A Novel Logic-Based Approach for Quantitative Toxicology Prediction

Ata Amini,[†] Stephen H. Muggleton,[‡] Huma Lodhi,[‡] and Michael J. E. Sternberg*,[†]

Structural Bioinformatics Group, Centre for Bioinformatics, Division of Molecular Biosciences, and
Computational Bioinformatics Laboratory, Department of Computing, Imperial College London,
London SW7 2AZ, U.K.

There is a pressing need for accurate in silico methods to predict the toxicity of molecules that are being introduced into the environment or are being developed into new pharmaceuticals. Predictive toxicology is in the realm of structure activity relationships (SAR), and many approaches have been used to derive such SAR. Previous work has shown that inductive logic programming (ILP) is a powerful approach that circumvents several major difficulties, such as molecular superposition, faced by some other SAR methods. The ILP approach reasons with chemical substructures within a relational framework and yields chemically understandable rules. Here, we report a general new approach, support vector inductive logic programming (SVILP), which extends the essentially *qualitative* ILP-based SAR to *quantitative* modeling. First, ILP is used to learn rules, the predictions of which are then used within a novel kernel to derive a support-vector generalization model. For a highly heterogeneous dataset of 576 molecules with known fathead minnow fish toxicity, the cross-validated correlation coefficients (R^2_{CV}) from a chemical descriptor method (CHEM) and SVILP are 0.52 and 0.66, respectively. The ILP, CHEM, and SVILP approaches correctly predict 55, 58, and 73%, respectively, of toxic molecules. In a set of 165 unseen molecules, the R^2 values from the commercial software TOPKAT and SVILP are 0.26 and 0.57, respectively. In all calculations, SVILP showed significant improvements in comparison with the other methods. The SVILP approach has a major advantage in that it uses ILP automatically and consistently to derive rules, mostly novel, describing fragments that are toxicity alerts. The SVILP is a general machine-learning approach and has the potential of tackling many problems relevant to chemoinformatics including in silico drug design.

INTRODUCTION

With more than 70 000 chemicals in use today and many more being synthesized, it is vital that there are effective methods to assess the effect of these compounds on the environment and on human health.[1,2] In the development of pharmaceuticals, many potential leads are dropped due to their toxicity after millions of dollars have been invested.[3] Experimental testing is both time-consuming and expensive, and accordingly, there is a pressing requirement for accurate in silico methods to provide an initial screen that generates alerts for toxicity.[4] Often the strategy to develop these predictors follows a more general approach to derive qualitative/quantitative structure activity relationships (SAR).[5] Thus many of the toxicity prediction methods are based on regression from chemical properties, advanced machine learning, or expert-derived rule-based systems.[6,7] One particular machine-learning approach that has successfully been used for toxicity prediction is based on inductive logic programming (ILP).[8] However a major limitation of ILP is that the resultant logic rules yield yes or no predictions, and thus the capacity for quantitative prediction is limited. Here we report the development of a new *quantitative* SAR method (SVILP)[9] which combines ILP with support vector (SV) programming. Using a recently available dataset of

toxicity DSSTox,[10] which provides the toxicities of 576 chemicals for fathead minnow, we show that SVILP yields significantly better accuracies than ILP, regression from chemical descriptors, and an industry standard method TOPKAT.[11] We also compare the SVILP results with two more studies which have investigated the prediction of toxicity for fathead minnow.

The toxicities of several classes of chemicals can, as a first approximation, be estimated from their hydrophobicities (using the logarithm of the octanol/water partition coefficient LOGP) and their electrophilicities (using the lowest-unoccupied molecular orbital, LUMO, energies and dipole moments).[12–14] There are also expert systems available which are used for prediction of toxicities predominantly in pharmaceutical companies, universities, and government agencies:[15] for example, HazardExpert, which predicts the toxicity according to the dissociation constant (pK_a), LOGP, and structure of compounds and DEREK (deductive estimation of risk from existing knowledge),[16] which is based on toxicophores (chemical substructures with known toxic effect) and physicochemical properties. TOPKAT (toxicity prediction by computer-assisted technology) is a well-known software for toxicity prediction that uses the quantitative SAR model of prediction derived using statistical methods such as linear regression of structural descriptors.[11]

ILP has been used to derive SAR in several systems,[17–20] and it has been shown that the logic-based approach remedies many of the problems associated with many other SAR

* To whom correspondence should be addressed. Phone: +44 (0)20 7594 5212. Fax: +44 (0)20 7594 5264. E-mail: m.sternberg@imperial.ac.uk.
† Structural Bioinformatics Group.
‡ Computational Bioinformatics Laboratory.

TOXICOLOGY PREDICTION USING LOGIC-BASED APPROACH

J. Chem. Inf. Model., Vol. 47, No. 3, 2007 **999**
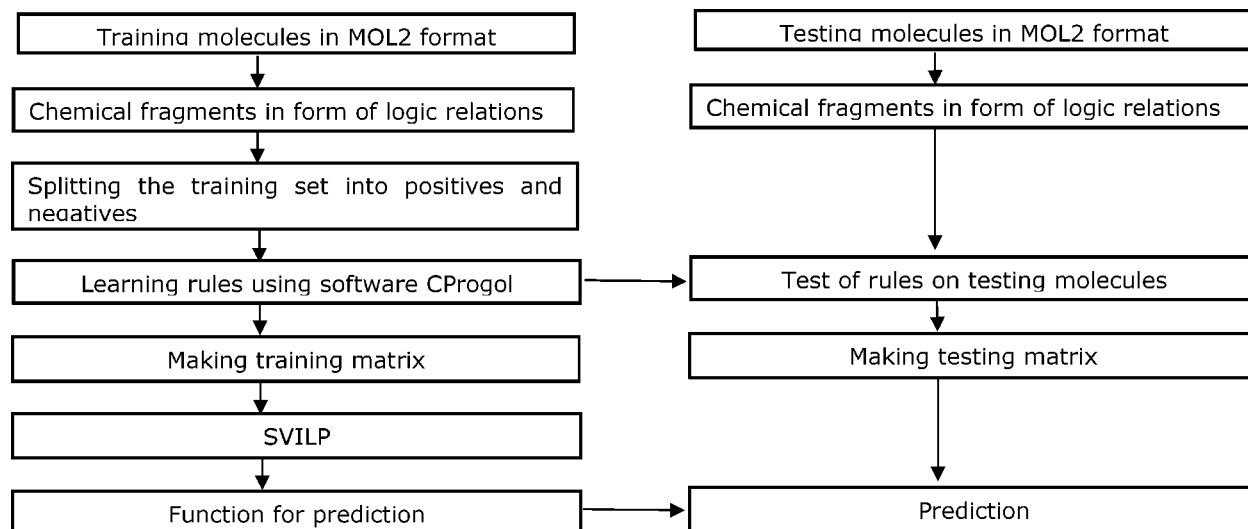
Figure 1. Process of construction and selection of logic rules by CProgol using the ILP system, followed by quantification of the logic rules using SVILP and testing on the unseen molecules.

methods including the need for molecular superposition as in CoMFA, problems of handling diverse data sets and the lack of chemical insights in learned SAR. In the ILP approach, there is no need to define all of the chemical features as in many other methods, since new features can be learned by the use of logic rules. One encodes that atom A is bonded to atom B, and atom B is bonded to atom C. The program can infer that atom A is connected to atom C via atom B without this having to be explicitly encoded. Importantly, the learned logic rules are readily amenable to interpretation as chemical substructures related to activity and thereby provide extensive chemical insights.

ILP is essentially a qualitative method yet the aim is to generate a quantitative toxicity predictor and, more generally, to generate quantitative SAR. A step toward generating quantitative SAR using ILP-based SAR was introduced by King and co-workers.[21] By inspection, they identified a few ILP-derived rules which were used as features in a linear regression to obtain a *quantitative* SAR.[21] However, their approach did not encapsulate the diversity of logic-learned rules and therefore would be difficult to apply to a diverse dataset, such as the one required for toxicity prediction. In this study, we report a novel approach which quantifies logic rules learned by ILP method using the SVILP method. This approach benefits from often-observed improved accuracy of the support vector methodology over linear regression. Moreover it can include a large number of logic-derived features and therefore is applicable to diverse datasets.

METHODS

The SVILP approach[9] uses ILP for learning logic rules, followed by quantitative modeling based on support vector technology as shown in Figure 1. The first step is to prepare the background knowledge, namely, the chemical fragments in the form of logic relations. The logic relations identify the chemical fragments according to the atom and bond details of the MOL2 structures. For example, we define a hydroxyl group as a combination of an oxygen atom and a hydrogen atom that are single-bonded in the form of a logic relation. In addition to background knowledge, the chemicals in the training set are classified into more toxic (positives)

and less toxic (negatives) according to the observed toxicities (see below for more details). ILP learning can now be conducted using the background knowledge and the observations. The software CProgol automatically learns logic rules using background knowledge and experimental observations. These learned rules form the input for quantitative prediction using the newly developed method SVILP. Furthermore, ILP provides extensive chemical insight into the cause of toxicity for each chemical. More details about each step are outlined in the following sections.

Dataset. The Distributed Structure-Searchable Toxicity (DSSTox) database from the U.S. Environmental Protection Agency (www.epa.gov; accessed Jan 30 2007) provides various databases including EPA Fathead Minnow Aquatic Toxicity Database (EPAFHM), which currently contains structures of 613 chemicals of which 576 have designated toxicity and therefore were used in this study. The set is expressed as highly diverse since it covers chemicals of various organic classes, that is, hydrocarbons, alcohols, aldehydes, esters, acids, etc., and furthermore, the molecules show many modes of action such as narcotics, oxidative phosphorylation uncouplers, respiratory inhibitors, electrophiles/proelectrophiles, acetylcholinesterase inhibitors, or central nervous system seizure agents. Supportiing Information Table SI-1 provides the structures of all of the chemicals. The toxicity end-points are based on the 96 h LC_{50} (mmol/L) value for the fathead minnow.[10] The LC_{50} is the aqueous concentration associated with 50% individual survival of a test population within a specified period.[13] The concept of toxicity in this manuscript is not general and is based entirely on the fathead minnow end-point.

The structures of chemicals in DSSTox are stored as SMILES strings, which we converted to 3D structures using CONCORD[22] via its implementation in the TRIPOS software. The 3D structures were then reoptimized using the PM3 semiempirical method,[23] and the structures were then saved as SYBYL Mol2 formats which provides all the necessary information for ILP (including atom types, atomic hybridizations, partial charges, *x*, *y*, and *z* coordinates of all of the atoms, and types of chemical bonds). The program BioMedCache[24] was used to derive the three chemical

Table 1. Chemical Fragments Used in This Study to Construct the Logic Rules

atoms[a]	rings[b]	alkyl groups	functional groups[c]
O(SP3), N(SP3), N(tertiary), N(quaternary), N(ar), F(ar), F(nar), F, Cl(ar), Cl(nar), Cl, Br(ar), Br(nar), Br, I(ar), I(nar), I, hydrophobic_hydrogen, S, P, Sn, ic	phenyl, hetar6ring, hetnar6ring, hetnar5ring, hetnar5ring, cyclohexane	methyl, ethyl, propyl, butyl, big_alkyl, alkyl, *tert*-butyl, *iso*-butyl, *iso*-propyl,	aldehyde, ether, thioether, ester, carboxylic acid, NH₂, amide, ketone, alcohol, nitro, alkene, alkyne, conjugated alkene, cyanide, CCl₃, CCl₂, NH, N═N, C═N, distance, edg, ewg, positive charge, negative charge, polar, hydrophobic

a ar stands for aromatic (for example, N(ar) means an aromatic nitrogen, but Cl(ar) means a chlorine atom connected to an aromatic atom), and "nar" stands for nonaromatic; ic (isolating or hydrophobic carbons) are carbon atoms which are not double- or triple-bonded to a heteroatom; hydrophobic_hydrogen is a hydrogen which is bonded to an ic. *b* hetnar6ring stands for hetero-nonaromatic-6-membered-ring and similar definitions are used for other rings. *c* edg and ewg stands for e-donating-groups and e-withdrawing-groups, respectively.

descriptors used, that is, LUMO, LOGP,[25] and dipole moment. LOGP reflects the hydrophobicity of compounds, and the mechanism of toxicities of highly hydrophobic molecules is often based on their accumulation in the nonpolar lipid phase of the biomembranes. LUMO and dipole moment describe electrophilicities of compounds and are two of the common chemical descriptors used in literature.[26,27] In this study, we used only a single conformation of the structures since most of structures in the set are rigid and therefore are not critically dependent on conformational changes. Furthermore, the logic-based SAR calculates the distance between fragments considering ±1.0 Å as the range which reduces the dependency of the model on conformational changes.

Background Knowledge. The first step for ILP learning is to provide generic logical relations that define the chemical fragments of the molecules: the background knowledge. The structures of molecules are prepared in Mol2 format and thus provide all the knowledge we need for a logic-based SAR, that is, details of atoms and bonds. The formats of atoms and bonds as transformed into PROLOG format (the logic programming language) are atom (M,Label,Atom_type,-Hybrid,X,Y,Z,Charge) and bond(M,Label1,Label2,Bond_type), where M represents molecules (e.g., m1, m2, ...); Label represents the unique label for each atom in each molecule (e.g., a1, a2, ...) via an atom number; Atom_type represents the type of atoms (e.g., c, h, cl, ...); Hybrid represents the hybridization state of each atom (e.g., 3 for sp3, 2 for sp2, ...); X, Y, Z represents the *x*, *y*, *z* coordinates of each atom; Q represents the partial charge of each atom; and Bond_type represents the type of bond ("1" for single bonds, "2" for double bonds, "3" for triple bond, and "ar" for aromatic bonds).

The atom−bond information provides elementary knowledge about the chemicals involved in the set. We then define chemical fragments (e.g., phenyl ring, aldehyde, carboxylic acids) to use as the main features for the ILP calculations. These chemical substructures are defined as relations in the PROLOG language using the background knowledge (atom−bond data). For example, the hydroxyl group can be defined as the following logic relation: hydroxyl(M,A) = atom-(M,A,o,o3,3.21,0.01,1.21,-0.32), atom(M,B,h,h,4.01,0.01, 1.02,0.22), bond(M,A,B,1).

This means that any molecule (M) could have an OH group, if it has a sp3 oxygen and a hydrogen that are single bonded. Table 1 lists the name of the chemical fragments used in this study.

Classification of Molecules. The dataset of 576 molecules was randomly divided into five folds with 115−116 molecules in each fold. One fold is used as a testing set, and the

four other folds are for training. Calculations were repeated five times for a 5-fold cross-validation. The dataset was then partitioned into positive and negative examples based on their toxicities. Molecules are distributed in a range of pLC₅₀ [−log(LC₅₀)] between −3 (LC₅₀ = 918 mmol/L) to 6.4 (LC₅₀ = 4.2 × 10⁻⁷ mmol/L). All molecules above the mean value of toxicities in the training set are considered to be positive (more toxic), and the remaining are considered to be negative (less toxic). Therefore, the cutoff value is different for each fold.

Learning theories using PROGOL. ILP learns from known examples or observations (i.e., it employs the reasoning known as induction).[28] The observations, the background knowledge, and the resultant rules are expressed as first-order logic programs, such as "compound m21 contains oxygen atom". CProgol[29,30] is a state-of-the-art ILP system. CProgol's input consists of positive and negative examples which belong to more toxic and less toxic molecules, respectively, together with background knowledge The output of CProgol is a set of logic rules which describe the positive and negative examples using the information provided in the background knowledge. In CProgol, the first positive example is randomly selected, and on the basis of the background knowledge, hypotheses are constructed; then the hypothesis with maximum compression is selected as the results of search. Compression, C, for each clause is defined as

$$C = P[p - (n + l)]/p$$

where C, P, p, n, and l are compression, total number of positive examples, number of positive examples covering by the clause, number of negative examples covering by clause, and length of clause (the number of features in each rule), respectively. Compression is a suitable measure for finding those rules which have predictive power, and it avoids overly specific rules (i.e., long clauses). The calculation is continued on the next positive example, but the redundant examples relative to the previously learned rules are removed. One of the advantages of the logic-based method is that it both constructs and selects the hypotheses. The selection is based on the value of compression that is defined automatically for each rule. At the end of the ILP calculation, all of rules with positive compression are used for regression.

SVILP. Support vector inductive logic programming (SVILP) is at the intersection of two areas of machine learning, namely, support vector machines (SVMs) and inductive logic programming (ILP). It is a novel machine-learning approach which combines the dimensionality-

TOXICOLOGY PREDICTION USING LOGIC-BASED APPROACH

J. Chem. Inf. Model., Vol. 47, No. 3, 2007 **1001**
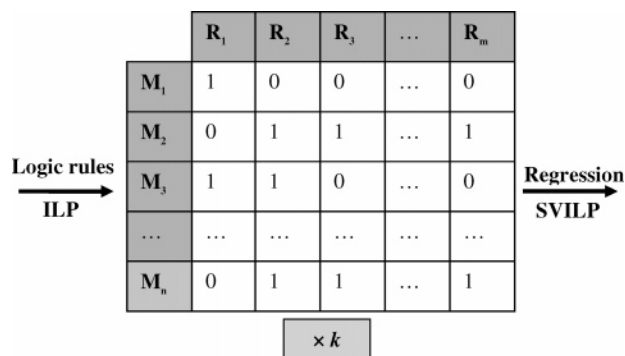
Figure 2. Support vector inductive logic programming (SVILP) for a system of n molecules and *m* learned rules: M_1, M_2, ..., M_n are the list of molecules; R_1, R_2, ..., R_m are the logic rules; the initial matrix is binary, "1" when it covers the molecule and "0" otherwise. The whole table is multiplied by a *k* factor.

independence advantages of SVMs with the expressive power and flexibility of ILP. In particular, we proposed a kernel[9] which is an inner product between two mapped examples. As with normal ILP, background knowledge and hypothesized clauses are encoded as logic programs. The approach we suggest differs from the existing relational kernels suggested in ref 31 by our use of logical background knowledge. The SVILP approach is a form of generalization relative to background knowledge, although the final combining function for the ILP-learned clauses in an SVM rather than by a logical conjunction. Figure 2 outlines the SVILP method. CProgol learns rules as we described in previous section. The learned logic rules are converted into a binary matrix (Figure 2). Each rule is tested for each molecule. If the rule covers the molecule, a number "1" is assigned, otherwise a "0" value is given. The whole matrix is multiplied by a *k* factor which is defined as

$$k = \frac{1}{\sqrt{m}}$$

where *m* is the number of rules used in the matrix. The training matrix is made by addition of the pLC_{50} of the molecules as the dependent variable to the first column of the matrix shown in Figure 2. The support vector machine method provided by SVMTorch (http://www.idiap.ch/machine_learning.php?content=Torch/en_SVMTorch.txt; accessed Jan 30 2007) is then used to make the model. The testing matrix is made using the same procedure, and the model is tested on this matrix for prediction. The results are presented using the squared linear correlation coefficient (R^2_{CV}) and the mean squared error (MSE) on the cross-validated data. The method has been reported in separate publications.[9] We chose the widely used approach partial least square (PLS) for comparison with SVILP. PLS is used when the number of variables exceeds the number of observations.

In SVILP and PLS, we require that the training set is further divided into a smaller training and a validating set. We use 25% of the training set for validation and the remaining 75% for making the validation model. For PLS calculations, we encoded a program using the algorithm described by Geladi and Kowalski.[32]

Chemical Descriptors. The toxicity of compounds can be modeled using various chemical descriptors such as LOGP, LUMO, and dipole moments.[11,26,27] To compare with this approach, we derived a (cross-validated) multilinear regression of pLC_{50} [$-\log(LC_{50} \text{ (mol/L)})$] with the above descriptors that we termed CHEM because it has been among chemical descriptors used in other studies.[26,27] These chemical features were calculated using the methods described previously.

TOPKAT. TOPKAT (toxicity prediction by computer-assisted technology), developed by Enslein et al.,[11] uses the quantitative SAR model of prediction including linear regression using the structural descriptors. The software accepts the structures of the molecules in the SMILES string, automatically splits the molecule into different fragments, and uses these fragments, as well as some chemical descriptors such as LOGP and shape index, molecular weight, and symmetry, for predictions. The program uses the above descriptors for quantitative toxicology modeling for over 18 endpoints. The Fathead minnow LC_{50} (version 3.2) was used as the model in this study, and the submodels were chosen by the program. TOPKAT validates its assessment by univariate analysis of the descriptors, multivariate analysis of the query structure in optimum prediction space, and finally similarity searching. To make a fair comparison of the above methods with the commercial software TOPKAT, we must ensure that we only consider predicted accuracies for molecules that were not included in the training data of either method. We therefore excluded any of the DSSTox molecules that TOPKAT had in its database leaving 165 unseen molecules.

Sign Test. The sign test compares the success of two methods under the null hypothesis that method 1 has the same chance of success as method 2. Random distributions of successes of a method follow a binomial distribution and the one tail provides the measure of significance.

McNemar Test. The McNemar test[33] was used to find the reliability of the classification methods. The McNemar test is a simple and standard approach for finding the statistical significance by evaluation of the probability of χ^2

$$\chi^2 = (b - c)^2/(b + c)$$

where *b* is the number of times that the prediction of the first method is wrong and the prediction of the second method (the case method) is correct and *c* is the number of times that the prediction of the first method is correct and the prediction of the second method is wrong. A prediction is significant if $\chi^2 < 0.05$.

RESULTS

The average accuracies of predictions over five folds using chemical descriptor method (CHEM), ILP rules in combination with PLS, and SVILP are given in Table 2a. The cross-validated square of correlation coefficients for CHEM, PLS, and SVILP are 0.52, 0.59, and 0.66, respectively. On the basis of the statistical sign test method, this improvement for SVILP and PLS is highly significant with respect to the CHEM method. The SVILP also shows significant improvement in comparison with PLS method. The numbers of features that are learned by ILP and used by SVILP and PLS are 1526, 1883, 1802, 1996, and 2095 for calculations 1−5, respectively

Table 2. (a) Accuracies of Quantitative Predictions by CHEM, PLS, and SVILP, (b) Results of Qualitative Classification of Molecules as Toxic or Nontoxic Using Three above Methods Plus ILP, and (c) Comparison of Results from Various Methods Introduced in This Study with TOPKAT

	(a) regression (N = 576)[a]			
	accuracy		significant improvement[g]	
	R^2_{CV}	MSE	CHEM	PLS
CHEM[d]	0.52	0.81		
PLS[e]	0.59	0.67	**0.005**	
SVILP[f]	0.66	0.57	***0.000001***	**0.02**

	(b) classification (N = 220)[b]			
	recall	significant improvement[h]		
	%	ILP	CHEM	PLS
ILP	55			
CHEM	58	**0.41**		
PLS	71	***0.00005***	***0.00005***	
SVILP	73	***0.00005***	***0.00005***	0.28

	(c) comparison with TOPKAT (N = 165)[c]			
	accuracy		significant improvement[g]	
	R^2	MSE	TOPKAT	CHEM
TOPKAT	0.26	2.2		
CHEM	0.48	1.04	**0.01**	
PLS	0.47	1.03	***0.001***	**0.02**
SVILP	0.57	0.8	***0.0001***	***0.0005***	***0.0001***

[a] The average of correlation coefficients on the five folds plus the mean square error (MSE) values for three methods of calculations. The significant (P < 0.05) improvements are shown in **bold**. The ***bold italic*** values are highly significant (P < 0.01) improvements. The numbers are the one-tail probabilities. N is the number of samples used for cross-validation. [b] The results of classification for the toxic class of molecules (pLC_{50} > mean) for four methods described in the text. [c] Comparison of toxicities of 165 molecules predicted by the different methods including the commercial software TOPKAT. [d] CHEM stands for chemical descriptors (LOGP, LUMO, and dipole moment). [e] Partial least square (PLS). [f] Support vector inductive logic programming (SVILP). [g] Using sign test. [h] Using McNemar method.

Table 3. Prediction of Toxic and Nontoxic Classes of Molecules by Methods Described in the Text

method	TP[a]	FN[b]	TN[c]	FP[d]	recall[e]	specificity[f]	accuracy[g]
CHEM	128	92	324	32	0.58	0.91	0.78
PLS	156	64	303	53	0.71	0.85	0.80
SVILP	161	59	310	46	0.73	0.87	0.82
ILP	121	99	313	43	0.55	0.88	0.75

[a] True positives. [b] False negatives. [c] True negative. [d] False positive. [e] TP/(TP + FN). [f] TN/(TN + FP). [g] (TP + TN)/(TP + TN + FP + FN).

In the second part of this study, the molecules were classified into two groups based on their toxicities: that is, toxic ($pLC_{50} \geq$ mean) and nontoxic ($pLC_{50} <$ mean), where "mean" is the average of toxicities of molecules in the training set. This study assesses a qualitative prediction of whether a molecule is toxic and also enables us to compare the accuracy of SVILP with ILP. The numbers of correct and incorrect predicted samples were found using three above-described methods, as well as the original ILP approach. The details of predicted values are given in Tables 2b and 3. According to Table 3, the SVILP has the largest recall, which means that more positives have been predicted correctly in comparison with other methods. The CHEM method is the best method among others regarding the

specificity, which means that less negative examples has been predicted as toxic by this approach, however, with no significant difference. But the total accuracy is in favor of SVILP based on the results of Table 3. Table 2b compares the recalls and shows significant improvements of the SVILP and PLS with respect to the CHEM method.

In the next study, the quality of the above methods was evaluated by comparison of the results with an extensively used software, TOPKAT. The SVILP, PLS, and ILP procedures were retrained using the 411 molecules of DSSTox database that are present in the databases of TOPKAT, and the remaining 165 molecules were used for testing. TOPKAT gives an error in prediction of toxicities of 33 molecules for various reasons such as absence of fragments and being outside of premium predictions space. Exclusion of these molecules improves the R^2 value for TOPKAT to 0.31. Because the change of accuracy resulting from the exclusion of 33 molecules is not significant, we have reported and compared the results based on all of 165 molecules in Table 2c. According to the results of Table 2c, the SVILP and TOPKAT methods have the highest and lowest accuracy of predictions, respectively. According to sign test, SVILP shows highly significant improvement in comparison with all of other approaches. The improvements for PLS in comparison with TOPKAT and CHEM and for CHEM in comparison with TOPKAT are significant. According to the results of Figure 3, TOPKAT is unable to discriminate between the toxic and nontoxic molecules; furthermore, some nontoxic molecules in top left corner of the Figure 3a have been predicted as very toxic, while in other two methods, we do not see the same problem. The advantage of SVILP in comparison with CHEM is in better prediction of more toxic molecules (top right corner of Figure 3b).

Chemical Insights. A major advantage of the logic-based approach is that the ILP-phase yields rules that provide chemical insights. The power of these rules is quantified by compression. The rules with high compression are expected to have a greater contribution to the predicted toxicity of compounds. Tables 4 and 5 show a few sample rules with high compression, and in Figures 4 and 5, these rules are shown on appropriate chemical structures. We have examined these rules and via a literature search assessed whether these features have previously been identified. Several of the rules had been previously identified as chemical alerts, but it is important that these are now discovered automatically and consistently using ILP. In addition, several new features have been identified. We now report some of these chemical alerts.

Hydrophobic Features. In our first series of rules, the molecule is toxic if it has hydrophobic features; however, this does not mean that all of molecules with hydrophobic features are toxic. These features have been found in different rules, and here, we summarize a few of the most important ones in Table 4. Some of the obvious forms of rules found in this study are a high value of LOGP, large number of hydrophobic hydrogen and hydrophobic carbon atoms, and two hydrophobic elements (carbon or hydrogen) in a range of 8−12 Å from each other. In Figure 4a, we show two highly hydrophobic molecules with a large number of hydrophobic elements. These rules emphasize the importance of having hydrophobic groups in increasing of toxicities which is a well-known phenomenon in the field of toxicol-

Toxicology Prediction using Logic-Based Approach

J. Chem. Inf. Model., Vol. 47, No. 3, 2007 **1003**

Figure 3. Comparison of prediction of a testing set of including 165 molecules using (a) TOPKAT and (b) SVILP. The lines inside the graphs separate the chemicals with high toxicity ($pLC_{50} > 0$) and the remaining with less toxicity. Note that the pLC_{50} of 0 is not the cutoff we used for rule generation.

Table 4. Sample Logic Rules with High Compression Expressed in English That Shows the Effects of Hydrophobicity on Toxicity[a]

rule	p	n	C	rule	ref
I	43	0	42	LOGP > 4	34
II	71	11	57	two hydrophobic elements (carbon or hydrogen) and distance between them is 10.7 ± 1.0 Å	34
III	40	6	33	number of hydrophobic hydrogen ≥ 20	34
IV	35	5	29	number of hydrophobic carbon ≥ 13	34
V	40	6	35	number of phenyl rings ≥ 2	35,36

[a] C is the compression; p and n are the number of positives and negatives, respectively, covered by the rule.

Table 5. Sample Logic Rules with High Compression Expressed in English That Shows the Electronic Effects on Toxicity[a]

rule	p	n	C	rule	ref
VI	32	2	27	a phenyl group and an electron-donating group and distance between them is 7.3 ± 1.0 Å	37
VII	41	6	32	a hydrophobic atom and an electron-withdrawing group and distance between them is 9.1 ± 1.0 Å	
VIII	27	2	22	an electron-donating group and an electron-withdrawing group and distance between them is 7.9 ± 1.0 Å	
IX	23	1	19	two chlorine atoms on an aromatic ring and distance between them is 3.0 ± 1.0 Å	36
X	28	4	21	a Cl−C−C−Cl group and an electron-donating group and distance between them is 3.2 ± 1.0 Å	38
XI	23	2	18	a alcohol group and a polar group and distance between them is 5.8 ± 1.0 Å	36
XII	17	0	14	a phenyl group and an electron-withdrawing group and distance between them is 5.7 ± 1.0 Å	
XIII	15	0	12	a phenyl group and an ester group and distance between them is 4.3 ± 1.0 Å	41, 42
XIV	15	0	14	number of chlorine atoms in an aromatic rings ≥ 2	36
XV	16	2	13	number of polar groups ≥ 6	
XVI	14	2	11	number of alkenes ≥ 2.	40
XVII	14	2	11	number of chlorine atoms ≥ 3	36

[a] C is the compression; p and n are the number of positives and negatives covered by the rule, respectively.

ogy. However, this is a new way to describe the importance of hydrophobicity and, in particular, the relational aspect of rules.[34] The exact value for the distance is obtained from the positive molecule used to generate the rule, and the range of ±1.0 Å is a parameter in the learning. One therefore interprets the distance with less precision (e.g., 10.5 ± 1.0 Å). In the literature, molecules with more than one phenyl ring are usually classified in the class of toxic compounds.[35] A clear example is the polyhalogenated aromatic compounds which are a group of the high contaminants distributed in variety of environments.[36] Most of molecules with more than a phenyl ring in the DSSTox dataset are very toxic; however, as we shown in Figure 4b, there are molecules with two phenyl rings and toxicities less than the cutoff value (LC_{50} about 10 mg/L) but not far away from this cutoff value.

Electronic Features. In the selected rules of Table 5, the electronic effects of chemical groups on toxicities are seen in different forms such as electron-donating groups (e.g., SP3 oxygen and alkyl groups), electron-withdrawing groups (such as halogen atoms and functional groups with carbonyl), and polar groups (Figure 5a−f). Electron-donating groups such as SP3 oxygen are the source of electrons and could alter toxicities of molecules.[37] 1-(2-Nitrophenoxy)-4-chlorobenzene (Figure 5d) is an example of these molecules. According to rule **VI**, a molecule is toxic if there is a phenyl ring and

an electron-donating group with the distance between them being 7.3 ± 1.0 Å (Figure 5a). The combination of a hydrophobic carbon atom with an electron-withdrawing group was found as a logic rule in **VII** and is shown in Figure 5b. A molecule with an electron-withdrawing group and an electron-donating group 7.9 ± 1.0 Å apart from each other could be toxic according to rule **VIII** (Figure 5c). We define these rules as novel features learned by ILP since we have not found similar features in the literature. For majority of rules, the distances between the chemical fragments are also defined, thereby identifying the relative location of the

Figure 4. Graphical representation of logic rules in Table 2: (a) positives (more toxic molecules) and (b) negatives (less toxic molecules). The arrows show the hydrophobic features

groups. For example, rule **XI** in Table 5 identifies the chlorine atoms in the 2 and 6 positions with respect to alcoholic groups as the most important, which is consistent with observed toxicities[36] (see Figure 5e). Rules **VII**, **VIII**, **IX**, **X**, **XI**, **XIV**, **XV**, and **XVII** directly or indirectly emphasize the importance of chlorine atoms in increasing toxicity. Chlorinated compounds have been used for many purposes in industry, for example, solvents, propellants, polymers, and pesticides. Such chlorinated compounds show toxicity, particularly in aromatic compounds.[38] Rules **IX** and **XIV** clearly identify the chlorine atoms in aromatic rings. The studies have shown that the biological activities of polychlorinated biphenyls are highly dependent upon the number of chlorine atoms, as well as on the substitution pattern.[36] Pentachlorophenol, for example, is used as preservative against fungus decay or termites, and this compound can be absorbed via the skin, lung, and gastrointestinal tract and is one of the known members of fungicides (pesticides).[39] The mechanism of action of this molecule is

through an increase in oxidative metabolism from uncoupling of oxidative phosphorylation.

Molecules including the alkene groups (C=C) could cause toxicity by different mechanisms, such as by formation of epoxides, which for example occurs in vinyl chloride. The epoxide could target DNA and RNA, but it may form aldehyde which is highly reactive against DNA and proteins. Alkadienes and, in particular, those that are halogenated are very well-known toxic chemicals (e.g., nephrotoxicity). Hexachlorobutadiene is a potent nephrotoxin with the kidney being the main target.[40] ILP has identified the effects of alkene and alkadiene groups in rule **XVI** (Figure 5f).

Rule **XIII** shows the effects of ester groups on toxicities (Figure 5c). Phthalates that include ester groups are described as the most abundant man-made environmental pollutants; they have been used extensively in consumer products and medical devices.[41] They can cause toxicity in many aspects, for example, by binding to estrogen receptors and inducing estrogen receptor-mediated responses.[42]

There are also many other features that had been learned and some of them are, to our knowledge, novel. For instance (as a known feature), organophosphorus compounds inhibit some enzymes, for example, cholinesterases in which the P=O (P=S) group binds to a serine group at the esteratic site of the cholinesterase enzyme followed by release of thiol or alcohol to produce the phosphorylated enzyme.[43] These compounds are used as pesticides and are rapidly absorbed through ingestion. We discovered rules which confirm these molecules are toxic. Thus, the logic rules are able to describe both the hydrophobicity of the molecules and the electronic effects.

Comparison with Other Methods. In the previous sections, we compared the SVILP with four methods: that

Figure 5. Graphical representation of logic rules in Table 3. The discovered chemical alerts are shown by arrows. Electron-withdrawing groups and electron-donating groups are shown by "ewg" and "edg", respectively.

TOXICOLOGY PREDICTION USING LOGIC-BASED APPROACH

J. Chem. Inf. Model., Vol. 47, No. 3, 2007 **1005**

is, ILP, CHEM, PLS, and TOPKAT. DSSTOX is a recently developed database. Previously, Kaiser and Niculescu[44] used a dataset of 865 chemicals to investigate the acute toxicity of the fathead minnow, and we expect there to be substantial similarities between their dataset (TerraTox)[45] and DSSTOX. They reported an average MSE of 0.96 after 5-fold cross-validation (from the model which they describe as the worst case scenario) which can be compared to the MSE of 0.57 that we obtained in this study using SVILP. Klopman et al.[46] also used a dataset of 658 molecules that we expect would also have large similarity with DSSTOX. They only report the results of their prediction on 78.6% of their dataset. We are therefore unable to compare our results with theirs. We note that typically 80% of chemicals in datasets such as DSSTOX have a very good correlation with LOGP and with simply the number of fragments, and accordingly, for these 80%, their toxicities can be predicted with high accuracy (typically >90%). Therefore the power of various predictors should be compared based on their accuracies on the 20% of the datasets that does not have a strong correlation with hydrophobicity.

DISCUSSION

We introduced a new quantitative logic-based method, support vector inductive logic programming (SVILP), which uses the logic-based technology to learn logic rules followed by regression. The method is based on a collection of logic rules, automatically learned by a well-studied system, inductive logic programming (ILP). ILP learns rules using the experimental observations and chemical structures. The support vector machine technology is then used for regression with logic rules as its features and observed toxicities as the dependent variable. The results of this study on a large, public, and diverse dataset show that SVILP predicts the toxicities with higher accuracy than other tested models. For the same number of features, SVILP showed significant improvement with respect to PLS; however SVILP and PLS use more features than CHEM. The CHEM method includes LOGP as the chemical descriptor, which has been the well-known feature of majority of toxicology predictors such as TOPKAT[11] and ECOSAR (www.epa.gov). One could interpret the higher accuracy of the SVILP and PLS as a consequence of using more features. However, it is suggested that more features do not necessarily improve the accuracy of prediction, and various feature selection methods have been designed to reduce the dimension and improve the predictions.[47] SVILP performs the feature selection after construction of hypotheses according to the compression values of the logic rules. The rules are chemically understandable and describe the chemical alerts which are the cause of activity/toxicity. The program automatically and consistently detects chemical substructures and properties by construction of rules which are general. The learned rules address features, some of which are exclusive for a particular mode of action such as narcotics and majority of which describe the system as a whole, not on the basis of their particular modes of action. The logic-based method can derive SAR for molecules in diverse dataset; no preliminary work such as superposition is necessary for this method, and it provides extensive chemical insight of activities/toxicities. We conclude that the SVILP approach provides both

accuracy of prediction combined with chemical insights and might have a wide range of applications in chemoinformatics.

Supporting Information Available: The chemical structures and their observed LC_{50} values in mg/L. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Mwense, M.; Wang, X. Z.; Buontempo, F. V.; Horan, N.; Young, A.; Osborn, D. Prediction of Noninteractive Mixture Toxicity of Organic Compounds Based on a Fuzzy Set Method. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1763−1773.

(2) Smith, D. A. Assessing Toxicology Quickly and Efficiently. *Chem. Eng.* **2000**, *107*, 125−126.

(3) Escher, B. I.; Hermens, J. L. M. Modes of Action in Ecotoxicology: Their Role in Body Burdens, Species Sensitivity, QSAR, and Mixture Effects. *Environ. Sci. Technol.* **2002**, *36*, 4201−4217.

(4) Klopman, G.; Chakravarti, S. K.; Zhu, H.; Ivanov, J. M.; Saiakhov, R. D. ESP: A Method to Predict Toxicity and Pharmacological Properties of Chemicals Using Multiple MCASE Databases. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 704−715.

(5) Smiesko, M.; Benfenati, E. Predictive Models for Aquatic Toxicity of Aldehydes Designed for Various Model Chemistries. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 976−984.

(6) Helma, C.; Kramer, S. A Survey of the Predictive Toxicology Challenge 2000−2001. *Bioinformatics* **2003**, *19*, 1179−1182.

(7) Dourson, M. L.; Teuschler, L. K.; Durkin, P. R.; Stiteler, W. M. Categorical Regression of Toxicity Data: a Case Study Using Aldicarb. *Regul. Toxicol. Pharmacol.* **1997**, *25*, 121−129.

(8) King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J. E. Drug Design by Machine Learning: The Use of Inductive Logic Programming to Model the Structure−Activity Relationships of Trimethoprim Analogs Binding to Dihydrofolate Reductase. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 11322−11326.

(9) Muggleton, S. H.; Lodhi, H.; Amini, A.; Sternberg, M. J. E. Support Vector Inductive Logic Programming. In *Discovery Science*, Proceedings of the 8th International Conference, DS 2005, Singapore, October 2005; Hoffmann, A., Motoda, H., Scheffer, T., Eds.; Springer: Berlin, 2005; pp 163−175.

(10) Richard, A. M.; Williams, C. R. Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network: A Proposal. *Mutat. Res.* **2002**, *499*, 27−52.

(11) Enslein, K.; Gombar, V. K.; Blake, B. W. Use of SAR in Computer-Assisted Prediction of Carcinogenicity and Mutagenicity of Chemicals by the TOPKAT Program. *Mutat. Res.* **1994**, *305*, 47−61.

(12) Ren, S. Modeling the Toxicity of Aromatic Compounds to *Tetrahymena pyriformis*: The Response Surface Methodology with Nonlinear Methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1679−1687.

(13) Verhaar, H. J. M.; Solbe, J.; Speksnijder, J.; van Leeuwen, C. J.; Hermens, J. L. M. Classifying Environmental Pollutants: Part 3. External Validation of the Classification System. *Chemosphere* **2000**, *40*, 875−883.

(14) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermister, D. E.; Drummond, R. A. Predicting Modes of Toxic Action from Chemical Structure: Acute Toxicity in the Fathead Minnow (*Pimephales promlas*). *Environ. Toxicol. Chem.* **1997**, *16*, 948−967.

(15) Benfenati, E.; Gini, G. Computational Predictive Programs (Expert Systems) in Toxicology. *Toxicology* **1997**, *119*, 213−225.

(16) Sanderson, D. M.; Earnshow, C. G. Computer Prediction of Possible Toxic Action from Chemical Structure; the DEREK System. *Hum. Exp. Toxicol.* **1991**, *10*, 261−273.

(17) King, R. D.; Muggleton, S. H.; Srinivasan, A.; Sternberg, M. J. E. Structure-Activity Relationships Derived by Machine Learning: The Use of Atoms and Their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 438−442.

(18) Finn, P.; Muggleton, S.; Page, D.; Srinivasan, A. Pharmacophore Discovery Using the Inductive Logic Programming System PROGOL. *Mach. Learn.* **1998**, *30*, 241−270.

1006 *J. Chem. Inf. Model., Vol. 47, No. 3, 2007*

AMINI ET AL.

(19) Sternberg, M. J. E.; Muggleton, S. H. Structure Activity Relationships (SAR) and Pharmacophore Discovery Using Inductive Logic Programming (ILP). *QSAR Comb. Sci.* **2003**, *22*, 527−532.

(20) Srinivasan, A.; King, R. D.; Muggleton, S. H.; Sternberg, M. J. E. The Predictive Toxicology Evaluation Challenge. In *Fifteenth International Joint Conference on Artificial Intelligence*; Morgan Kaufmann: San Francisco, 1997; pp. 4−9.

(21) Marchand-Geneste, N.; Watson, K. A.; Alsberg, B. K.; King, R. D. New Approach to Pharmacophore Mapping and QSAR Analysis Using Inductive Logic Programming. Application to Thermolysin Inhibitors and Glycogen Phosphorylase Inhibitors. *J. Med. Chem.* **2002**, *45*, 399−409.

(22) Pearlman, R. S. *Concord User's Manual*; Tripos, Inc.: St Louis, MO, 2000.

(23) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. I. Method. *J. Comp. Chem.* **1989**, *10*, 209−220.

(24) *BioMedCAChe,* version 6.1.10; Fujitsu Ltd: Kawasaki, Kanagawa, Japan, 2000−2004; Oxford Molecular Ltd: Oxford, U.K. ,1989−2000.

(25) Ghose, A. K.; Prichett, P.; Crippen, G. M. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships III: Modelling Hydrophobic Interactions. *J. Comput. Chem.* **1988**, *9*, 80−90.

(26) Katritzky, A. R.; Lobanov, V. S. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, *24*, 279−287.

(27) Cronin, M. T. D.; Netzeva, T. I.; Dearden, J. C.; Edwards R.; Worgan, A. D. P. Assessment and Modeling of the Toxicity of Organic Chemicals to *Chlorella vulgaris*: Development of a Novel Database. *Chem. Res. Toxicol.* **2004**, *17*, 545−554.

(28) Muggleton, S. H.; De Raedt, L. Inductive Logic Programming: Theory and Methods. *J. Logic Programming* **1994**, *19*, 629−679.

(29) Muggleton, S. H. Inverse Entailment and Progol. *New Gener. Comput.* **1995**, *13*, 245−286.

(30) Muggleton, S. H.; Bryant, C. H. In *Proceedings of the Tenth International Conference on Inductive Logic Programming (ILP-2000)*; Cussens, J., Frisch, A. M., Eds.; Springer-Verlag: Berlin, 2000; pp 130−146.

(31) Gartner, T.; Flach, P. A.; Kowalczyk, A.; Smola, A. J. Multi-Instance Kernels. In *Proceedings of the 19th International Conference on Machine Learning*; Sammut, C., Hoffmann, A., Eds.; Morgan Kaufmann: San Francisco, CA, 2000; pp 79−186.

(32) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(33) McNemar, Q. Note on the Sampling Error of the Difference Between Correlated Proportions of Percentages. *Psychometrica* **1947**, *12*, 153−157.

(34) Leo, A. J.; Hansch, C. Role of Hydrophobic Effects in Mechanistic QSAR. *Prosp. Drug Discovery Des.* **1999**, *17*, 1−25.

(35) Jankowiak, R.; Rogan, E. G.; Cavalieri, E. L. Role of Fluorescence Line-Narrowing Spectroscopy and Related Luminescence-Based Techniques in the Elucidation of Mechanisms of Tumor Initiation by Polycyclic Aromatic Hydrocarbons and Estrogens. *J. Phys. Chem. B* **2004**, *108*, 10266−10283.

(36) Chana, A.; Concejero, M. A.; de Frutos, M.; González, M. J.; Herradon, B. Computational Studies on Biphenyl Derivatives. Analysis of the Conformational Mobility, Molecular Electrostatic Potential, and Dipole Moment of Chlorinated Biphenyl: Searching for the Rationalization of the Selective Toxicity of Polychlorinated Biphenyls (PCBs). *Chem. Res. Toxicol.* **2002**, *15*, 1514−1526.

(37) Gamage, S. A.; Figgitt, D. P.; Wojcik, S. J.; Ralph, R. K.; Ransijn, A.; Mauel, J.; Yardley, V.; Snowdon, D.; Croft, S. L.; Denny, W. A. Structure−Activity Relationships for the Antileishmanial and Antitrypanosomal Activities of 1′-Substituted 9-Anilinoacridines. *J. Med. Chem.* **1997**, *40*, 2634−2642.

(38) Sixt, S.; Altschuh, J.; Bruggemann, R. Quantitative Structure−Toxicity Relationships for 80 Chlorinated Compounds Using Quantum Chemical Descriptors. *Chemosphere* **1995**, *30*, 2397−2414.

(39) Williams, P. L.; James, R. C.; Roberts, S. M. *Principles of Toxicology, Environmental and Industrial Applications*; John Wiley & Sons, Inc.: New York, 2000.

(40) Timbrell, J. *Principles of Biochemical Toxicology*; Taylor & Francis: London, 2001.

(41) Kaun-Yu, L.; Fu-Wei, T.; Chia-Jung, W.; Pei-Sham, L. Suppression by Phthalates of the Calcium Signaling of Human Nicotinic Acetylcholine Receptors in Human Neuroblastoma SH-SY5Y Cells. *Toxicology* **2004**, *200*, 113−121.

(42) Harris, C. A.; Henttu, P.; Parker, M. G.; Sumpter, J. P. The Estrogenic Activity of Phthalate Esters in Vitro. *Health Prospects* **1997**, *105*, 802−811.

(43) Hathway, D. E. *Molecular Aspects of Toxicology*; Royal Society of Chemistry: London, 1984.

(44) Kaiser, K. L. E.; Niculescu, S. P. Using Probabilistic Neural Networks to Model the Toxicity of Chemicals to the Fathead Minnow (*Pimephales promelas*): A Study Based on 865 Compounds. *Chemosphere* **1999**, *38*, 3237−3245.

(45) *The TerraTox rM & TerraFit Software Suite*, version 1.502; TerraBase Inc.: Burlington, ON, Canada, 1998.

(46) Klopman, G.; Salakhov, R.; Rosenkranz, H. S. Multiple Computer-Automated Structure Evaluation Study of Aquatic Toxicity II. Fathead Minnow. *Environ. Toxicol. Chem.* **2000**, *19*, 441−447.

(47) Mitchell, T. M. *Machine Learning*; McGraw Hill, New York, 1997.