Journal of Computer-Aided Molecular Design, 11 (1997) 571–580.
KLUWER/ESCOM
© 1997 *Kluwer Academic Publishers. Printed in The Netherlands.*

J-CAMD 415

The discovery of indicator variables for QSAR using inductive logic programming

Ross D. King[a,*] and Ashwin Srinivasan[b]

[a]*Department of Computer Science, The University of Wales Aberystwyth, Penglais, Aberystwyth, Ceredigion SY23 3DB, Wales, U.K.*
[b]*Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, U.K.*

Received 7 April 1997
Accepted 1 July 1997

Keywords: Artificial intelligence; Machine learning; Regression

Summary

A central problem in forming accurate regression equations in QSAR studies is the selection of appropriate descriptors for the compounds under study. We describe a novel procedure for using inductive logic programming (ILP) to discover new indicator variables (attributes) for QSAR problems, and show that these improve the accuracy of the derived regression equations. ILP techniques have previously been shown to work well on drug design problems where there is a large structural component or where clear comprehensible rules are required. However, ILP techniques have had the disadvantage of only being able to make qualitative predictions (e.g. active, inactive) and not to predict real numbers (regression). We unify ILP and linear regression techniques to give a QSAR method that has the strength of ILP at describing steric structure, with the familiarity and power of linear regression. We evaluated the utility of this new QSAR technique by examining the prediction of biological activity with and without the addition of new structural indicator variables formed by ILP. In three out of five datasets examined the addition of ILP variables produced statistically better results ($P < 0.01$) over the original description. The new ILP variables did not increase the overall complexity of the derived QSAR equations and added insight into possible mechanisms of action. We conclude that ILP can aid in the process of drug design.

Introduction

Many techniques have been used to describe compounds in quantitative structure–activity relationships (QSARs). Traditionally, QSAR studies have used descriptors based on 1-octanol–water partition coefficients to model the ‘hydrophobic effect’ and Hammett substituent constants to model electronic effects [1,2]. However, there has been no consensus on the best way to model steric interactions, and many different approaches have been suggested, e.g. topological descriptors [3,4], quantum mechanical descriptors [5], substructural units [6,7], molecular shape (MS [8]) and molecular fields (CoMFA [9]). One simple and successful way of modelling steric interactions is to use the ‘expert’ judgement of a computational chemist to identify the important structural features in a dataset, and to use these features as ‘indicator variables’ (Boolean attributes that are 1 when the structure is present in a compound and 0 when it is absent). This technique has been successfully applied to a large number of QSAR problems [5,10]. The advantage of this technique, when it works, is that it allows the use of simple regression techniques and forms equations that can be easily understood and interpreted. The technique’s main disadvantage is that it is not automatic, and its success or failure depends on the skill of the chemist at identifying indicator variables. In this paper we present an automatic method of deriving indicator variables for QSAR.

Inductive logic programming (ILP) algorithms are a class of machine learning algorithm that have been successfully applied to a number of SAR problems. Initial work was done using the program Golem [11] to form

*To whom correspondence should be addressed.

Program availability: The ILP programs Golem and Progol (implemented in C or Prolog) and the data used in this paper can be obtained by request from: Ashwin Srinivasan (Ashwin.Srinivasan@comlab.oxford.ac.uk); they are freely available to academics.

TABLE 1
CHARACTERISTICS OF DATASETS USED

Dataset	Number of examples	Number of descriptors
Pyrimidines	55	6
Triazines	186	11
MUT1	188	5
MUT2	20	5
CCA	36	2

CCA is the calcium-channel dataset, MUT1 is the mutagenesis dataset with 188 compounds, and MUT2 is the mutagenesis dataset with 20 compounds.

SARs for the inhibition of dihydrofolate reductase by pyrimidines [12,13] and triazines [14]. This work was extended by the development of the program Progol [15] and its adaptation for application to non-congeneric SAR problems. Progol has been successfully applied to predicting the mutagenicity of a series of structurally diverse nitro-aromatic compounds [16] and the prediction of carcinogenesis [17]. ILP SAR methods are designed to produce easily understandable rules (structural alerts) that provide insight into the chemical mechanisms. Most existing SAR methods describe chemical structure using *attributes* – general properties of objects. For example, in the traditional Hansch approach to QSARs [1,2] the attributes are properties such as LogP and π, which are global properties of the molecule or substituted group; in the CoMFA [9] approach to QSARs, the attributes are points in space which are global properties of the coordinate system used. Such descriptions are lists of attributes (technically tuples) representing points in N-dimensional space (a fundamental feature of such a representation is that the order of the attributes in the tuple is not important). This type of description is unsuited to representing the structure of compounds. For example, in CoMFA the use of attributes requires the imposition of a coordinate system (as attributes are general properties of objects – in this case the coordinate system) and the use of a large number of explicit points (as there is no way to implicitly define objects using attributes). A more efficient and general method of representing chemical structure is to use *relations*. In a relational description the basic elements are relations between objects (logic programs – a general form of compute program). This increased generality allows a more direct mapping from chemical steric structure to its representation. For example, bond relations could be defined between atoms, and atom relations between atom IDs. and the properties of atoms. In a CoMFA-like representation using relations, there would be no need to explicitly align the compounds, nor would it be necessary to use a large number of points to define the fields as they could be implicitly defined using a relation (logic program). Formally, the difference in descriptive language between attributes and relations corresponds to the difference between propositional and first-order predicate logic [18]. Existing learning methods for QSARs (linear regression, PLS, neural networks, genetic algorithms, etc.) are all based on propositional logic: ILP uses the more powerful representation language of predicate logic, equivalent to the ability to learn general computer programs for prediction. To illustrate the difference between attributes and relations, consider the following hypothesis: an active compound requires a double bond conjugated to an aromatic ring. Such a hypothesis could be directly discovered and represented by a relational QSAR 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 represented in an attribute-based language without specifically precoding the attribute ‘double bond conjugated with an aromatic ring’. ILP algorithms have been shown for many learning problems to generate more concise, understandable and accurate rules than those based on attributes.

The most similar approaches to ILP are CASE [6], MULTICASE [7] and related substructural methods, and the symbolic machine learning approaches of Bahler and Bristol [19] and Lee et al. [20]. These methods are based on describing compounds using attributes. The use of relation gives ILP a theoretical advantage in describing the steric properties of compounds. However, ILP methods have been previously restricted to qualitative predictions of activity (i.e. ‘high’, ‘low’, etc.), and have not been able to make quantitative predictions (i.e. predict real numbers, regression). This restriction is why we refer to the previous work using ILP as learning SARs. In this paper we show how ILP can be synergistically combined with existing regression techniques to produce a QSAR method that can learn relational features and make quantitative predictions.

TABLE 2
DESCRIPTION OF THE PROPERTIES USED TO CHARACTERIZE THE DATASET OF 2,4-DIAMINO-5-(SUBSTITUTED-BENZYL)PYRIMIDINES [10]

Variable	Description
Biological activity	Association constant to DHFR from MB1428 *E. coli*
$MR'_{3,5}$	Composite molar refractivity of substituents at positions 3 and 5
$MR'_{3,4}$	Composite molar refractivity of substituents at positions 3 and 4
MR'_{4}	Molar refractivity of substituent at position 4
$\Pi_{3,4,5}$	Composite hydrophobicity of substituents at positions 3, 4 and 5
$\Pi_{3,4}$	Composite hydrophobicity of substituents at positions 3 and 4
$\log_{10}(1.318 \times 10^{\Pi 3,4,5} + 1)$	Constructed to allow bilinear hydrophobic relationship

TABLE 3
DESCRIPTION OF THE PROPERTIES USED TO CHARACTERIZE THE DATASET OF 4,6-DIAMINO-1,2-DIHYDRO-2,2-DIMETHYL-1(X-PHENYL)-*s*-TRIAZINES [21]

Variable	Description
Biological activity	log(1/C), where C represents the molar concentration which produces 50% reversible inhibition of DHFR from L1210 mouse leukaemia cells and Walker 256 rat tumours
Π_3	Hydrophobicity of substituent at position 3
Π_4	Hydrophobicity of substituent at position 4
MR_3	Molar refractivity of substituent at position 3
MR_4	Molar refractivity of substituent at position 4
$\sigma_{3,4}$	Composite sigma effect of substituents at positions 3 and 4
I_1	Indicator variable: 1 if Walker's enzyme involved
I_2	Indicator variable: 1 if compound has an ortho-substitution
I_3	Indicator variable: 1 if compound has a rigid group to position 3
I_4	Indicator variable: 1 if compound has a rigid group to position 4
I_5	Indicator variable: 1 if compound has flexible bridges between phenyl rings
I_6	Indicator variable: 1 if compound has other (non-flexible) bridges

Materials and Methods

Data

Five QSAR datasets were used in this study; they are summarised in Table 1. The datasets concern the following: the inhibition of *E. coli* dihydrofolate reductase (DHFR) by 2,4-diamino-5-(substituted-benzyl)pyrimidine analogs [10]; the inhibition of mouse/rat tumour DHFR by 4,6-diamino-1,2-dihydro-2,2-dimethyl-1(X-phenyl)-*s*-triazine analogs [21]; the modulation of transmembrane calcium movement by methyl 2,5-dimethyl-4-[2-(phenylmethyl) benzoyl]-1*H*-pyrrole-3-carboxylate analogs [22]; and the mutagenicity of aromatic and heteroaromatic nitro compounds belonging to two disparate groups of 188 and 42 chemicals [5]. Of the 42 compounds in the latter set, actual activity values are only available for 20. These datasets were chosen as they have all been extensively studied by QSAR methods, and qualitative predictions of activity have been previously obtained using ILP algorithms on all except the modulation of transmembrane calcium movement. The basic descriptors (attributes/properties) used in the above previous studies of the datasets are listed in Tables 2–5. These descriptors consist of both standard QSAR descriptors (such as molar refractivity) and specific indicator variables formed by the authors to fit the particular datasets.

TABLE 4
DESCRIPTION OF THE BASIC PROPERTIES USED TO CHARACTERIZE THE DATASET OF METHYL 2,5-DIMETHYL-4-[2-(PHENYLMETHYL)BENZOYL]-1*H*-PYRROLE-3-CARBOXYLATE ANALOGS [22][a]

Variable	Description
Biological activity	Concentration needed to increase developed tension to 50% of the isoprenaline maximum in the 1 Hz paced guinea pig atria
CLOGP	Hydrophobicity of compound
CMR	Molar refractivity of compound

[a] The original study was carried out using CoMFA and no use is made of the CoMFA description of the compounds.

Algorithms used

The basic prediction method chosen was that of stepwise linear regression implemented in the SPSS package [23]. Any other robust regression method (method that predicts a real number) could have been employed, e.g. neural networks or genetic algorithms. However, stepwise linear regression is perhaps the simplest regression method, and produces easily understandable output.

Two ILP programs were used to extract structural constructs from the background knowledge. The program Golem [11] was used on the pyrimidine and triazine datasets; this is in keeping with earlier ILP experiments on these datasets and meant we could use exactly the same background knowledge as before [12–14]. Golem is incapable of reasoning with the type of background knowledge available for the calcium movement and the muta-

TABLE 5
DESCRIPTION OF THE PROPERTIES USED TO CHARACTERIZE THE DATASET OF MUTAGENICITY OF NITRO-AROMATIC COMPOUNDS [5]

Variable	Description
Biological activity	Mutagenesis using the Ames test with *S. typhimurium* TA98
LUMO	Energy level of lowest unoccupied molecular orbital in compound
log(P)	Hydrophobicity of compound
I_1	Indicator variable: 1 if compound contains three or more benzyl rings
I_a	Indicator variable: 1 if compound is an acenthrylene
$\log_{10}(10^{(\log(P)-5.48)}+1)$	Descriptor constricted to allow bilinear hydrophobic relationship

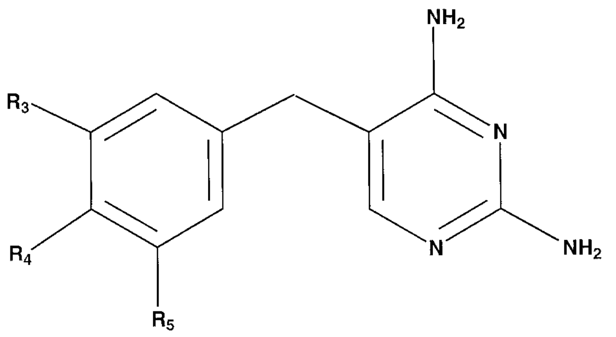

Fig. 1. Core structure of the 2,4-diamino-5-(substituted-benzyl)pyrimidine analog series.

genesis datasets. For these the program Progol [15] was used, also in keeping with earlier ILP experiments on these datasets [16].

ILP

In ILP all the inputs and outputs are logical statements in the computer language Prolog. Such statements are readily understandable as they closely resemble natural language. For any application the inputs to an ILP method are a set of positive examples (i.e. for SAR the active compounds), negative examples (i.e. non-active compounds) and background knowledge about the problem. The ILP method outputs a hypothesis expressed as a set of rules that predicts the positive and negative examples using the background knowledge. The hypothesis is found by a search through a partially ordered generalisation sublattice. Constraints on the search make it computationally feasible. The most general element in the lattice is the hypothesis that all compounds are active. This correctly predicts all the active compounds but makes mistakes on all the inactive compounds; the hypothesis is clearly too general. The most specific example is a hypothesis that either predicts one (Progol) or two (Golem) active compounds; such a hypothesis is probably too specific. The best hypothesis lies between the most general and most specific hypothesis; it is chosen by information compression. This 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 idea of using compression to evaluate hypotheses comes from the field of algorithmic information theory. This is based on quantifying Occam's razor, and using universal Turing machines to define simplicity. 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), i.e. it is a compromise between sensitivity and specificity. In Progol it can be shown that there is a guarantee that the most compressive hypothesis will be output. In Golem the search is stochastic and it is only probable that a highly compressive rule will be found. After generation of a single rule, the examples covered by the rule are removed from consideration and other rules are generated until all the examples are removed or no more compressive (statistically significant) rules can be found.

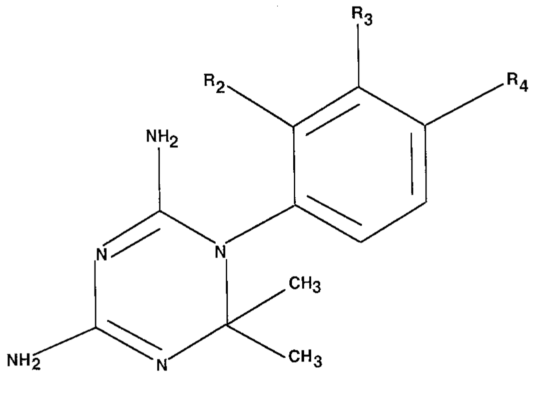

Fig. 2. Core structure of the 4,6-diamino-1,2-dihydro-2,2-dimethyl-1(X-phenyl)-*s*-triazine analog series.

Background knowledge for ILP

ILP systems use 'background knowledge' to describe problems. This background knowledge consists of logical statements about the data, e.g. the atomic structure of the chemicals involved.

In the pyrimidine dataset the existence of a template with only three possible substitution positions (Fig. 1) gives a relatively small structural component to the background knowledge. The background knowledge consists of statements that define the chemical structures substituted at each position. For example, a Prolog statement of the form

struc(drug_55, Cl, NH_2, CH_3)

represents the fact that drug 55 has a chlorine substituted at position 3, an amino group substituted at position 4, and a methyl at position 5. Also included in the background knowledge are the properties of the different chemical groups used. Full details of these predicates used can be found in Ref. 13.

In the triazine dataset the compounds can also be considered to have a common template structure (Fig. 2). However, the chemical groups substituted onto the template are more complicated, and 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. Each substituent can, in turn, itself contain a ring structure. In this case, further substitutions are possible into positions 3 and 4 of these rings. Several types of statement were used to encode the structure of the triazines. For example,

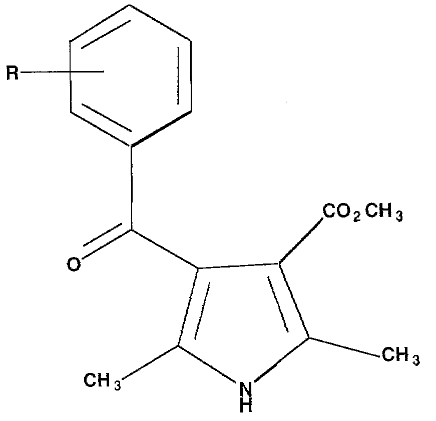

Fig. 3. Core structure of the methyl 2,5-dimethyl-4-[2-(phenylmethyl)benzoyl]-1*H*-pyrrole-3-carboxylate analog series.

struc3(drug_217, Cl, absent)
struc4(drug_217, $(CH_2)_4$, subst14)
subst(subst_14, SO_2F, Cl)

The relation 'struc3' represents substitutions at position 3 on the basic template: a chlorine is present and there is an absence of a further phenyl ring. The relation 'struc4' represents substitutions at position 4 on the basic template: there is a $(CH_2)_4$ bridge to a second phenyl ring (implicit in the representation). This second phenyl ring has an SO_2F group substituted at position 3 and a Cl group substituted at position 4. This is represented using the linker constant subst14 and the relation 'subst'. No substitutions at position 2 on the basic template were considered. As in the pyrimidines, the background knowledge also included basic properties of the substituent groups [14].

The remaining datasets (calcium movement and mutagenesis) were described using a general descriptive method that does not need the use of a template (Figs. 3 and 4). For mutagenesis, two related nitro-aromatic datasets were studied [5], one with 188 compounds that has been shown to be suitable for regression (MUT1), and one with 20 compounds that has proven difficult for regression (MUT2) [16]. The general descriptive method used for calcium movement and mutagenesis allows the compounds to be more heterogeneous than one based on templates; i.e. allow non-congeneric studies. The background knowledge used was the atom/bond structural descriptions of the molecules and the generic chemical knowledge of the structural groups. The former include the atom and bond structures in each molecule, along with typing information automatically obtained from the modelling program Quanta^{TM}. This results in the structure being represented by statements of the form

atom(drug_127, 127_1, carbon, 22, 0.191)
bond(drug_127, 127_1, 127_6, aromatic)

which state that in compound 127, atom number 1 is a carbon atom of Quanta type 22 (aromatic carbon in a six-membered carbon ring) with a partial charge of 0.191, and atoms 1 and 6 are connected by an aromatic bond. The generic structural definitions used provide definitions of methyl groups, nitro groups, aromatic rings, heteroaromatic rings, connected rings, ring length, and the three distinct topological ways to connect three benzene rings. Complete listings of these definitions can be found in Srinivasan et al. [25]. A more complete set of generic descriptors has been developed for use on other problems and is available on request.

Methodology

The basic methodology is designed to test whether ILP methods can be used to form indicator variables for regression equations in QSAR that give improved results over conventional approaches. This is done as follows:

(1) Form stepwise linear regression equations using the descriptors (attributes/properties) employed by the original authors (Tables 2–5). These descriptors consisted of both standard QSAR properties (e.g. molar refractivity) and specific hand-crafted indicator variables.

(2) Use ILP to form new indicator variables for the data. Then do stepwise linear regression using the original descriptors and the new ILP indicator variables.

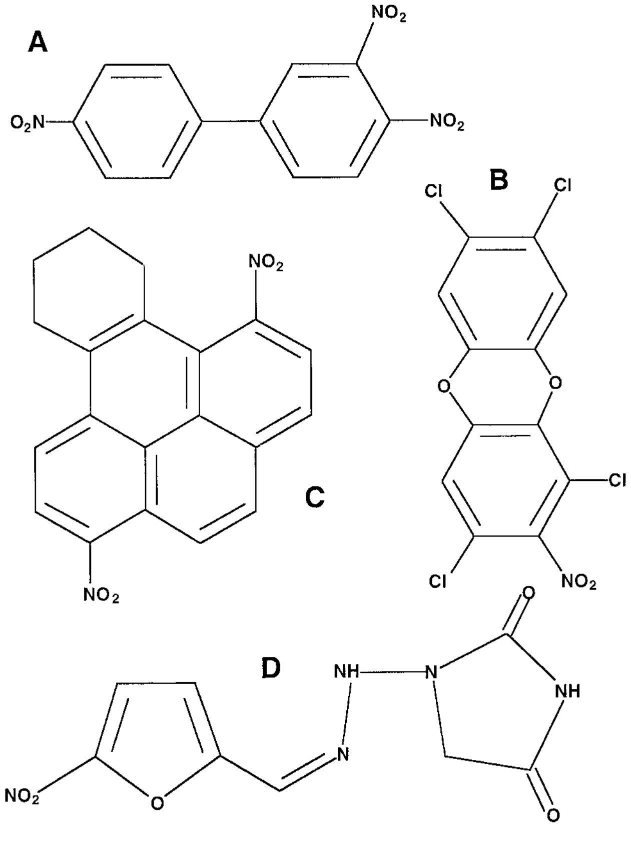

Fig. 4. Examples of the diverse dataset of nitro aromatic and heteroaromatic compounds.

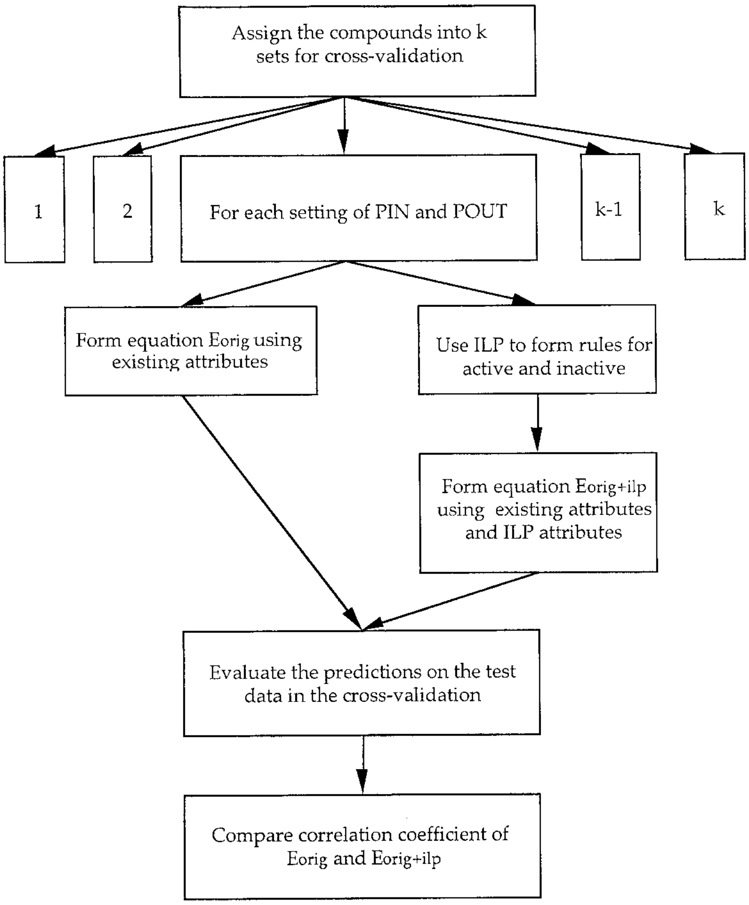
Assign the compounds into k sets for cross-validation
1
2
For each setting of PIN and POUT
k-1
k
Form equation Eorig using existing attributes
Use ILP to form rules for active and inactive
Form equation Eorig+ilp using existing attributes and ILP attributes
Evaluate the predictions on the test data in the cross-validation
Compare correlation coefficient of Eorig and Eorig+ilp

Fig. 5. Overview of the methodology used to test the utility of forming indicator variables using ILP.

(3) Test if there is a statistically significant improvement in the regression equations formed using ILP indicator variables compared to those using the original descriptors.

To test for a significant difference between the two regression equations, we adopted a two-stage approach based on model selection and validation. This was designed to avoid the danger that chance effects would allow the addition of extra ILP indicator variables fortuitously to improve the regression equations (see Topliss and Edwards [26], although details in the paper are flawed by the use of a bad random number generator). In the model selection stage we find 'good' significance levels for the stepwise linear regression. We did this using cross-validation to test the predictions of the regression equations formed with and without using ILP indicator variables. The use of cross-validation for parameter estimation was preferred over distribution-based tests, such as the F-test, as it directly considers predictive performance [27]. Validation of the models was then done for each of the datasets where cross-validation showed that there was a difference in accuracy on the addition of ILP indicator variables. Cross-validation could not be used to directly do this, because repeated reuse of the sample within the procedure could introduce dependencies that may violate assumptions of distribution-based tests of significance. We therefore used a classical statistical procedure for the final assessment.

Model selection

The following k-fold cross-validation design is adopted to explore for the best datasets and most reasonable significance levels; see Fig. 5.

(1) Randomly assign the compounds to k (approximately) equal partitions. Each partition will in turn be withheld to form a 'test' set. The compounds in the other partitions will provide the 'training' data for constructing the equation for predicting the activity of a compound. The value of k depends on the dataset. For pyrimidines $k=5$, for triazine and calcium-channel agonists $k=6$, and for mutagenesis with 188 compounds $k=10$. These values are in keeping with other experiments that have been performed on these datasets both with and without ILP. No cross-validation is needed for mutagenesis with 20 compounds (MUT2), as no significant equation could be formed using the original descriptors (the indicator variable discovered by Progol had already been shown using cross-validation to be significant [16]).

(2) Test different settings of significance for inclusion

TABLE 6
RESULTS OF THE EXPLORATORY CROSS-VALIDATION

Dataset	(PIN,POUT) = (0.05,0.10)		(PIN,POUT) = (0.01,0.05)	
	R_{orig}	$R_{orig+ilp}$	R_{orig}	$R_{orig+ilp}$
Pyrimidines	**0.77**	**0.83**	0.77	0.77
Triazines	0.63	0.64	0.64	0.64
CCA	0.80	0.81	**0.80**	**0.87**
MUT1	0.89	0.89	0.89	0.89
MUT2	–	**0.64**	–	0.64

PIN and POUT are parameters in the stepwise linear regression. R_{orig} and $R_{orig+ilp}$ are the correlation of predicted to actual activity for equations formed with and without ILP indicator variables, respectively. In bold are the results which suggest that the addition of ILP indicator variables significantly improves accuracy. In the calcium movement dataset it was shown that the CoMFA methodology [9] can significantly improve activity predictions over the results for the basic descriptors shown. For the mutagenesis dataset with 20 compounds (MUT2), it was only possible to produce statistically significant regression equations with the ILP indicator variables; it is therefore unnecessary to perform cross-validation.

and exclusion of terms in the formation of equations. These parameters relate to the level of significance that has to be achieved above (or below) which a term is retained (or removed from) the model. There is no prescribed setting for these parameters (termed PIN and POUT, respectively, in SPSS). Experimental results in this paper explore two settings: $S_1 = (PIN,POUT) = (0.05,0.10)$ (default settings for SPSS) and $S_2 = (PIN,POUT) = (0.01, 0.05)$. These settings are reasonable, given that the settings refer to significance values.

(3a) Construct an equation E_{orig} relating activity to the original descriptors.

(3b) Divide the range of activities in the (training) dataset into two intervals (namely, 'active' and 'inactive'). (The given activity value 0 was chosen as the point of division into two intervals.) For each of the intervals, use an ILP system to obtain rules for predicting the active and inactive classes. These rules are then used to form new ILP indicator variables. Finally, construct an equation $E_{orig+ilp}$ relating activity to all the descriptors available (i.e. original and ILP indicator variables).

(4) For each setting S_i, form records of actual and predicted values of biological activity on the k test datasets.

(5) Calculate the correlation of predicted to actual activity for equations E_{orig} and $E_{orig+ilp}$; these are termed R_{orig} and $R_{orig+ilp}$.

Model validation

The following procedure was followed to test the utility of new ILP indicator variables in the datasets identified in the model selection (cross-validation) stage:

(1) Datasets for which there appears to be no difference between R_{orig} and $R_{orig+ilp}$ were removed from further consideration. For these datasets ILP indicator variables are assumed to contribute nothing and no further quantitative assessment was required.

(2) For datasets where there does appear to be a difference between R_{orig} and $R_{orig+ilp}$, find ILP indicator variables using *all* of the data.

(3) For each dataset, using the setting that yielded the highest correlation R_{orig} and $R_{orig+ilp}$, obtain an equation $E_{orig+ilp}$ using all available descriptors (i.e. original and ILP indicator variables).

(4) Partition the descriptors in $E_{orig+ilp}$ into a set containing only original descriptors and a set containing only ILP indicator variables. Assess the changes in goodness of fit, and determine if the change caused by the addition of the ILP indicator variables is significant using the partial F-test [23].

Results

The results of the model selection (exploratory cross-validation) stage are given in Table 6. These show that in the datasets of pyrimidines, calcium-channel agonists and mutagenesis with 20 compounds, improved QSAR equations were formed using the automatically generated ILP indicator variables. In the datasets of triazines and mutagenesis with 188 compounds, no improvement using ILP indicator variables was found. Experimentation suggests that this would continue to hold even for other settings of PIN and POUT; however, it is interesting that ILP indicator variables were selected by the stepwise linear regression.

The quantitative results of the model validation stage are given in Table 7. These show, as indicated by the exploratory cross-validatory stage, that for the datasets of pyrimidines, calcium-channel agonists and mutagenesis

TABLE 7
ASSESSMENT OF THE CONTRIBUTION MADE BY ILP INDICATOR VARIABLES

Dataset	r^2_{orig}	$r^2_{orig+ilp}$	r^2_{ch}	C_{ilp}	Significance
Pyrimidines	0.54 (0.53)	0.84 (0.82)	0.30	65	$P < 0.01$
CCA	0.69 (0.68)	0.84 (0.82)	0.15	48	$P < 0.01$
MUT2	–	0.40 (0.37)	0.40	40	$P < 0.01$

r^2_{orig} and $r^2_{orig+ilp}$ are the coefficients of multiple determination before and after inclusion of the ILP indicator variables that appear in the final equation (adjusted values are in parentheses); r^2_{ch} is the importance of the ILP indicator variables; C_{ilp} measures (as a percentage) the proportion of unexplained variation that the change in r^2 constitutes; and the final column is the result of testing that the true population value for a change in r^2 on adding the ILP indicator variables is 0.

TABLE 8
EQUATIONS OBTAINED IN THE QUANTITATIVE ASSESSMENT STAGE WHEN ILP INDICATOR VARIABLES AUGMENT EXISTING DESCRIPTORS

Dataset	Equation with ILP indicator variables	No. of terms in equation with ILP	No. of terms in original equation
PYR	$\text{Act.} = 1.68\text{MR}'_{3,5} - 1.66\text{ILPa} + 1.15\text{ILPb} - 0.70\text{ILPc} - 0.69\text{ILPd} + 6.03$	6	6
CCA	$\text{Act.} = 0.92\text{CLOGP} - 0.89\text{ILPe} + 2.84$	3	3
MUT2	$\text{Act.} = 1.84\text{ILPf} - 0.24$	2	–

The last column refers to the number of terms (including the constant) in the equations formed before and after the addition of the ILP indicator variables.

with 20 compounds, there are significant improvements in prediction accuracy in the QSAR equations by the addition of the ILP indicator variables compared to the use of the existing descriptors. The actual equations formed using the ILP indicator variables are given in Table 8. It can be seen that the equations formed are simple and easy to comprehend. The addition of ILP indicator variables has not made the regression equations any more complicated than those obtained using the existing descriptors, as stepwise regression has in all the cases removed other descriptors to make way for the addition of the new ILP indicator variables. It is interesting that none of the original indicator variables appear in Table 8; they have, in effect, been replaced by the ILP ones. In a statistical sense, the ILP indicator variables are therefore more informative than the original author's ones.

It is relatively straightforward to interpret the coefficients of indicator variables in linear QSAR equations as the magnitude of the coefficient indicates the contribution of the presence of the corresponding structural feature. The descriptions of the indicator variables used are in Table 9. These are direct translations of the six automatically derived indicator variables. In the pyridine dataset five new indicator variables were formed by ILP. As there are crystallographic studies of the complex formed between trimethoprim (2,4-diamino-5-(3,4,5-trimethoxybenzyl)pyrimidine) and DHFR from *E. coli* [28,29], it is possible to compare the QSAR models with the X-ray stereochemistry of interaction. ILP indicator variables ILPa, ILP0c and ILPd all disfavour activity and require no substitution at position 4. This is in keeping with the suggestion of Roth et al. [30] that an important role of the 4 position is to force the 3 and 5 positions away from the 4 position. The combination of indicator variables ILPb and ILPd shows an interesting non-linear effect, that the flexibility of the substituent group at position 3 is favoured until it reaches a level of 7, and then it is less favoured (as any group makes ILPb true and a group of flexibility level of at least 7 makes ILPb and ILPd true). This suggests that the group has to be flexible enough to reach a certain position on DHFR and that too much flexibility interferes with binding in some way.

In the calcium movement dataset the new indicator variable formed by ILP is the existence of a double bonded oxygen with a partial charge of at least –0.252 (the exact charge is an artefact of the use of Quanta); this indicator variable disfavours activity. An examination of the position of these atoms in the compounds shows that they occur in the carboxylate group of the template. This is of interest, as groups added to the template are modifying the properties of the template. This region of the template is far from the region postulated to be important by Davis et al. [22] using CoMFA; their best regression equation had an r^2 value of 0.86, which is only slightly more than the r^2 value of 0.84 obtained using the ILP indicator variables (Table 7). However, the ILP model is considerably simpler than that formed by CoMFA as it consists of only three descriptors.

In the mutagenesis dataset with 20 compounds, the new indicator variable formed by ILP is the existence of a double bond conjugated to a five-membered aromatic ring in Ref. 16. This indicator variable is a strong structural indicator of mutagenesis. It is possible to rationalise by the conjugated double bond stabilising the five-membered aromatic ring, and so enabling the compound to reach the target site unchanged.

TABLE 9
THE INDICATOR VARIABLES CONSTRUCTED USING ILP

Dataset	Feature	Description
PYR	ILPa	Compound has no substituent at position 4 and a substituent at position 3
	ILPb	Compound has no substituents at positions 4 and 5, and a substituent at position 3
	ILPc	Compound has no substituent at position 4 and the substituent at position 3 is a hydrogen donor of at least level 1
	ILPd	Compound has no substituents at positions 4 and 5, and the substituent at position 3 has a flexibility of at least level 7
CCA	ILPe	Compound has a double bonded oxygen with a partial charge of at least –0.252
MUT2	ILPf	Compound has a double bond conjugated to a five-membered aromatic ring via a carbon atom

Discussion

There are two components to deriving a QSAR: the choice of representation used to describe the chemical structure of the compounds, and the analysis algorithm employed. Advances in QSAR studies have been made by improving both aspects, e.g. the introduction of quantum mechanical descriptors [5] was an advance in representation, and the introduction of neural network learning methods [31,32] and genetic algorithms [33,34] was an advance in prediction methodology. However, both the choice of representation and the prediction algorithm are intimately linked. Every analysis algorithm works best with certain forms of data [35] as they make explicit or implicit assumptions about data – it is a case of 'horses for courses'. For example, it is not wise (and sometimes impossible) to use the molecular fields description of compounds in CoMFA with standard linear regression (the assumption of statistical independence of the input attributes would be grossly violated and inaccurate predictions made). The molecular fields description requires some form of statistical resampling [36], either explicitly e.g. in PLS or implicitly e.g. in neural networks. In general, the more appropriate a particular description of data, the less complicated the learning algorithm needed to produce accurate results on that data. Much effort has been expended in statistics and machine learning to find ways of improving descriptions of data [37,38]. The use of ILP in this paper is aimed at forming more appropriate descriptions of QSAR data.

The methodology described in this paper is different from that customarily used in employing sophisticated learning methods in QSAR. Instead of using the learning technique to compensate for a poor description of the compounds under study, as is normally done, for example with neural networks or genetic algorithms, ILP is used to transform the description of the compounds into a more suitable description. This has the same result of producing an accurate QSAR, but has the added advantage of providing a better and more understandable description of the problem. In the transformation, ILP uses the assumption that a logical relationship between a few of the descriptors is important; this tends to give ILP the ability to form simpler and therefore more easily understood new descriptors. The most similar previous approach to that taken in this paper is that employed by CASE [6] and MULTICASE [7]. In this work an attribute-based algorithm specific to QSAR is used to identify structurally based indicator variables; these variables can then be used in forming a regression equation. The advantage of using ILP over CASE or MULTICASE is the greater generality of ILP as it is based on using relations and not attributes. The use of relations means that, in principle, there will always be indicator variables that can be found using ILP that cannot be found using an attribute-based method. However, it is still an open empirical question how far the theoretical advantages of ILP carry over to the practice of forming QSARS [39]. A recent illustration of the generality of the ILP approach in drug design is the extension of the method to forming 3D QSARs. This was achieved by simply adding three-dimensional coordinate information and trigonometry to the background knowledge and with no change to the basic Progol algorithm [40].

Conclusions

We have described a new two-step approach to forming accurate QSARs. ILP was used to improve the representation used to describe the chemical structure of the compounds under study. This was done by identifying useful indicator variables (attributes) for the compounds. Then standard linear regression (or any other regression method, e.g. PLS, neural networks, genetic algorithms) is used to form a QSAR. The role of ILP was to transform the original description of the compounds into one that is more appropriate. This new QSAR method synergistically combines the strength of ILP at using relational descriptors of compounds, while allowing familiar prediction techniques to be employed. It also avoids the difficulties ILP had in predicting real numbers (regression), and its necessity to use logical rules. The utility of the approach was assessed by examining the prediction of biological activity using standard linear regression with and without the addition of new structural indicator variables formed by ILP. In three out of the five datasets examined, the addition of ILP variables produced statistically better results ($P < 0.01$). The new ILP variables were found not to increase the complexity of the final QSAR equations and to give possible insight into the modes of action of the compounds.

The QSAR approach described in this paper is very general. In principle, it could be applied to almost any type data where, for whatever reason, the original description is considered inadequate and standard prediction methods did not employ it efficiently.

Acknowledgements

This research was supported by the EPSRC project 'Experimental Application and Development of ILP'. R.D.K. was supported by the Imperial Cancer Research Fund. The calcium movement dataset was provided by Darren R. Flower of ASTRA. Finally, thanks are due to Donald Michie for his guidance and advice, to Steve Moyle for the generous use of his personal computer, and to Penny Moyle for bringing to our attention facilities within the SPSS package for performing the statistical tests used here.

References

1 Hansch, C., Maloney, P.P., Fujita, T. and Muir, R.M., Nature, 194 (1962) 178.

2 Martin, Y.C., Quantitative Drug Design: A Critical Introduction, Marcel Dekker, New York, NY, U.S.A., 1978.

3 Ramsden, C. (Ed.) Comprehensive Medicinal Chemistry, Vol. 4, Pergamon, Oxford, U.K., 1990.

4 Trinajstic, N., Chemical Graph Theory, CRC Press, Boca Raton, FL, U.S.A., 1983.

5 Debnath, A.K., Lopez de Compadre, R.L., Debnath, G., Shusterman, A.J. and Hansch, C., J. Med. Chem., 34 (1991) 786.

6 Klopman, G., J. Am. Chem. Soc., 106 (1984) 7315.

7 Klopman, G., Quant. Struct.–Act. Relatsh., 11 (1992) 176.

8 Hopfinger, A.J., J. Am. Chem. Soc., 102 (1980) 7196.

9 Cramer, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.

10 Hansch, C., Li, R.-I., Blaney, J.M. and Langridge, R., J. Med. Chem., 25 (1982) 777.

11 Muggleton, S. and Feng, C., In Proceedings of the First Conference on Algorithmic Learning Theory, Japanese Society of Artificial Intelligence, Tokyo, Japan, 1990, pp. 368–381.

12 King, R.D., Muggleton, S., Lewis, R.A. and Sternberg, M.J.E., Proc. Natl. Acad. Sci. USA, 89 (1992) 11322.

13 Hirst, J.D., King, R.D. and Sternberg, M.J.E., J. Comput.-Aided Mol. Design, 8 (1994) 405.

14 Hirst, J.D., King, R.D. and Sternberg, M.J.E., J. Comput.-Aided Mol. Design, 8 (1994) 421.

15 Muggleton, S.H., New Gen. Comput., 13 (1995) 245.

16 King, R.D., Muggleton, S.H., Srinivasan, A. and Sternberg, M.J.E., Proc. Natl. Acad. Sci. USA, 93 (1996) 438.

17 King, R.D. and Srinivasan, A., Environ. Health Perspect., 104 (Suppl. 5) (1996) 1031.

18 DeLong, H., A Profile of Mathematical Logic, Addison-Wesley, Reading, MA, U.S.A., 1970.

19 Bahler, D. and Bristol, D.W., In Intelligent Systems for Molecular Biology-93, AAI/MIT Press, Menlo Park, CA, U.S.A., 1993.

20 Lee, Y., Buchanan, B.G., Mattison, D.M., Klopman, G. and Rosenkranz, H.S., Mutat. Res., 328 (1995) 127.

21 Silipo, C. and Hansch, C., J. Am. Chem. Soc., 97 (1975) 6849.

22 Davis, A.M., Gensmantel, N.P., Johansson, E. and Marriott, D.P., J. Med. Chem., 37 (1994) 963.

23 Norusis, M.J., SPSS: Base System User Guide, Release 6.0, SPSS Inc., Chicago, IL, U.S.A., 1994.

24 Wallace, C.S. and Freeman, P.R., J. R. Statist. Soc., B49 (1987) 195.

25 Srinivasan, A., Muggleton, S.H., Sternberg, M.J.E. and King, R.D., A.I. Journal, (1997) in press.

26 Topliss, J. and Edwards, R.P., J. Med. Chem., 22 (1979) 1238.

27 Wold, S., Technometrics, 20 (1978) 397.

28 Champness, J.N., Stammers, D.K. and Beddell, C.R., FEBS Lett., 199 (1986) 61.

29 Matthews, D.A., Bolin, J.T., Burridge, J.M., Filman, D.J., Volz, K.W., Kaufman, B.T., Beddell, C.R., Champness, J.N., Stammers, D.K. and Kraut, J., J. Biol. Chem., 260 (1985) 381.

30 Roth, B., Aig, E., Rauckman, B.S., Srelitz, J.Z., Phillips, A.P., Ferone, R., Bushby, S.R.M. and Siegel, C.W., J. Med. Chem., 24 (1981) 933.

31 Andrea, T.A. and Kalayeh, H., J. Med. Chem., 34 (1991) 2824.

32 So, S.-S. and Richards, W.G., J. Med. Chem., 35 (1992) 3201.

33 Kubinyi, H., Quant. Struct.–Act. Relatsh., 13 (1994) 285.

34 Glen, R.A. and Payne, A.W.R., J. Comput.-Aided Mol. Design, 9 (1995) 181.

35 Michie, D., Spiegelhalter, D.J. and Taylor, C.C., Machine Learning, Neural and Statistical Classification, Ellis Horwood, London, U.K., 1994.

36 Frank, I.E. and Friedman, J.H., Technometrics, 35 (1993) 109.

37 Michalski, R.S., In Michalski, R.S., Carbonnel, J. and Mitchell, T. (Eds.) Machine Learning: An Artificial Approach, Morgan Kaufmann, Los Altos, CA, U.S.A., 1986, pp. 83–134.

38 Lavrac, N. and Dzeroski, S., Inductive Logic Programming Techniques and Applications, Ellis Horwood, London, U.K., 1994.

39 King, R.D., Srinivasan, A. and Sternberg, M.J.E., New Gen. Comput., 13 (1995) 411.

40 Muggleton, S., Page, D. and Srinivasan, A., In Inductive Logic Programming 96, Stockholm, Sweden, 1996.