the modifying groups is unclear, but they are essential, as ester groups on their own have no discriminatory power. Rules 2, 6, and 11 use the generic background knowledge that was first used in applying *Progol* to predicting mutagenesis (*21*).
- Rule 4 is concerned with ether oxygens with high partial charges. All such groups are bonded to aromatic rings, suggesting the involvement of electrophilic substitution in activity.
- Rule 5 identifies an ether group in a 6-membered ring. These cyclic ethers may also be involved in electrophilic reactions.
- Rules 6 and 7 identify reactive halides as indicators of carcinogenesis; such compounds have been widely recognized as potential carcinogens.
- Rule 8 identifies an aldehyde group as an indicator of carcinogenicity. Aldehyde groups are potentially very reactive.
- In rule 9, the aromatic amine group indicates high reactivity, as does the low partial charge on the unsaturated carbon (it is associated with a double bond to an oxygen group).
- The high partial charge on the unsaturated carbon in rule 10 occurs in reactive alkenes.
- Rule 11 occurs in substituted cyclohexenes; note the similarity with rule 5.
- Rule 12 occurs when a 6-membered aromatic ring is bonded to a nonaromatic ring.
- Rule 13 occurs when a carbon atom in a single 6-membered aromatic ring is bonded to an amine or carbon-substituted amine group.
- Rule 15 occurs in chlorinated alkane groups; see rule 6.
- Rule 16 occurs when a hydroxyl group is attached to an aliphatic carbon.
- In rule 17 the indicator of a halide atom is attached a tetrahedral carbon. This is the only rule that uses the structural alerts from Ashby et al. (*4*), It is possible that this rule is an artifact, since there appears to be no chemical reason why 4 halide atoms should be chosen instead of, say, 3 or 5.
- Rules 14 and 18 may also be artifacts because there appears to be no chemical rationale for them.

Discussion

Prediction of Results of Ongoing NTP Studies

The *Progol* rules were used to predict the compounds in the second round of the NTP test of strategies for predicting chemical carcinogenesis in rodents; the 25 organic compounds were predicted but no prediction was made for the 5 inorganic compounds (Table 4). The predictions made by *Progol* are only tentative for the compounds with negative Ames tests because of the limited number of examples covered by the rules (low compression and low statistical reliability). The predictions based on rule 8 are probably the least reliable because there are no other supporting rules and the rule is almost certainly an over-generalization. The extent of the agreement between the predicted results

Table 4. *Progol* predictions for the second round of the NTP trial of strategies for predicting chemical carcinogenesis in rodents.

CAS no.	Name	Prediction	Rule*
6533-68-2	Scopolamine hydrobroamide	+	2,5
76-57-3	Codeine	–	
147-47-7	1,2-Dihydro-2,2,4-trimethyequinoline	+	9,14
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	+	1
78-84-2	Isobutyraldehyde	+	8
127-00-4	1-Chloro-2-propanol	–	
11-42-2	Diethanolamine	–	
77-09-8	Phenolphthalein	–	
110-86-1	Pyridine	–	
1300-72-7	Xylenesulfonic acid, Na	–	
98-00-0	Furfuryl alcohol	–	
125-33-7	Primaclone	+	1
111-76-2	Ethylene glycol monobutyl ether	–	
115-11-7	Isobutene	–	
93-15-2	Methyleugenol	–	
434-07-1	Oxymetholone	–	
84-65-1	Anthraquinone	+	1
518-82-1	Emodin	+	1
5392-40-5	Citral	+	8
104-55-2	Cinnamaldehyde	+	8
10026-24-1	Cobalt sulfate heptahydrate	Not predicted	
1313-27-5	Molybdenum trioxide	Not predicted	
1303-00-0	Gallium arsenide	Not predicted	
7632-00-0	Sodium nitrite	Not predicted	
1314-62-1	Vanadium pentozide	Not predicted	

+, positive; –, negative. *Rule used in the prediction.

and those experimentally predicted will indicate how relevant the assumptions underlying the *Progol* predictions are.

A major statistical problem in trying to predict the results of this NTP trial is that the distribution of compounds in the trial is not the same as that from which the rules were learned, e.g., the percentage of compounds with positive Ames tests is only 16% (4 of 25) compared with 42% for the compounds previously tested. The change in distribution between training and test data for NTP trials has previously been noted by Tennant et al. (*3*). This is called concept drift in machine learning and it is a problem because almost all statistical methods are based on the assumption of an underlying constant distribution.

Comparison with Other SAR results

The estimated accuracy of 63% for predicting carcinogenesis by *Progol* is higher but not statistically significantly higher than the results obtained using other SAR methods that do not incorporate results from rodent biological tests. This confirms the results of Benigni (*15*), who showed that all the SAR approaches to carcinogenicity had similar prediction profiles. The relatively low prediction accuracy of ≈60% is probably due to the diversity of mechanisms of action and the complexity of interactions *in vivo*.

Comparison of the Ashby et al. Structural Alerts and Those Generated by *Progol*

The Ashby et al. structural alerts (*2–4,28*) and those generated by *Progol* differ fundamentally in their formation and application. The Ashby alerts were generated by a human expert and applied subjectively. The *Progol* alerts were generated automatically by machine and are applied objectively. The Ashby structural alerts are based on electrophilic attack on DNA. This means that they are not statistically independent of the Ames test (*4*), and there is some redundancy between the Ames test and the structural alerts. The *Progol* structural alerts were selected so that they covered compounds not covered by the Ames test. This makes them much more independent of each other than those of Ashby.

Many of the structural alerts found by *Progol* are similar to those identified by Ashby, e.g. Ashby recognizes forms of esters (rules 2 and 3), ethers (rules 4 and 5), halogenated compounds (rules 6, 7, and 15), and aldehydes (rule 9) as structural alerts. The exact forms of the alerts differ significantly between Ashby and *Progol*. This strongly suggests that it may be possible to develop a system for predicting chemical carcinogenesis that combines the best features of human-based prediction with the objectivity and speed of the *Progol* rules to develop a superior SAR system.

The results for prediction of carcinogenesis, taken together with the previous successful applications of predicting mutagenicity in nitroaromatic compounds and inhibition of angiogenesis by suramin analogues, show that *Progol* has a role to play in understanding the SARs of cancer-related compounds.

Program Availability

The ILP program *Progol* (implemented in PROLOG) and the data used in the current study can be obtained from Ashwin Srinivasan, Oxford Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK; e-mail at ashwin.srinivasan@comlab.oxford.ac.uk. They are freely available to academics. A version of *Progol* is also available that is implemented in C, available by anonymous ftp from ftp.comlab.ox.ac.uk in directory pub/Packages/ILP/progol4.1. Additional information about *Progol* and ILP can be found at the World Wide Web address http://www.comlab.ox.ac.uk/oucl/groups/machlearn

Appendix: Illustration of Use of the *Progol* Algorithm

To outline the *Progol* algorithm, we illustrate its use on a simple SAR problem. The example is not completely trivial, as it could not be dealt with using simple linear regression/discrimination. The problem data consist of eight compounds. The positive examples (active compounds) are high(drug3), high(drug6), high(drug7), high(drug8). The negative examples (inactive compounds) are high(drug1), high(drug2), high(drug4), high(drug5). Each drug can have a substituent at three positions: subst1(drug1), subst1(drug2), subst1(drug3), subst1 (drug8), subst2(drug1), subst2(drug2), subst2(drug6), subst2(drug7), subst3(drug1), subst3(drug3), subst3(drug4), subst3(drug7).

The *Progol* algorithm starts by randomly selecting a positive example. The order of example selection does not affect the final theory, only the efficiency of the learning process, e.g., drug number 3: high(drug3). *Progol* generalizes the example using inverse resolution to construct the most specific rule that explains the example in terms of the background knowledge. This rule logically implies the original example. The rule is

high(X):

subst1(X),
not subst2(X),
subst3(X).

In plain language, this rule states that

A drug has high activity if
it has a substitution at position 1 and
it does not have a substitution at position 2 and
it has a substitution at position 3.

This rule covers 1 positive example and no negative examples. *Progol* further generalizes this rule by removal of redundant parts of its body (literals) to find the maximally compressive rule using a complete top-down search. The most compressive rule for our example is

high(X):

subst1(X),
not subst2(X).

This rule is the optimal in terms of compression and the descriptive language used. It covers two positive examples and no negative examples. This rule is added to the knowledge base and the examples covered by it removed. Steps 1 to 4 are then repeated until no more compression is possible.

The final theory produced is

A drug has high activity if
it has a substitution at position 1 or
it has a substitution at position 2.

This is called the *exclusive or* rule.

REFERENCES

1. Huff J, Haseman J. Long-term chemical carcinogenesis experiments for identifying potential human cancer hazards. Environ Health Perspect 96:23–31 (1991).
2. Ashby J, Tennant RW, Zeiger E, Stasiewicz S. Classification according to chemical structure, mutagenicity to *Salmonella* and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the US National Toxicology Program. Mutat Res 223:73–103 (1989).
3. Tennant RW, Spalding J, Stasiewicz S, Ashby J. Prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 44 chemicals by the National Toxicology Program. Mutagenesis 5:3–14 (1990).
4. Ashby A, Tennant RW. Definitive relationships among chemical carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. Mutat Res 257:229–306 (1991).
5. Sanderson DM, Earnshaw CG. Computer prediction of possible toxic action from chemical structure. Hum Exp Toxicol 10:261–273 (1991).
6. Klopman G. Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. J Am Chem Soc 106:7315–7321 (1984).
7. Klopman G. MULTICASE: 1. A hierarchical computer automated structure evaluation program. Quant Struct Act Relat 11: 176–184 (1992).
8. Bahler D, Bristol DW. The induction of rules for predicting chemical carcinogenesis in rodents. In: Intelligent Systems for Molecular Biology-93 (Hunter L, Searls D, Shavlick J, eds). Cambridge, MA:AAI/MIT Press, 1993; 29–37.
9. Bakale G, McCreary RD. Prospective K_e screening of potential carcinogens being tested in rodent bioassays by the US National Toxicology Program. Mutagenesis 7:91–94 (1992).
10. Lewis DFV, Ionnides C, Parke DV. A prospective toxicity evaluation (COMPACT) on 40 chemicals currently being tested by the National Toxicology Program. Mutagenesis 5:433–436 (1990).
11. Enslein K, Blake BW, Borgstedt HH. Prediction of probability of carcinogenicity for a set of ongoing NTP bioassays. Mutagenesis 5:305–306 (1990).
12. Benigni R. QSAR prediction of rodent carcinogenicity for a set of chemicals currently bioassayed by the US National Toxicology Program. Mutagenesis 6:423–425 (1991).
13. Jones TD, Easterly CE. 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).
14. Ashby J, Tennant RW. Prediction of rodent carcinogenicity for 44 chemicals: results. Mutagenesis 9:7–15 (1994).
15. Benigni R. Predicting chemical carcinogensis in rodents: the state of the art in the light of a comparative exercise. Mutat Res 334:103–113 (1995).
16. King RD, Muggleton S, Lewis RA, Sternberg MJE. 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 Natl Acad Sci USA 89:11322–11326 (1992).
17. Hirst JD, King RD, Sternberg MJE. Quantitative structure-activity relationships by neural networks and inductive logic programming. 1. The Inhibition of dihydrofolate reductase by pyrimidines. J Comput Aided Mol Des 8:405–420 (1994).
18. Hirst JD, King RD, Sternberg MJE. Quantitative structure-activity relationships by neural networks and inductive logic programming. II. The inhibition of dihydrofolate reductase by triazines. J Comput Aided Mol Des 8: 421–432 (1994).
19. Muggleton SH. Inverse entailment and Progol. New Gen Comput 13:245–286 (1995).
20. King RD, Srinivasan A, Sternberg MJE. Relating chemical activity to structure: an examination of ILP success. New Gen Comput 13:411–433 (1995).
21. King RD, Muggleton SH, Srinivasan A, Sternberg MJE. Structure-activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity using inductive logic programming. Proc Natl Acad Sci USA 93:438–442 (1996).
22. Lee Y, Buchanan BG, Mattison DM, Klopman G, Rosenkranz HS. Learning rules to predict rodent carcinogenicity of non-genotoxic chemicals. Mutat Res 328: 127–149 (1995).
23. DeLong H. A Profile of Mathematical Logic. Reading, MA:Addison-Wesley, 1970.
24. Wallace CS, Freeman PR. Estimation and inference by compact coding. J R Stat Soc B 49:195–209 (1987)

25. Rosenkranz HS, Klopman G. Prediction of the carcinogenicity in rodents of chemicals currently being tested by the US National Toxicology Program. Mutagenesis 5:425–432 (1990).
26. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrica 12: 153–157 (1947).
27. Ashby J. Two million rodent carcinogens? The role of SAR and QSAR in their detection. Mutat Res 305:3–12 (1994).
28. Ashby J, Paton D. The influence of chemical structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogen exposures. Mutat Res 286:3–74 (1993).
29. Richard AM. Application of SAR methods to noncongeneric data bases associated with carcinogenicity and mutagenicity: issues and approaches. Mutat Res 305:73–97 (1994).