Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 438–442, January 1996
Chemistry

Structure–activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming

ROSS D. KING*†, STEPHEN H. MUGGLETON‡, ASHWIN SRINIVASAN‡, AND MICHAEL J. E. STERNBERG*

*Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, Lincoln's Inn Fields, P.O. Box 123, London, WC2A 3PX, United Kingdom; and ‡Computing Laboratory, University of Oxford, Wolfson Building, Parks Road, Oxford, OX1 3QD, United Kingdom

Communicated by Walter Bodmer, Imperial Cancer Research Fund, London, U.K., September 19, 1995 (received for review May 11, 1995)

ABSTRACT **We present a general approach to forming structure–activity relationships (SARs). This approach is based on representing chemical structure by atoms and their bond connectivities in combination with the inductive logic programming (ILP) algorithm PROGOL. Existing SAR methods describe chemical structure by using *attributes* which are general properties of an object. It is not possible to map chemical structure directly to attribute-based descriptions, as such descriptions have no internal organization. A more natural and general way to describe chemical structure is to use a *relational* description, where the internal construction of the description maps that of the object described. Our atom and bond connectivities representation is a relational description. ILP algorithms can form SARs with relational descriptions. We have tested the relational approach by investigating the SARs of 230 aromatic and heteroaromatic nitro compounds. These compounds had been split previously into two subsets, 188 compounds that were amenable to regression and 42 that were not. For the 188 compounds, a SAR was found that was as accurate as the best statistical or neural network-generated SARs. The PROGOL SAR has the advantages that it did not need the use of any indicator variables handcrafted by an expert, and the generated rules were easily comprehensible. For the 42 compounds, PROGOL formed a SAR that was significantly ($P < 0.025$) more accurate than linear regression, quadratic regression, and back-propagation. This SAR is based on an automatically generated structural alert for mutagenicity.**

A structure–activity relationship (SAR) models the relationship between activities and physicochemical properties of a set of compounds and is fundamental to many aspects of chemistry. SAR modeling has been applied to a multitude of biological systems and has aided the development of many new drugs (see refs. 1 and 2). To guide rational drug design a SAR should be both reliable and comprehensible. This paper presents an approach to forming SARs based on the machine learning program PROGOL (3). This approach allows the use of a rich representation of chemical structure and leads to SARs that are both accurate and simple to understand.

There are two components to deriving a SAR: the choice of representation to describe the chemical structure of the compounds and the learning algorithm employed. The form of learning algorithm restricts the representation that can be employed. Widely used learning algorithms include linear regression (4), partial least-squares regression (PLS) (5), neural networks (6, 7), and decision trees (8). These algorithms have been applied to a variety of descriptions of chemical structure—e.g., Hansch-type parameters (4, 9), topological descriptors (2, 10), quantum mechanical descriptors (9), substructural units (11, 12), molecular shape (MS) (13), and molecular fields (CoMFA) (14).

A key feature of all the above representations is that they are based on *attributes*—i.e., general properties of objects. For example, in the traditional Hansch approach to SARs the attributes are properties such as LogP and π, which are global properties of the molecule or substituted group, whereas in the CoMFA approach to SARs, the attributes are points in space which are global properties of the coordinate system used. Each compound is described as a list (technically a tuple) of attributes. However, this form of data representation is not well suited to describing the steric structure of chemicals, as it is difficult to map efficiently atoms and their connectivities onto a list.

A more general way to describe objects is to use *relations.* In a relational description the basic elements are substructures and their associations. This increased generality allows a more direct mapping from chemical steric structure to its representation. Fully relational descriptions of chemical structure have not previously been used in SARs because existing learning algorithms cannot use them. Inductive logic programming (ILP) algorithms (15) are designed to learn (i.e., induce) from examples encoded as logical relations. For many learning problems, relational descriptions have been shown to produce more concise and accurate rules than those based on attributes (16). Formally, the difference in descriptive language between attributes and relations corresponds to the difference between propositional and first-order predicate logic (17). To illustrate this difference between attributes and relations consider the following hypothesis: *An active compound requires a double bond conjugated with an aromatic ring.* Such a hypothesis could be directly discovered and expressed by a relational SAR system using only simple atom and bond types (e.g., atom A in an aromatic ring is connected by a single bond to atom B, which is connected by a double bond to atom C). It could not be found or expressed in an attribute-based language without specifically precoding the attribute "double bond conjugated with an aromatic ring."

Recently we have developed the ILP algorithm PROGOL (3), whose features (see below) enable us to implement a general relational method for describing chemical structure in SARs. This method is based on using atoms and their bond connectivities and is simple, powerful, and generally applicable to any SAR. It is particularly well suited to forming SARs that are dependant on molecular shape (shape is the relationship between objects in space), and SARs that are easily understood, as chemists are used to relating chemical properties and functions for groups of atoms. The method also appears robust and suited to SAR problems that are difficult to model conventionally. We present a benchmark of this ILP approach on a system that has been studied by several existing algo-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: SAR, structure–activity relationship; ILP, inductive logic programming.

†To whom reprint requests should be addressed.

rithms, the mutagenicity of aromatic and heteroaromatic nitro compounds.

METHODS

Data. Debnath *et al.* (9) studied the SAR of a set of 230 aromatic and heteroaromatic nitro compounds tested for mutagenesis by the Ames test using *Salmonella typhimurium* TA98. The compounds are heterogeneous and cannot be superimposed onto a common template and thus present a challenge to existing SAR methods. Debnath *et al.* identified two subsets of the data, 188 compounds considered to be amenable to regression, and 42 compounds that could not easily be fitted by regression. We have used this split of the data in the present study. This split of the data was also used by Villemin *et al.* (18), who examined only the 188 dataset using neural networks. These previous studies used an attribute-based representation which consisted of two real-valued attributes and two binary-valued indicator variables. The real-valued attributes were the energy of the lowest unoccupied molecular orbital (LUMO) and the molecular hydrophobicity (octanol/water partition coefficient, LogP. [LUMO energies are used in preference to the physically more meaningful highest occupied molecular orbital (HOMO) energies to allow direct comparison with previous work.] Explicit information about the molecular shape of the compounds is not captured by LUMO or LogP. Therefore Debnath *et al.* (9) augmented the description of the compounds by creation of two binary-valued indicator variables: I_1, set to 1 for all compounds containing three or more fused rings, and I_a, set to 1 for the five examples of acenthrylenes (as they had lower than expected activity). Note that this representation was not formed automatically but was selected by experienced chemists after detailed inspection to fit this particular dataset.

We split the dataset of 188 compounds into 10 cross-validation sets for testing. The dataset of 42 compounds was tested by leave-one-out cross-validation. Cross-validation was used as an empirical check of the significance tests used by the different learning methods. The compounds were split into two classes, high mutagenic [log (no. of revertants/nmol) > 0.0] and low mutagenic. There were 138 compounds considered in class high mutagenic and 92 compounds in class low mutagenic.

PROGOL. In ILP, logical relationships expressed as a subset of predicate calculus (17) are used to represent rules. Predicate calculus is expressive enough to describe most mathematical concepts and has a strong link with natural language. PROGOL implements predicate logic in the computer language PROLOG.

In any application, the input to PROGOL is a set of positive examples (i.e., for SAR, the active compounds), negative examples (i.e., inactive compounds), and background knowledge about the problem (e.g., the atom/bond structure of the compounds). PROGOL outputs a hypothesis, expressed as a set of rules which explain the positive and negative examples in terms of the background knowledge. The rule found for each example is optimal in terms of simplicity (information compression, see below) and the language used to describe the examples. However, the final output hypothesis is not necessarily optimal, because a simpler hypothesis may exist that is suboptimal for each individual example. 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 (19). The use of compression balances accuracy (no. of correct predictions/no. of total predictions) and coverage (no. of examples predicted by the rule/no. of examples). Full details of PROGOL are available in ref. 3.

Compound Representation for PROGOL. In our PROGOL study we used a generic relational representation based on atoms and their bond connectivities. The atomic structure of each compound was input into the molecular modeling program QUANTA (Molecular Simulations, Burlington, MA) by using its chemical editing facility. QUANTA was then used to automatically transform the representation by typing the atoms and by adding partial charges. This representation is richer than the original atomic structure because it takes advantage of the chemical knowledge in QUANTA. The choice of QUANTA was arbitrary and any similar molecular modeling package would have been suitable.

Two basic relations were used to represent structure: *atom* and *bond*. For example, for compound 127 (3,4,3′-trinitrobiphenyl), *atom*(*127, 127_1, C, 22, 0.191*). states that in compound 127, atom no. 1 is a carbon atom of QUANTA type 22 with a partial charge of 0.191. Equivalently, *bond*(*127, 127_1, 127_6, 7*). states that in compound 127, atom no. 1 and atom no. 6 are connected by a bond of type 7 (aromatic). The relation representation is completely general for chemical compounds and no special attributes need to be invented. The structural information of these compounds was represented by ≈18,300 facts of background knowledge.

The PROGOL algorithm allows for the inclusion of complex background knowledge that can be either explicit as facts or in the form of computer programs. This allows the addition in a unified way of any information that is considered relevant to learning the SAR. Generally the input programs are in the language PROLOG, but they could be in any language that can be linked to PROLOG—e.g., a FORTRAN program to assign partial charges. We have investigated the importance of background knowledge in the learning process by adding a set of PROLOG programs to the background knowledge that define some higher level chemical structures formed from atoms and bonds. Definitions/programs for the following high-level chemical concepts were formed in PROLOG: methyl group, nitro group, ring length (five- or six-membered), aromatic ring, heteroaromatic ring, connected rings, and the three distinct topological ways to connect three benzenes. It is important to appreciate that encoding PROLOG programs to define these concepts is not the same as including them as attributes. 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, an attribute-based representation would be able to use only the absence or presence of the different groups, not a bonded combination of them. To represent compounds to the equivalent level of detail using attributes would require several orders of magnitude more attributes than needed for only the simple atom/bond representation (see *Discussion*).

Two versions of the atom/bond representation were tested: representation I (atoms, bonds) and representation II (atoms, bonds, LogP, LUMO, and the above PROLOG programs).

Other SAR Algorithms Used for Comparison with PROGOL. The dataset of aromatic nitro compounds has previously been studied by linear regression (9) and the neural network algorithm back-propagation (18). We have repeated these studies to allow cross-validated comparison with our work. We applied regression methods by using the Minitab package (Minitab, Pennsylvania State University). Two variations of regression were used: basic linear regression, and regression using the dependant variables plus their squares (this allows simple nonlinear behavior and was found in initial trials to be as effective as quadratic regression). We applied the neural network algorithm back-propagation using the NN program (written by J. D. Hirst of the Imperial Cancer Research Fund and incorporating the GEAR algorithm to solve sets of stiff differential equations). We used the same network topology as previously used by Villemin *et al.* (18) (three hidden units). We also applied to the data the nonparametric decision tree

algorithm CART (20). In CART each node corresponds to a split of the data based on an attribute; the leafs correspond to classes. The CART algorithm was taken from the Ind package (National Aeronautics and Space Administration Ames Research Center, MS 269-2, Moffat Field, CA 94035-1000). Taken together, linear regression, back-propagation, and CART pose a formidable challenge to PROGOL, as they have all been shown to be accurate and robust general classification algorithms (21, 22) and they have also been shown to be successful on SAR problems (6–8).

Statistical Evaluation of Methods. To compare two prediction methods the McNemar test for changes was used (23). This is a binomial test based on the discrepant predictions of the methods. The null hypothesis is that the number of cases where method 1 predicts class high mutagenic and method 2 predicts low mutagenic is the same as the number of cases where method 1 predicts class low mutagenic and method 2 predicts high mutagenic. To show that a prediction method was better than random we used the McNemar test to compare the method with default accuracy (predicting all examples to be in the largest class). Note that this test is stronger than a simple χ^2 test, as that would test the weaker default hypothesis of random proportional guessing.

RESULTS

Table 1 gives the results for the different methods on the 188 dataset and the 42 dataset with and without indicator variables.

Representation I on the 188 Dataset. PROGOL applied to the 188 compounds and using atoms, bonds, and numerical inequalities found a theory with an estimated accuracy of 81.4% that consisted of five rules: A compound is highly mutagenic if

(*i*) it has an aliphatic atom carbon attached by a single bond to a carbon atom which is in a six-membered aromatic ring, or
(*ii*) it has a carbon atom in an aryl–aryl bond between two benzene rings with a partial charge ≥ 0.010, or
(*iii*) it has an oxygen atom in a nitro (or related) group with a partial charge ≤ 0.406, or
(*iv*) it has a hydrogen atom with a partial charge of 0.146, or
(*v*) it has a carbon atom that merges six-membered aromatic rings with a partial charge ≤ 0.005.

Table 1. Cross-validation prediction results

Dataset	Theory	Accuracy, % Without indicators	Accuracy, % With indicators
188	REG	85.2	89.3
	REG+	83.0*	88.8
	NN	86.2	89.4
	CART	82.5*	88.3
	PROGOL I	81.4†	—
	PROGOL II	87.8	—
42	REG	66.7†‡	66.7†‡
	REG+	71.8†‡	69.0†‡
	NN	64.3†‡	69.0†‡
	CART	83.3	83.3
	PROGOL I	85.7	—
	PROGOL II	83.3	—

Accuracy is defined as (no. of correct predictions)/(no. of predictions made) (for all drugs predicted). REG, linear regression; REG+, linear regression plus squares; NN, back-propagation; PROGOL I, PROGOL with representation I; and PROGOL II, PROGOL with representation II.
*Accuracy significantly worse ($P < 0.1$) than PROGOL I.
†Accuracy significantly worse ($P < 0.025$) than PROGOL I.
‡Accuracy significantly worse ($P < 0.025$) than PROGOL II.

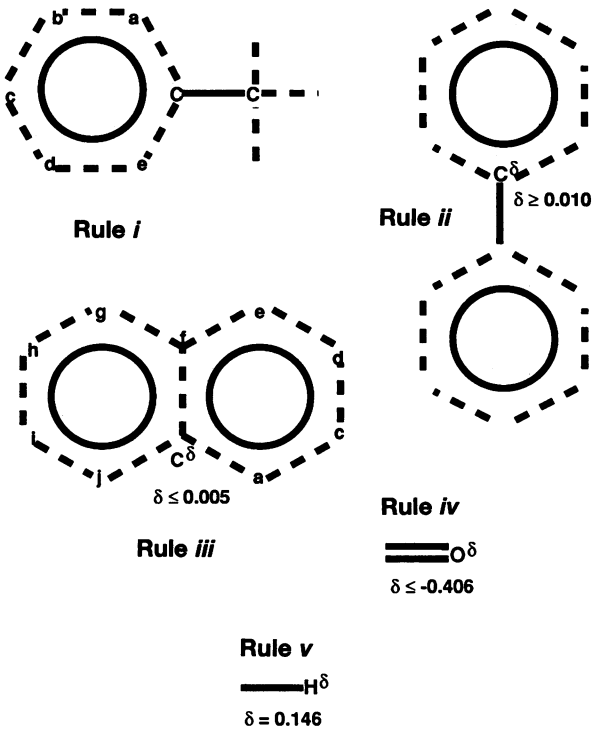

FIG. 1. The structural features of the theory learned by PROGOL using representation I on the 188 dataset. The dashed lines represent structure implied by the PROLOG rule; lowercase letters represent atoms that are not necessarily carbon.

A structural interpretation of the theory is given in Fig. 1. The PROGOL theory has the advantage of providing direct insight into the mutagenesis process. Rule *i* is a shape-based (steric) indicator of mutagenesis; it does not appear to be explainable by hydrophobic or electronic effects. Rule *ii* is a combination of a biphenyl shape-based feature and the electronic effect of a high partial charge on the aromatic carbon. This electronic effect supports the idea that electron-withdrawing rings can promote mutagenesis by promoting the initial reduction of a nitro group [a possible rate-limiting step in nitroarene activation (9)]. It has also been argued that electron-withdrawing rings should boost mutagenicity by increasing the lifetime of hydroxylamine, thereby giving it time to diffuse to DNA (24). The low partial charge on the oxygen in rule *iii* should mediate reduction to the amine. In rule *iv* the positive partial charges of the hydrogens may also indicate the presence of electron-withdrawing groups. Rule *v* is an indicator of high hydrophobicity; more hydrophobic rings have low partial charges on the aromatic carbons which overrides the expected electronic effect.

There is no significant difference (at $P < 0.1$) between the accuracy of this theory and the accuracy of the results obtained with any other method (see below) using only the LUMO and LogP attributes (i.e., excluding the indicator variables). This means that it is possible to do as well using PROGOL and a simple molecular representation, on a dataset especially selected to be suitable for regression, as it is with state-of-the-art statistical methods. It is possible to do significantly better than PROGOL only if indicator variables are included. However, this comparison is statistically biased against PROGOL, as the indicator variables were devised after visual inspection of the full dataset, not cross-validated subsections.

Representation II on the 188 Dataset. By using representation II a theory with an estimated accuracy of 87.8% was found that consisted of three rules: A compound is highly mutagenic if

(*i*) it has LUMO ≤ -1.870, or

(*ii*) it has LUMO ≤ -1.145 and a five-membered ring, or
(*iii*) it has LogP ≥ 4.180.

The PROGOL theory is simple and easy to understand. Rule *i* states that low values of LUMO indicate mutagenicity [as shown by Debnath *et al.* (9)]. Rule *ii* shows that this effect is modulated by the structural feature of a five-membered ring; this is very similar to the I_a indicator variable of Debnath *et al.* (9), who considered it before choosing the more specific structural feature. Rule *iii* states that high values of LogP indicate mutagenicity [also shown by Debnath *et al.* (9)].

No algorithm is significantly more accurate (at $P < 0.1$) with or without the indicator variables. This accuracy is significantly higher (at $P < 0.1$) than the results obtained by regression plus squares and CART without indicator variable (see below). By using PROGOL and a generic molecular representation it is possible to do as well as state-of-the-art statistical methods using a carefully hand-crafted representation and a dataset selected to be suitable for regression.

Comparative Results on the 188 Dataset. We summarize the results with the other methods. Applying linear regression produced results consistent with those of Debnath *et al.* (9). The approach with the highest estimated accuracy (89.3%) was basic linear regression with two indicator variables that produced the following model: log TA98 = $-(2.94 \pm 0.33) + (0.10 \pm 0.08)$LogP $- (1.42 \pm 0.16)$LUMO $- (2.36 \pm 0.50)I_a + (2.38 \pm 0.23)I_1$ ($n = 188, r = 0.844, s = 1.085, F = 113.42$). The high accuracy of this model indicates that the data is quite linear using these attributes.

The back-propagation neural network yielded results which are consistent with the work of Villemin *et al.* (18). The accuracies produced by back-propagation are higher than those produced by basic linear regression, but the differences are not significant at $P < 0.1$. It is not possible in a simple way to form an explicit SAR or to interpret chemically the meaning of a neural network.

The CART algorithm using the LogP and LUMO attributes and the indicator variables produced a cross-validated estimated accuracy of 88.3%. There is no significant difference (at $P < 0.1$) between the accuracy of CART and the accuracy of linear regression. CART produced very simple and easy-to-interpret decision trees.

PROGOL on the 42 Dataset. PROGOL found the same theory for the 42 compounds using representations I and II. This theory consists of a single rule and is the *optimal* theory possible given the descriptive languages and the compression measure. The rule states that an indicator for high mutagenicity is a double bond conjugated to a five-membered aromatic ring via a carbon atom (Fig. 2). This rule is a new structural indicator for high mutagenicity in chemical compounds. The conjugated double bond should stabilize the five-membered aromatic ring, and this may allow greater time for the compound to diffuse to the target site. The accuracy of this theory, estimated by leave-one-out cross-validation, is 85.7% for representation I and 83.3% for representation II. These differences are caused by chance effects causing PROGOL, for one of the splits, to find a more compressive theory on the training data that does not perform as well on the test data. Such effects are more likely with smaller datasets. The results for PROGOL are higher than for any other method with or without indicators. For representations I and II the results are significantly better (at $P < 0.025$) than for all other methods except CART (with or without indicator variables). This illustrates the robust nature of the basic atom/bond representation and machine learning.

Comparative Results on the 42 Dataset. It was not found possible, except with CART, to obtain results significantly better (at $P < 0.1$) than the default accuracy (that of the largest class). The results obtained with regression show that the indicator variables are not applicable to this dataset; i.e., they are not generic. The relationship between mutagenicity and LogP is reversed on the 42 dataset compared with the 188 dataset: high LogP is associated with low mutagenicity. Improvements in the results from neural networks may have been possible with a different training regime and net topology. CART produced a very simple tree with just one node based on LogP (hydrophobicity < 1.195, then active: hydrophobicity ≥ 1.195, then inactive). However, this tree provides little insight into the stereochemistry of mutagenesis.

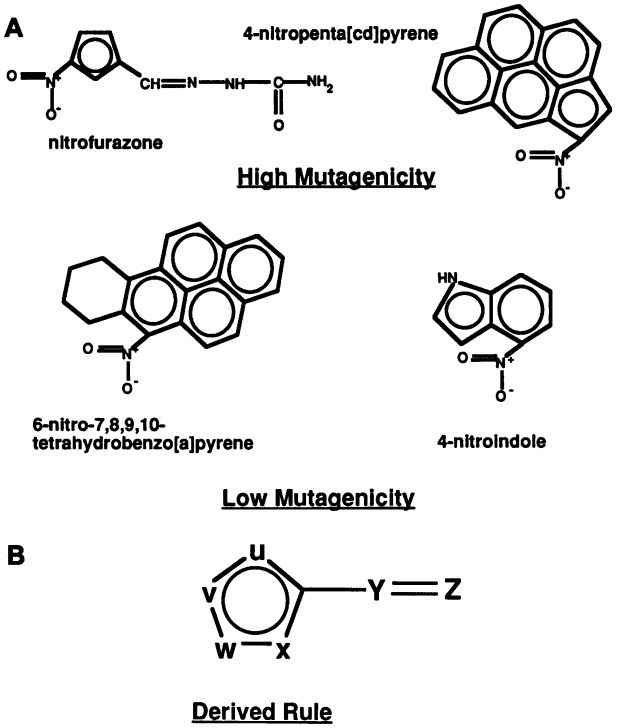

FIG. 2. (*A*) Example compounds with low mutagenicity explained by the structural feature. (*B*) The structural feature found by the three versions of PROGOL on the 42 dataset; atoms U–Z are not necessarily carbon.

DISCUSSION

Several workers have investigated SAR representations by using atoms and bond connectivities [e.g., substructural units (11)]. These previous representations were attribute-based and not relational. This meant that before learning proceeded it was necessary to enumerate (perhaps by use of a computer program) all potentially interesting combinations of atoms and bonds (fragments). Each combination of interest then became an attribute. This procedure potentially produces a prohibitive number of attributes for compounds with complex structure. For the data investigated in this paper we calculate that it would require 1,055,943 attributes per example to represent the compounds in the same level of detail as our simplest relational representation. This number of attributes is beyond the capabilities of any learning algorithm (statistical, neural network, or symbolic machine learning). In practice, attribute-based learners have to compromise on the number of substructures they can consider, and attributes that have not been precoded will not be learned.

The PROGOL algorithm marks an important advance in the field of ILP. In our previous work applying machine learning to SARs we used the program GOLEM (25), which had the ability to use certain types of relational information. However, this ability was limited by a determinacy constraint (15), which meant that when chemical compounds were represented each substructure could be connected only to at most one other substructure. GOLEM could therefore not use the basic atom/

bond connectivities representation. In addition, GOLEM could use only knowledge expressed as facts, whereas PROGOL can use both facts and rules (arbitrary PROLOG programs). A further advantage of PROGOL over GOLEM is its ability to do a complete search of the rules space, thereby finding the optimal rule explaining an example.

A major limitation of the work is that three-dimensional structure is not used. The simplest way to include this would be to extend the atom predicate by including Cartesian coordinates and by including background knowledge about Euclidean space (Pythagoras' theorem and trigonometry). No more information is necessary for PROGOL to learn rules about three-dimensional structures. Such a SAR method would have advantages over existing methods because it would not be necessary to align the compounds to a coordinate frame. Alignment is necessary in existing methods because compounds are described by using attributes, and so the only way that a point on one compound can be related to one on another is by the aligned grid—a universal property of all the compounds. If the compounds were represented relationally there would be no need to align the compounds explicitly, as the logical process of induction aligns the represented structures of the compounds internally in a computationally efficient way. A similar relational approach could be applied to CoMFA-type grid information (14). This might allow a more elegant formulation of the CoMFA idea by removing the need for explicit alignment of the compounds and the need to represent explicitly empty space.

Conclusions. This paper presents a generic relational method for representing compounds in SAR models. This method is based on an atom/bond representation that is simple, powerful, and generally applicable to any SAR problem. All previous atom/bond-based representations have used attributes. Relational representations are more powerful than those based on attributes. For atom/bond-based representations this increased power means that there is no need for explicit enumeration of all possible structures of interest.

It is clear from the theory and practice of statistics and machine learning that there is no single best SAR modeling algorithm (22, 26). The method that is best for a particular problem depends on the features of the data and the form of the required answer. This means that the computational chemist's toolbox should include a variety of methods. Specifically, we consider there is a role for the use of PROGOL and the relational atom/bond representation for problems where a small number of unknown shape/structure features are important, where it is important to obtain chemical insight, and where it has proven difficult to fit by using other techniques. If many independent attributes add together to produce activity, then the PROGOL approach is unlikely to be successful, as this would contradict the algorithm's inductive bias.

Finally, the advantages of using relational representations and automatic inference may not be limited to modeling SARs. It is possible that other branches of chemistry may benefit. For example, in synthesis planning it is possible to envisage a system that combines the deductive power of PROLOG with the induction of PROGOL.

Program Availability. The ILP program PROGOL (implemented in PROLOG) and the data used in this paper can be obtained by request from Ashwin Srinivasan, Oxford Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, United Kingdom, Ashwin.Srinivasan@comlab.oxford.ac.uk; they are freely available to academics. A version of PROGOL is also available that is implemented in C language.

We thank Professor D. Michie for suggesting analysis of regression-unfriendly data, Dr. R. Jackson for chemical advice, and Dr. J. D. Hirst for use of his NN program. This work was supported by the European Union's Information Technologies Program (6020), the Sciences and Engineering Research Council, and the Imperial Cancer Research Fund.

1. Martin, Y. C. (1978) *Quantitative Drug Design: A Critical Introduction* (Dekker, New York).
2. Ramsden, C. (1990) *Comprehensive Medicinal Chemistry 4* (Pergamon, Oxford).
3. Muggleton, S. H. (1995) *New Gener. Comput.* **13,** 245–286.
4. Hansch, C., Maloney, P. P., Fujita, T. & Muir, R. M. (1962) *Nature (London)* **194,** 178–180.
5. Frank, I. E. & Friedman, J. H. (1993) *Technometrics* **35,** 109–135.
6. Hirst, J. D., King, R. D. & Sternberg, M. J. E. (1994) *J. Comput. Aided Mol. Des.* **8,** 405–420.
7. Hirst, J. D., King, R. D. & Sternberg, M. J. E. (1994) *J. Comput. Aided Mol. Des.* **8,** 421–432.
8. King, R. D., Hirst, J. D. & Sternberg, M. J. E. (1994) *Appl. Artif. Intell.* **9,** 213–234.
9. Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J. & Hansch, C. (1991) *J. Med. Chem.* **34,** 786–797.
10. Trinajstic, N. (1983) *Chemical Graph Theory* (CRC, Boca Raton, FL).
11. Klopman, G. (1984) *J. Am. Chem. Soc.* **106,** 7315–7321.
12. Ormerod, A., Willet, P. & Bawden, D. (1989) *Quant. Struct. Act. Relat.* **8,** 115–129.
13. Hopfinger, A. J. (1980) *J. Am. Chem. Soc.* **102,** 7196–7206.
14. Cramer, R. D., Patterson, D. E. & Bunce, J. D. (1988) *J. Am. Chem. Soc.* **110,** 5959–5967.
15. Muggleton, S. (1991) *New Gener. Comput.* **8,** 295–318.
16. Lavrac, N. & Dzeroski, S. (1994) *Inductive Logic Programming Techniques and Applications* (Horwood, London).
17. DeLong, H. (1970) *A Profile of Mathematical Logic* (Addison–Wesley, Reading, MA).
18. Villemin, D., Cherqaoui, D. & Cense, J. M. (1993) *J. Chim. Phys.* **90,** 1505–1519.
19. Wallace, C. S. & Freeman, P. R. (1987) *J. R. Stat. Soc. B* **49,** 195–209.
20. Breiman, L., Friedman, J. H., Olshen, R. & Stone, C. J. (1984) *Classification and Regression Trees* (Wadsworth and Brooks, Monterey, CA).
21. King, R. D., Feng, C. & Sutherland, A. (1995) *Appl. Artif. Intell.* **9,** 289–334.
22. Michie, D., Spiegelhalter, D. J. & Taylor, C. C. (1994) *Machine Learning and Statistical Classification* (Horwood, London).
23. McNemar, Q. (1947) *Psychometrica* **12,** 153.
24. Vance, W. A., Okamoto, H. S. & Wang, Y. Y. (1988) in *Carcinogenic and Mutagenic Responses to Aromatic Amines and Nitroarenes*, eds. King, C. M., Romano, L. J. & Schetzle, D. (Elsevier, New York), p. 291.
25. King, R. D., Muggleton, S., Lewis, R. A. & Sternberg, M. J. E. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 11322–11326.
26. Weiss, S. M. & Kulikowski, C. A. (1991) *Computer Systems That Learn* (Kaufmann, San Mateo, CA).