New Generation Computing, 13 (1995) 411-433
OHMSHA, LTD. and Springer-Verlag

Relating Chemical Activity to Structure: An Examination of ILP Successes

Ross D. KING and Michael J. E. STERNBERG
Biomolecular Modelling Laboratory,
Imperical Cancer Research Fund,
44 Lincoln's Inn Fields, P. O. Box 123, London, WC2A 3PX, U.K.
Ashwin SRINIVASAN
Oxford University Computing Laboratory,
Wolfson Building, Parks Road, Oxford, OX1 3QD, U.K.

Received 6 October 1994
Revised manuscript received 1 February 1995

Abstract Problems concerned with learning the relationships between molecular structure and activity have been important test-beds for Inductive Logic programming (ILP) systems. In this paper we examine these applications and empirically evaluate the extent to which a first-order representation was required. We compared ILP theories with those constructed using standard linear reqression and a decision-tree learner on a series of progressively more difficult problems. When a propositional encoding is feasible for the feature-based algorithms, we show that such algorithms are capable of matching the predictive accuracies of an ILP theory. However, as the complexity of the compounds considered increased, propositional encodings becomes intractable. In such cases, our results show that ILP programs can still continue to construct accurate, understandable theories. Based on this evidence, we propose future work to realise fully the potential of ILP in structure-activity problem.

Keywords: ILP, inductive, chemistry, comparison, regression, decision-tree

§1 Introduction

A central concern of chemistry is understanding the relationships between chemical structure and activity. In most cases, these relationships cannot be derived solely from physical theory and experimental evidence is essential. Such empirically derived relationships are called Structure Activity Relationships (SARs). In a typical SAR problem, a set of chemicals of known structure and activity are given, and the problem is to construct a predictive

theory relating the structure of a compound to its activity. This relationship can them be used to select for structures with high or low activity. Typically, knowledge of such relationships form the basis for devising clinically effective non-toxic drugs.

For some time, structure activity problems have been an important area of application for programs developed within the framework of Inductive Logic Programming[6,7,13]. By using first-order logic as its representation language, Inductive Logic Programming (ILP) appears better suited to the efficient analysis of complex structured objects, like molecules, than programs that perform a feature-based analysis. Statistical methods, including traditional parametric techniques of regression, and modern non-parametric techniques of decision trees fall into this latter category. In this paper, we examine the case for ILP in the structure activity problems tackled to date. For these problems, we bring together results in historical sequence, and for each, we retrospectively consider the following:

Question
Was the first-order representation necessary, or is a propositional representation adequate?

To judge "adequacy", we consider recoding the problems propositionally and compare the predictive performance of theories learnt with this representation against their ILP counterparts. As representatives of propositional learners, we use standard linear regression and the decision-tree program CART.[2] The ILP program we first use is *Golem*.[11] Due to various restrictions within this program, we later adopt *Progol*.[10]

Firstly, it should be noted that a learner using the richer, first-order representation can perform *worse* than one using a propositional representation because of the restrictions used in many ILP learners. Further, a comparison on accuracy alone ignores the chemist's natural inclination to use a relational representation. Consequently, when judging adequacy, we also make a subjective assessment of the understandability and ease of use of the propositional representation.

For the problems considered, the results reported in this paper support the following answers to the question posed:

Answer 1
When propositional learning is feasible, ILP learners give results that are comparable in accuracy and chemically easy to understand; and

Answer 2
When propositional learning is infeasible, ILP learners still give good results.

Propositional learning is taken to be 'feasible' if the data can be encoded with a reasonable number of attributes (at most 1000).

The paper is organised as follows: Section 2 describes the structure activity problem and Sections 3-6 examine the Question posed in cases where ILP has yielded interesting results. The order of cases is historical, and describes a line of research that has progressively increased the complexity of molecules considered to the point where ILP learners appear essential. In Section 7 we describe future work to explore the capabilities of an ILP learner in this field. Section 8 summarises and concludes this report. We caution the reader that we have not attempted detailed descriptions of the materials and methods for each case study. For these, we provide adequate references to the literature.

§2 Chemical Structure Activity Relationship (SAR)

Forming a SAR is a basic step in drug design, and any method that can form improved SARs would speed up the process and perhaps help find better drugs. An average drug takes 12 years and $\approx$ £10^8 to develop.[8] SARs are also important in environmental toxicology. It is very expensive and time consuming to carry-out toxicological tests, e.g. a carcinogenicity test of a chemical takes $\approx$ 5 years to complete.[1] It is therefore desirable to have some method to predict toxicology from molecular structure. It is estimated that there 10^5 types of untested man-made potential carcinogens in the environment.[1]

The traditional method of developing SARs is to use linear regression.[4,9] While other more sophisticated statistical methods are also commonly used,[12] recently neural network and symbolic machine learning methods have also started to be applied. For these methods a large number of attribute based chemical representations have been developed e.g.: Hansch type parameters, topological descriptors, quantum mechanical descriptors, substructural units molecular shape (MS), and molecular fields (CoMFA).

It is widely recognised in the SAR community that existing methods are poor at representing chemical structure. There are two main methods used to try and get around the problem: the manual invention of new structural attributes, and structural alignment. The former relies heavilly on chemical experience, and to some extent, trial and error. The latter forms the basis of the CoMFA methodology. Here the chemicals are aligned to a 3 dimensional grid space, with the attributes being voxols of space. This idea has been successful, but it needs chemicals to be able to be aligned in a meaningful way. This is not possible for heterogeneous compounds. Further, the use of voxol attributes produces large numbers of attributes (possibly hundreds) and special resampling statistical methods need to be applied to avoid forming spurious SARs.

In contrast to the attribute-based representation, the constructs used by chemists in normal discourse are relational. These relational representations are either based on atom/bond connectivities or on Cartesian co-ordinates. Established knowledge of this form finds only restricted expression within feature based analysis tools, and hence the interest in the role of ILP in developing SARs. First, it may be possible to discover relationships that are inaccessible to

a conventional method (because the SAR is relational), and second, it is possible to express the SARs in a language easily understood by chemists. The question in Section 1, aims to examine the extent to which ILP has been successful on these two fronts.

§3 Case Study 1: E. Coli Dihydrofolate Reductase by Pyrimidines

3.1 Problem Description

We first consider the classic drug design problem of inhibition of E. Coli Dihydrofolate Reductase and Pyrimidines. Pyrimadine compounds are antibiotics. They act by inhibiting Dihydrolate Reductase, an enzyme on the pathway to forming DNA. They inhibit bacterial forms of the enzyme preferentially to the human form and therefore kill bacteria. Further details on experimental design are available in Ref. 6).

3.2 Representation

In this problem, the chemicals structures of all compounds used to induce the SAR can be considered to have a common template (see Fig. 1). To this template, chemical groups can be added at three possible substitution positions, 3, 4, and 5. A chemical group is an atom or set of structurally connected atoms that can be substituted together as a unit and have well defined chemical properties.

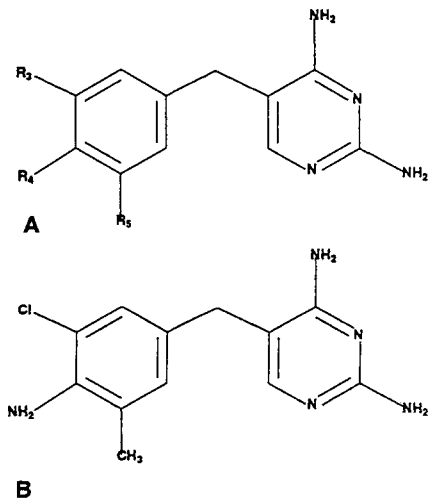


Fig. 1 Chemical structure of pyramidine compounds. (A) Generic template for all compounds in the study; and (B) An example compound: $3—Cl$, $4—NH_2$, $5—CH_3$.

(1) ILP representation

The existence of a template with only three possible substitution positions gives the pyrimidine problem a relatively small structural component. The

chemical structure of the example compound in Fig. 1 is adequately represented by a Prolog fact of the form:

struc(d55, cl, nh2 ch3)

which is intended to represent that drug 55 has a chlorine atom substituted at position 3, a amino group (NH_2) group substituted at position 4, and a methyl (CH_3) group at position 5.

In Ref. 6), the authors use 9 integer-valued attributes to represent chemical properties of the substituents. These are shown in Fig. 2. These were encoded as predicates of the form:

polar(br, polar3)

which states that a bromine atom has a polarity of value 3.

Positive examples are pairs of drugs, the activity of one being known to be higher than the other. For example:

great(d1, d2)

states that the E. Coli Dihydrofolate Reductase inhibition by drug $d1$ is higher than that by $d2$.

〔2〕 Propositional represention

The chemical structures used in Ref. 6) can be encoded by 27 integer-valued attributes (nine each for positions 3, 4, and 5). In order to learn the same pairwise comparison relationship considered by an ILP program, each example is actually a pair of drugs. The simple-minded approach of concatening the sets of attributes for each drug appears to be adequate. Thus each drug-pair can be represented by 54 attributes.

To be fully comparable to the ILP learner, a propositional learner also has to account for the fact that former has access to predicates that, for each drug pair, allow pairwise comparison of chemical properties for each of the positions. For each drug-pair comparisons are possible for substituent at six positions (three of each drug). For example, it is possible to express that the polarity of the substituent at position 3 is the same as, or greater than the polarity of the substituent at position 4. To be able to express the same, a propositional learner would require an additional $2 \times {}^6C_2 = 30$ boolean-valued attributes for each property considered (the multiplicative factor 2 arises from the need to have 1 attribute for equality, and 1 for expressing if one value is greater than the other). Mutal comparison of the 9 properties (Fig. 2), at 3 substituent positions therefore requires $9 \times 30 = 270$ boolean-valued attributes.

Our preliminary experiments however, suggested thet the predicates for comparison were not of much use for the ILP learner. Consequently, we did not encode these addidtional attributes for the propositional learners. As will be seen in the following section, this omission does not cost us unduly. One

Chemical Property	Number of Values (values in parentheses)
Polarity	6 (polar0, polar1,...,polar5)
Size	6 (size0, size1,...,size5)
Hydrogen-bond donor	3 (h_doner0, h_doner1, h_doner2)
Hydrogen-bond acceptor	3 (h_acceptor, h_acceptor, h_acceptor)
pi-donor	3 (p_doner0, p_doner1, p_doner2)
pi-acceptor	3 (p_acceptor0, p_acceptor1, p_acceptor2)
Polarisability	4 (polarisible0,...,polarisible3)
Sigma effect	6 (sigma0, sigma1,...,sigma5)

Fig. 2 Chemical properties of substituents and their valuea.

Number of compounds	55
Number of positive examples	1394
Number of negative examples	1394
Background clauses	2116

(A) ILP

Number of compounds	55
Number of positive examples	1394
Number of negative examples	1394
Attributes per example	55

(B) Propositional

Fig. 3 Data summaries for Dihydrofolate Reductase by pyrimidines.

additional boolean-valued attribute is required to specify if the first drug in the pair has higher activity than the second.

A summary of the sizes of background knowledge and examples for both ILP and propositional learners is in Fig. 3.

3.3 Experimental Results

The ILP program *Golem* was used to develop a SAR for the compounds. The mean accuracies calculated from a five-fold cross-validation study of the data is in Fig. 4. Linear regression was done stepwise using the program Minitab with variables and their squares (this allows for some simple non-linearities). Figure 5 gives a summary description of the theory resulting from each method. *Golem*'s mean Spearman rank correlation coefficients with the true activity ordering is 0.684. This is not significantly higher ($P = 0.05$) that the results

Method	Mean Spearman's rank
Golem	0.684(0.107)
Regression	0.654(0.104)
CART	0.499(0.294)

Fig. 4 Rank correlation of predicted activity ordering against true ordering of pyramidine compounds. Mean Spearman's rank correlation coefficient calculated from a five-fold cross-validatation. Standard deviations are in parentheses.

Method	Description of theory
Golem	9 clause Prolog program
Regression	Equation with 7 independent variables
CART	Classification tree with 49 leaves

Fig. 5 Description of theories produced for Dihydrofolate Reductase inhibition by pyramidine compounds. Sizes of theories are averaged over results from each data split of the cross-validation.

achieved by the propositional learning methods.

On the issue of chemical understandability, the Prolog rules found by *Golem* for the SAR are easily comprehensible and can be simply translated into meaningful rules for a chemist. An example is shown in Fig. 6.

```
great(Drug1,Drug2):-
        struc(Drug1,Pos_3,Pos_4,_),
        struc(Drug2,_,_,h),
        h_donor(Pos_3,h_don0),
        pi_acceptor(Pos_3,pi_acc0),
        polar(Pos_3,Polarity),
        great0_polar(Polarity),
        size(Pos_3,Size),
        less3_size(Size),
        polar(Pos_4,_).

Drug1 inhibits more than Drug2 if
        Drug1 has a position 3 substituent which
                is a hydrogen-bond donor with value 0;
                is a pi-acceptor with value 0;
                has polarity greater than 0;
                has size less than 3; and
        Drug1 has a substituent at position 4 (i.e. not hydrogen); and
        Drug2 does not have substituent at position 5
```

Fig. 6 A *Golem* SAR rule for comparing Dihydrofolate Reductase inhibition by a pair of pyramidine drugs, and its English translation.

From the *Golem* rules it is possible to deduce certain facts about the physical mechanism of inhibition: for example, that positions 3 and 5 of the template are partially buried within a hydrophobic environment. These deduc-

Regression:

$$Activity = 0.44 + 0.66(\pm 0.21)PO_3 - 0.82(\pm 0.23)PO_3^2 + 0.11(\pm 0.05)\pi A_4 + 0.79(\pm 0.15)SZ_5 - 0.94(\pm 0.10)FL_5 + 1.10(\pm 0.18)\pi A_5 - 0.18(\pm 0.14)\pi A_5^2 \quad (1)$$

```
CART:
a27<0.5:
|  a46<-0.5:
|  |  a1<-0.5:
|  |  |  a38<4:
|  |  |  |  a11<4:
|  |  |  |  |  a29<1.5:
|  |  |  |  |  |  a44<0.5:
|  |  |  |  |  |  |  a38<2.5:
|  |  |  |  |  |  |  |  a10<-0.5: +0+20
|  |  |  |  |  |  |  |  a10>=-0.5:
|  |  |  |  |  |  |  |  |  a10<4.5: +35+11
|  |  |  |  |  |  |  |  |  a10>=4.5: +0+20
|  |  |  |  |  |  |  a38>=2.5: +3+60
|  |  |  |  |  |  a44>=0.5:
      (etc)
```

Fig. 7 An example of propositional theories for Dihydrofolate Reductase inhibition by a pair of pyramidine drugs. The independent variables in the regression equation are as follows: PO_3: polarity of substituent at position 3; $\pi A_{4,5}$: pi-acceptor values of substituents at positions 4,5; SZ_5: size of substituent at position 5; and FL_5: flexibility of substituent at position 5.

```
Regression:
Drug1 inhibits more than Drug2 if
        Drug1 has a position 3 substituent which
                has a small size;
                has low flexibility; and
        Drug1 has a position 4 substituent which
                is involved in a pi-bond

CART:
Drug1 inhibits more than Drug2 if
        Drug1 has a position 3 substituent which
                is a hydrogen-bond donor with value 0;
                has polarity that is not 0; and
        Drug1 has a position 4 substituent which
                has polarity that is not 0; and
        Drug1 has a substituent at position 5 (i.e. not hydrogen)
```

Fig. 8 Some SARs extracted by propositional learners for comparing Dihydrofolate Reductase inhibition by a pair of pyramidine drugs.

tions have been shown to be true by physically co-crystallising Dihydrofolate Reductase with pyrimidines.

Interpreting the results of linear regression and CART are not as straightforward. An example of results from one of the cross-validation runs is in Fig. 7 (for reasons of space, only part of the CART tree is shown). Figure 8 shows some general chemical trends that we have been managed to extract from the output of the propositional learners.

§4 Case Study 2: E. Coli Dihydrofolate Reductase by Triazines

4.1 Problem Description

The study in Section 3 was extended to the inhibition of E. Coli Dihydrofolate Reductase by triazines. Triazines act as anti-cancer agents by inhibiting the enzyme Dihydrolate Reductase. They act by preferentially inhibiting reproducing cells. Further details on experimental design are available in Ref. 5).

4.2 Representation

Like the pyrimidines, the triazines can also be considered to have a common template structure (see Fig. 9). However, the chemical groups substituted onto the template are much more complicated than those in Section 3. Further, many of the substituting groups can more naturally be considered as sub-templates with substitutions. There are seven regions where a substituent might be present: the 2, 3, and 4 positions of the phenyl ring as shown in Fig. 9. Each substituent at positions 3 and 4 can in turn, itself contain a ring structure. In this case, further substitutions are possible into positions 3 and 4 of each additional rings.

(1) ILP representation

The first-order representation of the triazines is best explained using an

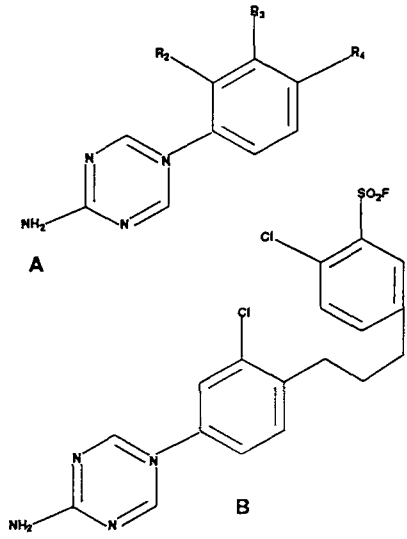


Fig. 9 Chemical structure of triazine compounds. (A) Generic template for all compounds in the study; and (B) An example compound: 3—*Cl*, 4—$(CH_2)_3C_6H_3$—4—*Cl*, 5—CH_3.

example. The example compound in Fig. 9 is represented by the following Prolog facts:

```
struc3(d217, cl, absent).
struc4(d217, (ch2)3, subst14).
subst(subst14, so2f, cl).
```

The first clause represents substitutions at position 3 on the basic template: a *Cl* is present and there is an absence of a further phenyl ring. The second clause represents substitutions at position 4 on the basic template: there is a $(CH_2)_3$ bridge to a second phenyl ring (implicit in the representation). This second phenyl ring has a SO_2F group substituted at position 3 and a *Cl* group substituted at position 4. This is represented using the linker constant subst14 to the third clause. There is no substitution at position 2 on the basic template. Each of the chemical groups had 10 attributes, 9 of which were the same as used in the study of pyrimidines. One further attribute was added to capture flexibility of a substituent. The degree of flexibility is represented by one of 9 values.

As in the previous study, examples are pairwise comparisons of the inhibitions of E. Coli Dihydrofolate Reductase.

(2) Propositional representation

We adopt the following propositional encoding for the data. Substitutions at the main position 2 are relatively rare and are ignored. 20 integer-valued attributes are required for the substitution positions 3, and 4; and 10 attributes

Number of compounds	186
Number of positive examples	17063
Number of negative examples	17063
Background clauses	1933

(A) ILP

Number of compounds	186
Number of positive examples	17063
Number of negative examples	17063
Attributes per example	121

(B) Propositional

Fig. 10 Data summaries for Dihydrofolate Reductase by triazines.

each for the possible substitutions on rings (if present) at the main positions 3 and 4. This gives a total of 60 attributes. As before, examples are actually drug-pairs, each of which can be represented by $60 + 60 = 120$ attributes.

Ignoring the substitution at the main position 2 leaves 6 possible substitution positions for a drug. For a drug pair therefore, pairwise comparison of substituent properties requires $10 \times 2 \times {}^{12}C_2 = 1320$ boolean attributes. However, as in Section 3, we found it unneccessary to include these binary attributes for comparison. Along with one additional attribute to specify the class value, each example is therefore a vector of 121 entries.

A summary of the sizes of background knowledge and examples for both ILP and propositional learners is in Fig. 10.

4.3 Experimental Results

As before, the ILP program *Golem* was used to develop a SAR for the compounds. The mean accuracies of the different methods, calculated from a six-fold cross-validation study of the data is in Fig. 11. Figure 12 gives a summary description of the theory resulting from each method. Again, no one method's rank ordering is significantly better ($P = 0.05$).

As before, the main advantage of the results obtained by *Golem* appears to be the comprehensibility of its SAR. The understandability of the rules contrast favourably with the other two methods examined. The rules make

Method	Mean Spearman's rank
Golem	0.431(0.166)
Regression	0.446(0.181)
CART	0.532(0.121)

Fig. 11 Rank correlation of predicted activity ordering against true ordering of triazine compounds. Mean Spearman's rank correlation coefficient calculated from a six-fold cross-validatation. Standard deviations are in parentheses.

Method	Description of theory
Golem	7 clause Prolog program
Regression	Equation with 10 independent variables
CART	Classification tree with 78 leaves

Fig. 12 Description of theories produced for Dihydrofolate Reductase inhibition by triazine compounds. Sizes of theories are averaged over results from each data split of the cross-validation.

several predictions, for example, that at position 4 there should be a substituent connected to a phenyl ring with polarity = 1. Unfortunately because triazines have not as yet been co-crystallised with Dihydrofolate Reductase, it is not possible to judge how well these rules correspond to physical reality. In contrast to *Golem*'s results, we have found it difficult to extract any meaningful chemical predictions from the regression equation. The relatively large size of the CART trees also pose a problem. We believe that they are broadly in consensus with the *Golem* rules, with additional constraints on the size of substituents at position 3.

§5 Case Study 3: Design of Tacrine Analogues

5.1 Problem Description

The relational representation used in the pyrimidine and triazine problems (Sections 3, 4) was then applied to a practical drug design problem of direct clinical interest. The problem concerns the design of analogues to the Alzheimer's disease drug tacrine. For Alzheimer's disease there is no known protein structure to inhibit, there is not even a well defined biological mechanism of action. Four biological/chemical properties are considered to be important and sets of rules are to be constructed for each of them: maximisation of acetyl cholinesterase inhibition, maximisation of inhibition of amine reuptake, maximisation of the reversal of scopolamine-induced memory impairment, and minimisation of toxicity. The individual rule-sets would then aid in developing, if possible, potential drugs that satisfy all of the sub-problems.* Further details on experimental design are available in Ref. 7.

5.2 Representation

A template based on the tacrine molecule is shown in Fig. 13. New compounds are created by substituting chemicals for R, X_6 and X_7. The X substitutions are much like the ones already studied in Sections 3 and 4. The R substitutions are more complicated. For the compounds in this study, the substitution consists of one or more single alkyl groups linked to one or more benzene rings. The linkage can either be direct, or through N or O atoms, or through a CH bond. As is readily observed, the compounds considered are more complicated than those in either problem described before. This makes the tacrine problem more relational than the studies in Sections 3, and 4.

(1) ILP representation

As before, we describe the relational representation using the example compound in Fig. 13. It has a Chlorine (Cl) substitution at position 7 and a complex substitution $(CH_2)_2CH(4-FC_6H_4)_2$ at position R. The position 7

* This work was done in collaboration with Chemistry and Pharmaceutical Departments of Strathclyde University.

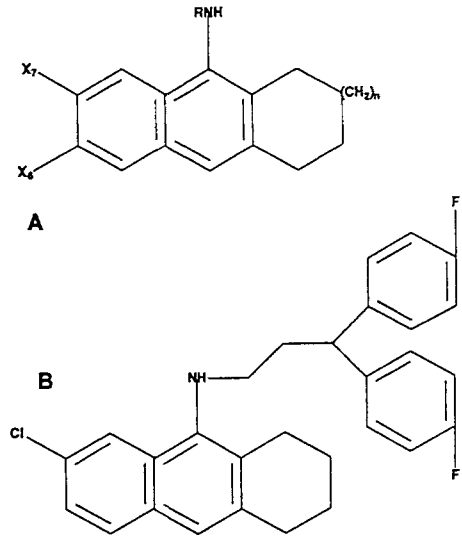


Fig. 13 Chemical structure of tacrine analogues. (A) Generic template for all compounds in the study; and (B) An example compound.

substitution is represented as follows:

x_subst(n1,7,cl).

In the compounds considered here, the *R* substitution can have up to 3 chemical units linked together. These units are either alkyl groups (like CH_2, CH) or ring (aromatic) structures. The latter can be benzyl derivatives that arise from substitutions within the basic benzene ring structure. There are 5 positions in the benzene structure at which a hydrogen atom may be substituted for a different chemical (one final position is reserved for linking up with the rest of the compound). The number of such substitutions in the basic benzene structure, the position of the substitutions on the actual substituent are provided. In the example compound considered here, the complex substitution at position *R* contains an alkyl chain and 2 substituted benzene groups. This is represented by the following facts:

alk_groups(n1, 4).
r_subst_1(n1, single_alk(2)).
r_subst_2(n1, double_alk(1)).
r_subst_3(n1, 3, aro(2)).
ring_substitution(n1, 1).
ring_subst_4(n1, f).

which state 1) that drug n1 has four alkyl groups; 2) the *R* substitution group in drug n1 starts with two methyl groups ((CH_2)$_2$; 3) this is followed by an ethyl

group (CH); 4) the final alkyl group in drug n1 has two substitutions; 5) drug n1 has two aromatic rings; 5) the aromatic rings have one substitution; and 6) the aromatic rings have substitutions at position 4.

The 10 attributes of chemical groups used in Section 4 are inherited here. For each of the biological/chemical properties analysed, examples are pairwise comparisons of the values of drugs for that property. For example:

less_toxic(d1,d2)

states that the toxicity of drug $d1$ is less than that of $d2$.

[2] Propositional representation

The X substituents are adequately described by 20 attributes (10 each for positions 6 and 7). The following encoding can capture the R substitution structure. One integer-valued attribute is required to specify the number of single alkyl groups at the start of the R substitution. A further 10 integer-valued attributes are required to specify the chemical properties of the group. One four-valued categoric attribute is sufficient to specify the linkage to the benzene rings. One integer-valued attribute is needed to specify the number of rings. Since all rings are identical, it is sufficient to have 1 integer to indicate the number of substitution and 50 further attributes to describe the ring substituents (10 each for each of the five possible substituents). This is a total for each compound of $20 + 1 + 10 + 1 + 1 + 50 = 83$ attributes. Following the ILP representation, examples are drug-pairs. Along with an attribute for class value, each such drug pair therefore requires 167 attributes.

Unlike the studies in Sections 3 and 4, our preliminary experiments suggested that pairwise comparison of chemical properties of the substituents were important. For each drug pair, there are 16 possible substituent positions (in order, for each drug, 2 X substituents, 1 at the start of the R substitution, and 5 possible substituents into a benzene ring), a comparison of any single property requires $2 \times {}^{16}C_2 = 240$ attributes. Mutual comparison of the 10 attributes, therefore requires 2400 attributes. In addition, for each drug pair, the following

Number of compounds	37
Number of positive examples	663 + 343 + 596 + 443 = 2045
Number of negative examples	663 + 343 + 596 + 443 = 2045
Background clauses	580

(A) ILP

Number of compounds	37
Number of positive examples	663 + 343 + 596 + 443 = 2045
Number of negative examples	663 + 343 + 596 + 443 = 2045
Attributes per example	2575

(B) Propositional

Fig. 14 Data summaries for tacrine analogues. Numbers of positive and negative examples shown as summations of numbers required for each of the four sub-problems.

R substitution comparisons are also possible: the number of alkyl groups, the number ring substitutions, the number of single and double alkyl bonds. These require $2 \times 4 = 8$ boolean attributes. therefore requires $167 + 2400 + 8 = 2575$ attributes.

A summary of the sizes of background knowledge and examples for both ILP and propositional learners is in Fig. 14.

5.3 Experimental Results

Given the large number of attributes required, propositional learning was deemed infeasible. Consequently, we do not report any comparative trials. Instead, we only describe the ILP-derived SAR and on a new drug suggested by the rules. For full details, we refer the reader to Ref. 7). An example rule for acetyl cholinesterase inhibition is in Fig. 15. This rule like many other found is quite relational and structural in nature.

The *Golem* rules have been used to suggest a number of compounds that could prove to be less toxic to the human body than tacrine. Figure 16 shows one such compound. Unfortunately, we are not aware of any laboratory tests having been performed to date on this compound.

```
better_acinh(Drug1, Drug2):-
        alk_groups(Drug1, N1),
        alk_groups(Drug2, N2),
        gt(N2, N1),
        ring_substitution(Drug2, 1).

Drug1 inhibits acetyl cholinesterase more than Drug2 if
        The R substitution of Drug2 has
                more alkyl groups than Drug1; and
                has a benzyl ring with 1 hydrogen atom substituted
```

Fig. 15 A *Golem* SAR rule for comparing acetyl cholinesterase inhibition by a pair of tacrine analogue drugs, and its English translation.

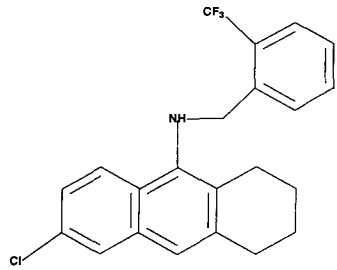


Fig. 16 A compound proposed by *Golem* rules to have similar efficacy to tacrine, but with lower toxicity.

§6 Case Study 3: Nitroaromatic and Heteroaromatic Mutagens

6.1 Problem Description

The problem concerns identification of mutagenic nitroaromatic and heteroaromatic compounds. Mutagenicity a strong indicator of carcinogenicity and is assessed using the Ames test. These compounds are very structurally heterogeneous (see Fig. 17).

The task here is to distinguish compounds with Ames test mutagenicity greater than 0, from those for which this value is at most 0. Further details on experimental design can be found in Ref. 13).

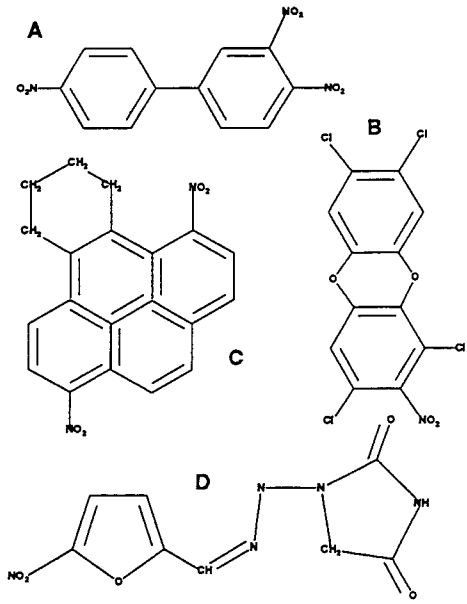


Fig. 17 Examples of the nitroaromatic and heteroaromatic compounds used in the mutagenesis study. (A) 3,4,4′-trinitrobiphenyl, (B) 2-nitro-1,3,7,8-tetrachlorodibenzo-1,4-dioxin, (C) 1,6-dinitro-9,10,11,12-tetrahydrobenzo[e]pyrene, (D) nitrofuraniton.

6.2 Representation

Representing the pyrimidine, triazine and tacrine analogues relied on the existence of a template on which substitutional groups are added. Using a template has the disadvantage that the representational scheme has to be tailored for each individual problem. A more fundamental problem with templates is that many SAR problems the compounds involved are so diverse that they cannot reasonably be said to be variations of a template.

(1) ILP representation

The obvious generic description of compounds is to use atoms and their bond connectivities. Unfortunately, such a description could not be used by algorithms life *Golem* which require a deterministic representation. This is not a property of the atom/bond representation, as a molecule is typically made up of many linked atoms.

The compounds were input into the molecular modelling program *QUANTA* using its chemical editing facility. *QUANTA* then automatically adds typing information and calculates approximate partial charges associated with each atom. The choice of QUANTA was arbitrary, any similar molecular modelling package would have been suitable. The result is that each compound is represented by a sets of facts of the form:

```
atom(127, 127_1, c, 22, 0.191)
bond(127, 127_1, 127_6, 7)
```

which states that in compound 127, atom number 1 is a carbon atom of QUANTA type 22 with a partial charge of 0.191, and atoms 1 and 6 are connected by a bond of type 7 (aromatic). This representation is not peculiar to the compounds in this study, and can be used to encode arbitrary chemical structures.

Positive examples are those compounds whose log mutagenicity (measured by the Ames test) is greater than 0. These compounds are termed 'active', and all others are deemed to be 'inactive'.

(2) Propositional representation

The lack of any well-defined template for the compounds poses some problems in capturing structural information in a propositional form. Consider the following straightforward method for encoding the data at hand.

From *QUANTA*'s typing mechanism, we can deduce that there are 726 distinct atom types. 726 integer-valued attributes can be used to indicate the number of atoms of each type present in a compound. We now turn to the bond connectivities. A pair of atoms may be connected to each other by one of 6 different bond-types. A further 6 integer-valued attributes will be needed to encode the number of bonds of each type for a molecule.

This does not, by itself capture all that can be said with the language restrictions used by the authors of Ref. 13). In that encoding, it is also possible to state:

(1) Bond connections between atoms of particular types; and
(2) That a compound has one or more atoms with at least (at most) some partial charge.

For the first, we need to associate the number and type of bonds between each pair of atom types. This requires a further 726^2 integer-valued attributes to

specify the number of bonds between atom-pairs, and 726^2 integer-valued attributes to specify the bond type. Secondly, *QUANTA* identifies 529 charge constants in the data. With 529 integer-valued attributes, it is possible to specify the number of atoms in the compound with partial charge that is at least that of the corresponding *QUANTA* constant. 529 more will be needed to specify the number with at most that value. The data available in Ref. 3) also provide 4 attributes identified by chemists as being potentially relevant. These can be used directly by bothe propositional and ILP learners. They are:

(1) The hydrophobicity of the compound (termed logP);
(2) The energy level of the lowest unoccupied moelcular orbital (termed LUMO);
(3) A boolean attribute identifying compounds with 3 or more benzene rings (termed I1); and
(4) A boolean attribute identifying a sub-class of compounds termed acenthryles (termed Ia).

One additional boolean-valued attribute is needed to indicate if the compound's Ames mutagenicity is greater than 0 oherwise. This accounts for a total of $726 + 6 + 726^2 + 726^2 + 529 + 529 + 4 + 1 = 1055947$ attributes to specify each compound.

A summary of the sizes of background knowledge and examples for both ILP and propositional learners is in Fig. 18.

Number of compounds	230
Number of positive examples	138
Number of negative examples	92
Background clauses	12203

(A) ILP

Number of compounds	230
Number of positive examples	138
Number of negative examples	92
Attributes per example	1055947

(B) Propositional

Fig. 18 Data summaries for mutagens

6.3 Experimental results

As in Section 5, propositional learning with all attributes is intractable. However, unlike the data in that Section, we do have access to four attributes that have been shown to be useful for a regression analysis. Using these attributes, the compounds had previously been split into two subsets by the authors of Ref. 3), 188 compounds that were known to be amenable to regression, and 42 that were not. We are therefore in a position to examine the performance of an ILP learner in a setting where linear modelling by regression is *known* to be inadequate for a subset of the data.

Given the non-determinate nature of the atom/bond representation (a given compounds has more than one atom, each associated with one or more bond connections), the ILP program *Golem* is unsuitable. Consequently, we use the more recent *Progol* system.[10] The propositional learners are as before.

However, they use the four attributes that have proven value. We further adopt the convention of learning rules for the minority class for each of the two distinct subsets identified by the chemists in Ref. 3). This is the 'inactive' class for the 188 compounds, and the 'active' class for the 42 compounds. This convention is in variance with that adopted in Ref. 13), where the authors learn rules for the 'active' class only. Figures 19 and 20 tabulate the comparative performance of the two forms of learning for the sets of data. Figures 21 and 22 give a summary description of the theory resulting from each method. There is no significant loss in accuracy ($P = 0.05$) in using *Progol* on the subset known to be well-suited for propositional learning. In contrast, as expected, linear regression performs significantly worse than either of the logic-based learners on the subset of 42 compounds.

The *Progol* SAR does have the advantage that it did not need the use of any specially tailored 'indicator' variables to include structural information. In contrast, these were crucial for the regression analysis. Figures 23 and 24 tabulate the comparative performance of the two forms of learning for the sets

Method	Estimated Accuracy
Progol	0.84
Regression	0.89
CART	0.88

Fig. 19 Estimated accuracy of theories for mutagenicity. Data are the 188 compounds known to be adequately explained by a regression analysis. Accuracy is estimated from a ten-fold cross-validation.

Method	Estimated Accuracy
Progol	0.83
Regression	0.69
CART	0.83

Fig. 20 Estimated accuracy of theories for mutagenicity. Data are the 42 compounds known to be unsuitable for a regression analysis. Accuracy is estimated from a leave-one-out procedure.

Method	Description of theory
Progol	4 clause Prolog program
Regression	Equation with 4 independent variables
CART	Classification tree with 3 leaves

Fig. 21 Description of theories produced for mutagenicity. Data are 188 compounds known to be adequately explained by a regression analysis.

Method	Description of theory
Progol	1 clause Prolog program
Regression	Equation with 4 independent variables
CART	Classification tree with 2 leaves

Fig. 22 Description of theories produced for mutagenicity. Data are 42 compounds known to be unsuitable for a regression analysis.

Method	Estimated Accuracy
Progol	0.84
Regression	0.83
CART	0.82

Fig. 23 Estimated accuracy of theories for mutagenicity. Data are the 188 compounds known to be adequately explained by a regression analysis. Indicator variables have been masked out for Regression and CART. Accuracy is estimated from a ten-fold cross-validation.

Method	Estimated Accuracy
Progol	0.83
Regression	0.71
CART	0.83

Fig. 24 Estimated accuracy of theories for mutagenicity. Data are the 42 compounds known to be unsuitable for a regression analysis. Indicator variables have been masked out for Regression and CART. Accuracy is estimated from a leave-one-out procedure.

```
inactive (A):-
        atm(A, C, c, 10, D),
        logp(A, B), lteq(B, 2.070).
```

A compound is not mutagenic if
it has an aliphatic carbon atom (QUANTA type 10); and
a hydrophobicity value of at most 2.070

Fig. 25 A *Progol* SAR rule for mutagenic (in)activity, and its English translation.

Regression:

$$Activity = -2.94(\pm 0.33) + 0.10(\pm 0.08)logP - 1.42(\pm 0.16)LUMO - 2.36(\pm 0.50)Ia + 2.38(0.23)I1 \quad (2)$$

```
CART:
I1 = 1: Active
I1 = 0:
|   LUMO <    -2.141: Active
|   LUMO >= -2.141: Inactive
```

Fig. 26 Propositional theories for mutagenic activity.

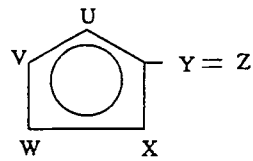


Fig. 27 Structural alert for mutagenicity found by *Progol*. Data are 42 compounds known to be unsuitable for a regression analysis.

of data, when the non ILP learners have been denied access to the indicator variables.

Figure 25 shows an example of a *Progol* rule obtained on the set of 188 compounds. Figure 26 shows the corresponding theories constructed by the propositional learners.

There appears to be little to choose in accuracy or simplicity between *Progol* and CART. The definition of inequalities (*lteq*/2, *gteq*/2) provided as background knowledge to *Progol* confines comparisons of hydrophobicity and energy values to constants that actually appear in the data. CART's comparisons are not restricted in this fashion. This appears to give it a certain robustness in obtaining good numerical thresholds. Of course, it is possible that numerical features like hydrophobicity and molecular energies may not always be available for new or complex molecular structures. In such cases, a *Progol* theory conveys much more information to a chemist, as it is readily visualised from atomic and bonding relationships. A case in point is the different SARs obtained by *Progol* and CART for the 42 'regression unfriendly' compounds. The *Progol* SAR identifies the structure shown in Fig. 27 as being mutagenic. This appears to be a new structural alert for mutagenicity. Contrast this with the SAR extracted by CART in Fig. 28.

```
hydrophobicity <  1.195: Active
hydrophobicity >= 1.195: Inactive
```

Fig. 28 Classification tree for mutagenicity found by CART. Data are 42 compounds known to be unsuitable for a regression analysis.

§7 Future Work

We believe that the ability to use atom and bond connectivities give ILP programs an edge over other simple structural analysis methods. We are currently looking to consolidate and extend the use of such information to problem involving heterogeneous compounds. To this end we plan to participate in an international competition to predict chemical carcinogenesis in rodents 1). The test compounds are currently being assayed in mice and rats. The compounds are very heterogeneous and should suit our atom/bond connectivities methodology.

We also intend to take advantage of *Progol*'s ability to use arbitrary Prolog programs as background knowledge. This allows us to give *Progol* more knowledge about higher level structural features, building on our simple atom/bond connectivities. For example, this could be Prolog programs encoding knowledge of aromatic and non-aromatic ring systems and how they can be connected. This information is generic and so can be reused in different SAR problems.

The next logical step in representing chemical structure is to add three-dimensional information for the compounds studied. We are studying the use of three-dimensional Cartesian co-ordinates for atoms. To test this new representa-

tion we intend to study a series Chloroflurocarbons (CFCs) in an international comparative trial on the dataset. The data consists of $\approx$ 50 compounds. For each compound the most stable conformations are given, giving in total around 200 structures. We intend applying *Progol* to this database. The atom and bond predicates previously used are extended to by adding a predicate that computes the distance in 3-D space between stoms, and a predicate that computes the angle between any three atoms. If this approach proves successful, it eliminates the need to explicitly align compounds, since in effect, this is automatically accomplished by the induction procedure.

A distant extension to our SAR work is to use CoMFA type representations. This would be similar to the proposed representation using distance and angle predicates, but would involve a more regular coverage of space.

§8 Conclusions

The advantages ILP has demonstrated as a SAR method are its ability to deal efficiently with inductive problems concerning complicated molecular structures, and its ability to produce rules that are readily comprehensible by chemists and translatable into the language of normal chemical discourse.

Four application problems are reviewed in this paper: inhibition of dihydrofolate reductase by pyrimidines, inhibition of dihydrofolate reductase by triazines, design of tacrine analogues, and the mutagenesis prediction of nitroaromatic and heteroaromatic compounds. The reasons for this ordering of the problems is both historical reflects an increasing diversity of molecules. In our first work on pyrimidine analogous, the basic structural representation was a simple sequence. In the triazine analogues, the representation was a simple branched tree. For the mutagens, the compounds were represented as arbitrary graphs.

While the case studies in this paper represent almost four years of ILP research into chemical SAR problems, we are still far from establishing ILP as a standard SAR method in chemistry laboratories. However, we believe that these studies *have* fulfilled an important role: professional SAR chemists, with no direct Machine Learning have started to take an interest in the method. We have collaborated with Professor Harris (Institute of Molecular Medicine, University of Oxford), on improvements of suramin analogues. There is also active interest in ILP at the pharmaceutical companies Rhone-Poulenc Rorer and Pfizer.

At various stages of this research, it was suggested that propositional representation schemes are sufficient for all the problems considered. Our aim here was to show that while this is true for simple template-based structures, with arbitrarily complex chemical structures and background knowledge, propositional representations become unmanageable. Admittedly, our study has been limited to fairly simple encodings for propositional learners that use fixed-length attribute lists. A class of propositional learners that could poten-

tially be of interest is that which allow for propositional learning in infinite attribute spaces. For such learners, problems such as those in Section 6 should be tractable. In that problem, on average, a compound only has 9 different atom types, 3 different bond types, and 8 different charge constants. On average, therefore a learner capable of variable length attribute lists would require only $9 + 3 + 9^2 + 9^2 + 8 + 8 + 1 = 191$ attributes for each compound. While we are unaware of any implementations of such learners, we believe that they could potentially construct useful theories.

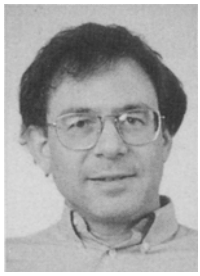
The advent of an efficient ILP program *Progol* that is capable of utilising non-determinate, non-ground background knowledge has opened up the possibility of analysing arbitrary chemical structures. With the generic representation scheme based on atoms and their bond connectivities, we believe that we are now in a position to employ fully the power of first-order learning to find principles that relate chemical activity to structure.

References

1) Bahler, D. and Bristol, D., "The Induction of Rules for Predicting Chemical Carcinogenesis," in *Proceedings of the 26th Hawaii International Conference on System Sciences*, Los Alamitos, IEEE Computer Society Press, 1993.
2) Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., *Classification and Regression Trees*, Wadsworth, Belmont, 1984.
3) Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Schusterman, A. J., and Hansch, C., "Structure-Activity Relationship of Mutagenic Aromatic and Heteroaromatic Nitro Compounds. Correlation with Molecular Orbital Energies and Hydrophobicity," *Journal of Medicinal Chemistry*, *34*, *2*, pp. 786-797, 1991.
4) Hansch, C., Malong, P. P., Fujita, T., and Muir, M., *Nature*, *194*, pp. 178-180, 1962.
5) Hirst, J. D., King, R. D., and Sternberg, M. J. E., "Quantitative Structure-Activity Relationships by Neural Networks and Inductive Logic Programming. ii. The Inhibition of Dihydrofolate Reductase by Triazines," *Journal of Computer-Aided Molecular Design*, *8*, pp. 421-432, 1994.
6) King, R., Muggleton, S., and Sternberg, M. J. E., "Drug Design by Machine Learning: The Use of Inductive Logic Programming to Model the Structure-Activity Relationships of Trimethoprim Analogues Binding to Dihydrofolate Reductase," *Proc. of the National Academy of Sciences*, *89*, *23*, pp. 11322-11326, 1992.
7) King, R. D., Muggleton, S., Srinivasan, A., Feng, C., Lewis, R. A., and Sternberg, M. J. E., "Drug Design Using Inductive Logic Programming," in *Proceedings of the 26th Hawaii International Conference on System Sciences*, Los Alamitos, IEEE Computer Society Press, 1993.
8) Marsh, P., "Prescribing All the Way to the Bank," *New Scientist*, 18 Nov. 1989.
9) Martin, Y. C., *Quantitative Drug Design: A Critical Introduction*, Marcel Drekker Inc., New York, 1978.
10) Muggleton, S., "Inverse Entailment and Progol," in *New Generation Computing*, *13*, Springer-Verlag and Ohmsha, Ltd., 1995.
11) Muggleton, S. H. and Feng, C., "Efficient Induction of Logic Programs," in *Proceedings of the First Conference on Algorithmic Learning Theory*, Tokyo, Ohmsha, 1990.
12) Ramsden, C. A., *Comprehensive Medicinal Chemistry (vol. 4)*, Pergamon Press, Oxford, 1979.

13) Srinivasan, A., Muggleton, S. H., King, R. D., and Sternberg, M. J. E., "Mutagenesis: ILP Experiments in a Non-Determinate Biological Domain," in *Proceedings of the Fourth International Inductive Logic Programming Workshop* (S. Wrobel, ed.), Gesellschaft fur Mathematik und Datenverarbeitung MBH, 1994. *GMD-Studien Nr 237*.

Ross D. King, Ph. D: He currently works as a postdoctoral fellow in the Biomolecular Modelling Laboratory of the Imperial Cancer Research Fund. His interests are in the application of Machine Learning to the understanding of the relationship between molecular structure and function. He received his B. Sc. in Microbiology from the University of Aberdeen in 1983, his M. Sc. in Computer Science from the University of Newcastle Upon Tyne in 1985, and his Ph. D. from the Turing Institute, University of Strathclyde in 1988.



Michael J. E. Sternberg, D. Phil.: He is head of the Biomolecular Modelling Laboratory at the Imperial Cancer Research Fund, London. He studied Natural Sciences at Cambridge obtaining a B. A. in 1972. He obtained an M. Sc. in Computing from Imperial College London in 1974. His research in computer modelling of biologcal molecules was his Ph. D. thesis research in Oxford (1974-78) and he has continued in this area.

Ashwin Srinivasan, Ph. D: He currently works as a postdoctoral reserchcr at the University of Oxford computing Laboratory. His interests are in developing ILP systems and applying them to difficult real-world problems in molecular biology and chemistry. He received his B. E. in Electrical Engineering and Computer Science from the University of New South Wales, where he also received his Ph. D in 1991.