Experiments in predicting biodegradability

Sašo Džeroski (1), Hendrik Blockeel (2), Boris Kompare (3),
Stefan Kramer (4), Bernhard Pfahringer (4), Wim Van Laer (2)

(1) Department of Intelligent Systems, Jozef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
(2) Department of Computer Science, Katholieke Universiteit Leuven
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
(3) Faculty of Civil Engineering and Geodesy, University of Ljubljana
Hajdrihova 28, SI-1000, Ljubljana, Slovenia
(4) Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria

Abstract. We present a novel application of inductive logic programming (ILP) in the area of quantitative structure-activity relationships (QSARs). The activity we want to predict is the biodegradability of chemical compounds in water. In particular, the target variable is the half-life in water for aerobic aqueous biodegradation. Structural descriptions of chemicals in terms of atoms and bonds are derived from the chemicals' SMILES encodings. Definition of substructures are used as background knowledge. Predicting biodegradability is essentially a regression problem, but we also consider a discretized version of the target variable. We thus employ a number of relational classification and regression methods on the relational representation and compare these to propositional methods applied to different propositionalisations of the problem. Some expert comments on the induced theories are also given.

1 Introduction

The persistence of chemicals in the environment (or to environmental influences) is welcome only until the time the chemicals fulfill their role. After that time or if they happen to be at the wrong place, the chemicals are considered pollutants. In this phase of chemicals' life-span we wish that the chemicals disappear as soon as possible. The most ecologically acceptable (and a very cost-effective) way of 'disappearing' is degradation to components which are not considered pollutants (e.g. mineralization of organic compounds). Degradation in the environment can take several forms, from physical pathways (erosion, photolysis, etc.), through chemical pathways (hydrolysis, oxydation, diverse chemolises, etc.) to biological pathways (biolysis). Usually the pathways are combined and interrelated, thus making degradation even more complex. In our study we focus on biodegradation in an aqueous environment under aerobic conditions, which affects the quality of surface- and groundwater.

The problem of properly assessing the time needed for ultimate biodegradation can be simplified to the problem of determining the half-life time of that process. However, very few measured data exist and even these data are not taken under controlled conditions. It follows that an objective and comprehensive database on biolysis half-life times can not be found easily. The best we were able to find was in a handbook of degradation rates [10]. The chemicals described in this handbook were used as the basis of our study.

Usually, authors try to construct a QSAR model/formula for only one class of chemicals, or congeners of one chemical, e.g. phenols. This approach to QSAR model construction has an implicit advantage that only the variation with respect to the class mainstream should be identified and properly modelled. Contrary to the described situation, our database comprises several families of chemicals, e.g. alcohols, phenols, pesticides, chlorinated aliphatic and aromatic hydrocarbons, acids, diverse other aromatic compounds, etc. From this point of view, the construction of adequate QSAR models/formulae is a much more difficult task.

We apply several machine learning methods, including several inductive logic programming methods, to the above database in order to construct SAR/QSAR models for biodegradability. The remainder of the paper is organized as follows. Section 2 describes the dataset and how the representations used by the different machine learning systems were generated. Section 3 lists the representation and the machine learning systems employed, and describes the experimental setup. Section 4 presents the experimental results, including expert comments on some of the induced rules. Section 5 gives further discussion, Section 6 comments on related work, and Section 7 concludes and gives some directions for further work.

2 The dataset

The database used was derived from the data in the handbook of degradation rates [10]. The authors have compiled from available literature the degradation rates for 342 widely used (commercial) chemicals. Where no measured data on degradation rates were available, expert estimation was performed. The main source of data employed was the Syracuse Research Corporation's (SRC) Environmental Fate Data Bases (EFDB), which in turn used as primary sources of information DATALOG, CHEMFATE, BIOLOG, and BIODEG files to search for pertinent data.

For each considered chemical the book contains degradation rates in the form of a range of half-life times (low and high estimate) for overall, biotic and abiotic degradation in four environmental compartments, i.e., soil, air, surface water and ground water. We focus on surface water here. The overall degradation half-life is a combination of several (potentially) present pathways, e.g., surface water photolysis, photooxydation, hydrolysis and biolysis (biodegradation). These can occur simultaneously and have even synergistic effects, resulting in a half-life time (HLT) smaller than the HLT for each of the basic pathways. We focus on biodegradation here, which was considered to run in unacclimated aqueous conditions, where biota (living organisms) are not adapted to the specific pollutant considered. For biodegradation, three environmental conditions were considered:

aerobic, anaerobic, and removal in waste water treatment plants (WWTP). In our study we focus on aqueous biodegradation HLT's in aerobic conditions.

The target variable for machine learning systems that perform regression was the natural logarithm of the arithmetic mean of the low and high estimate of the HLT for aqueous biodegradation in aerobic conditions, measured in hours.

A discretized version of the arithmetic mean was also considered in order to enable us to apply classification systems to the problem. Four classes were defined : chemicals degrade *fast* (mean estimate HLT is up to 7 days), *moderate*ly fast (one to four weeks), *slow*ly (one to six months), or are *resistant* (otherwise).

From this point on, we proceeded as follows. The CAS (Chemical Abstracts Service) Registry Number of each chemical was used to obtain the SMILES [22] notation for the chemical. In this fashion, the SMILES notations for 328 of the 342 chemicals were obtained.

The SMILES notation contains information on the two-dimensional structure of a chemical. So, an atom-bond representation, similar to the representation used in experiments to predict mutagenicity, can be generated from a SMILES encoding of a chemical. A DCG-based translator that does this has been written by Michael de Groeve and is maintained by Bernhard Pfahringer. We used this translator to generate atom-bond relational representations for each of the 328 chemicals. Note that the atom-bond representation here is less powerful than the QUANTA-derived representation, which includes atom charges, atom types and a richer selection of bond types. Especially the types carry a lot of information on the substructures that the respective atoms/bonds are part of.

A global feature of each chemical is its molecular weight. This was included in the data. Another global feature is logP, the logarithm of the compound's octanol/water partition coefficient, used also in the mutagenicity application. This feature is a measure of hydrophobicity, and can be expected to be important since we are considering biodegradation in water.

The basic atom and bond relations were then used to define a number of background predicates defining substructures / functional groups that are possibly relevant to the problem of predicting biodegradability. These predicates are: nitro ($-NO_2$), sulfo ($-SO_2$ or $-O-S-O_2$), methyl ($-CH_3$), methoxy ($-O-CH_3$), amine, aldehyde, ketone, ether, sulfide, alcohol, phenol, carboxylic acid, ester, amide, imine, alkyl_halide (R-Halogen where R is not part of a resonant ring), ar_halide (R-Halogen where R is part of a resonant ring), epoxy, n2n ($-N=N-$), c2n ($-C=N-$), benzene (resonant C_6 ring), hetero_ar_6_ring (resonant 6 ring containing at least 1 non-C atom), non_ar_6c_ring (non-resonant C_6 ring), non_ar_hetero_6_ring (non-resonant 6 ring containing at least 1 non-C atom), six_ring (any type of 6 ring), carbon_5_ar_ring (resonant C_5 ring) non_ar_5c_ring (non-resonant C_5 ring), non_ar_hetero_5_ring (non-resonant 5 ring containing at least 1 non-C atom), and five_ring (any type of 5 ring). Each of these predicates has three arguments: MoleculeID, MemberList (list of atoms that are part of the functional group) and ConnectedList (list of atoms connected to atoms in MemberList, but not in MemberList themselves).

3 Experiments

3.1 Representations

Molecular weight, logP and the abovementioned predicates form the basic relational representation (denoted by R1) considered in our experiments. Two propositional representations were derived from this. The first one (denoted P1) has an attribute $fgCount$ for each three-argument predicate fg of the background knowledge, which is the number of distinct functional groups of type fg in a molecule. Including logP and molecular weight, this representation has 31 attributes.

The second propositional representation (denoted P2) has been derived by counting all substructures of two and three atoms plus all four-atom substructures of a star-topology (no chains). Substructures that appear in at least three compounds (59 of them) are taken into account. For each such substructure we have a feature counting the number of distinct substructures of that kind in a molecule. The second propositional representation also includes logP and molecular weight.

Many of the functional groups have been selected from the PTE (predictive toxicology evaluation) domain theory [20], where the task is to predict carcinogenicity of chemicals. In this domain, the approach of Dehaspe and Toivonen [7] to discover (count) most frequent substructures that occur in the dataset and use these in conjunction with propositional learners has been among the most successful. Our small substructure representation has been derived along these lines.

3.2 Systems

A variety of classification and regression systems were applied to the classification, respectively regression version of the biodegradability problem. Propositional systems were applied to representations P1 and P2. For classification, these were the decision-tree inducer C4.5 [16] and the rule induction program RIPPER [6]. For regression, the regression-tree induction program M5' [21], a re-implementation of M5 [17] was used. It can construct linear models in the leaves of the tree.

Relational learning systems applied include ICL [8], which induces classification rules, SRT [15] and TILDE [1]. The latter are capable of inducing both classification and regression trees. ICL is an upgrade of CN2 [5] to first-order logic, TILDE is an upgrade of C4.5, and SRT is an upgrade of CART [3]. TILDE cannot construct linear models in the leaves of its trees; SRT can.

Finally, FFOIL [18] was also applied to the classification version of the problem. It used a representation (denoted R2) based on the atom and bond relations, designed to avoid problems with indeterminate literals. New predicates are introduced for conjunctions of the form $atom(M, X, Element1, \_, \_)$, $bond(M, X, Y, BondType), atom(M, Y, Element2, \_, \_)$. E.g., $o2s(M, X, Y)$ stands for $atom(M, X, o, \_, \_), bond(M, X, Y, 2), atom(M, Y, s, \_, \_)$.

Regarding parameter settings, default settings were employed for all systems wherever possible. Deviations from default parameter settings will be mentioned where appropriate in the results section.

3.3 Evaluation

Performance on unseen cases was estimated by performing five 10-fold cross-validations. The same folds were used by all systems. Performances reported are averages over the 5 cross-validations. Some of the induced models were inspected by B. Kompare, acting here as a domain expert, who provided some comments on their meaning and agreement with existing knowledge in the domain.

For the regression systems, correlation between the actual and predicted values of the log mean half-time of aerobic aqueous biodegradation is reported. We also measure classification accuracy (as described below) achieved by discretizing the real-valued predictions.

For the classification systems, classification accuracy is reported. We are dealing with ordered class values and misclassification of, e.g., fast as slow is a bigger mistake than misclassification of fast as moderate. We thus also record accuracy where only misclassification by more than one class up or down counts as an error (e.g., fast as slow, or resistant as moderate). This is denoted as Accuracy (+/-1) in Table 1.

4 Results

Table 1 gives an overview of the performance of the different classification and regression systems as applied to the problem of predicting biodegradability. SRT-C denotes SRT used to learn classification trees, while SRT-R denotes SRT used to learn regression trees. TILDE-C and TILDE-R have similar meaning. The first column lists the system applied, the second the representation used. The second column also lists some parameters changed from their default values. The representations are described in Section 3.1 (P1,P2, R1) and 3.2 (R2). The next three columns list performance measures as described in Section 3.3.

C4.5 was used on the two different propositional representations. Better performance was achieved using P2. Default parameters were used. The trees generated were too bushy for expert inspection. C4.5 performs worst in terms of large misclassification (e.g. fast as slow) errors, i.e. in terms of the measure Accuracy (+/-1).

RIPPER achieves highest accuracy of the classification systems applied. With its default parameters RIPPER prunes drastically, producing small rule sets. The rule set derived from the entire dataset for representation P2 is given in Figure 1, together with some comments provided by our domain expert.

The expert liked the rule-based representation and the concise rules very much (best of the representations shown to him, which included classification and regression trees induced by M5', SRT and TILDE, as well as clausal theories induced by ICL). The rules make sense, but are possibly pruned too much and cover substantial numbers of negative examples.

Table 1. Performance of machine learning systems predicting biodegradability.

System	Representation	Accuracy	Accuracy (+/-1)	Correlation (r)
C4.5	P1	55.2	86.2	-
C4.5	P2	56.9	82.4	-
RIPPER	P1 (-S0)	52.6	89.8	-
RIPPER	P2	57.6	93.9	-
M5'	P1	53.8	94.5	0.666
M5'	P2	59.8	94.7	0.693
FFOIL	R2	53.0	88.7	-
ICL	R1	55.7	92.6	-
SRT-C	P1	51.3	88.2	-
SRT-C	P1+R1	55.0	90.0	-
SRT-R	P1	49.8	93.8	0.580
SRT-R	P1+R1	52.6	93.0	0.632
TILDE-C	R1	51.0	88.6	-
TILDE-C	P1+R1	52.0	89.0	-
TILDE-R	R1	52.6	94.0	0.622
TILDE-R	P1+R1	52.4	93.9	0.623
BIODEG				0.607

Pruning was then turned down in RIPPER (option -S0), producing larger sets of longer rules, at a moderate loss of accuracy. The accuracy for representation P2 is in this case 54.8 % (again estimated by doing five 10-fold cross-validations).

M5' achieves best results among the systems applied in terms of both regression accuracy (almost 0.7) and classification accuracy (almost 60 %, respectively 95 %). M5' was used with pruning turned down (-f0.0), as this seemed to perform best in terms of accuracy. Linear models are by default allowed in the leaves of the trees. Trees generated with these settings were too large and cumbersome to interpret.

Trees were generated from the entire dataset with more intensive pruning to ensure they were of reasonable size for interpretation by the domain expert. The tree generated from representation P2 is shown in Figure 2. The setting -f1.2 was used for pruning. The numbers in brackets denote the number of examples in a leaf and the relative error of the model in that leaf on the training data. So LM1 was constructed from 80 examples and has 49.7 % relative error on these 80 examples.

Unsurprisingly, the most important feature turns out to be logP, the hydrophobicity measure. For compounds to biodegrade fast in water, it helps if they are less hydrophobic. When a compound is not very hydrophobic (logP < 4.005), molecular weight is an important feature. With relatively low molecular weight (< 111.77), the presence of an $-OH$ group indicates smaller half-life times. With no $-OH$ groups (LM1), halogenated compounds degrade more slowly and so do compounds with CN substructures (positive coefficients in LM1). This is also consistent with the expert comments on the RIPPER rules.

FFOIL uses the R2 representation (Section 3.2). The settings -d10 and -a65 were used; -d10 allows the introduction of "deeper variables" (this does not seem

```
resistant :- logP>=4.91, 'C[H]'<=15 (27/4).
  % Nonpolar (hydrophobic) compounds degrade less readily
resistant :- 'C[Cl]'>=3, mweight<=165.834 (7/1).
  % Halogenated compounds are resistant
fast :- mweight<=110.111, 'O[H]'>=1 (18/4).
  % Alcohols (alkyl -OH) are fast to degrade
fast :- mweight<=108.096, 'C=O'>=1 (15/7).
  % C=O readily degrades
slow :- 'N=O'>=1, mweight<=130.19 (10/0).
  % Compounds  with N(-)O degrade slowly
slow :- logP>=1.52, 'C[H]'<=5 (31/16).
slow :- 'CN'>=1, logP>=1.7, mweight>=249.096 (11/3).
  % Very heavy and possibly toxic
slow :- 'C=O'<=0, mweight>=121.182, 'CN'>=1 (23/15).
default moderate (85/51).
```

Fig. 1. Rules for predicting biodegradability induced by RIPPER.

to have any impact), and -a65 means that a clause must be 65% correct or better (FFOIL's default is 80 %, which seems too demanding in this domain).

FFOIL only uses the atom and bond relations, molecular weight and logP, but not the functional group relations/predicates. On the entire dataset, FFOIL induces 54 rules. It is interesting that some of these rules use negation. The rule `activity(A,fast):-mw(A,C), logp(A,D), not(c1cl(A,_1,_2)), C>104.151, D>1.52, C <=129.161, D<=3.45,!.` states that a compound A degrades fast if it is not halogenated, is relatively light, and relatively nonhydrophobic.

ICL was applied to representation R1. In terms of accuracy, it achieves better results than all other systems not using P2, and in terms of Accuracy (+/-1) it performs better than all classification systems except RIPPER on P2. Using R1+P1+P2 yields worse results than R1 alone. The theory induced from the entire dataset contains 87 rules.

An example rule is: `moderate(M) :- atom(M,A1,Elem1,_,_), Elem1 = s, mweight(M,MW),lt(MW,190),gt(MW,90)`. It states that a compound with a sulphur atom and molecular weight between 90 and 190 degrades moderately fast. The expert comments that sulphur slows down biodegradation.

Another rule states that a compound is fast to degrade if it contains a benzene and a phenol group and is lighter that 170. The expert comments that in this case degradability is probably due to hydrolisis and photolysis.

SRT upgrades CART to a relational representation, as mentioned above. From CART it inherits error-complexity pruning. It can construct linear models in the leaves and extends CART methodology by cross-validating these models. No linear models in the leaves were allowed in the experiments reported here.

The SRT results were not obtained by using default settings. Results for unmodified error-complexity pruning were not competitive. We thus forced SRT to overfit: from the sequence of pruned trees ordered by increasing complexity we took the first tree after the most accurate tree that was within one standard error of the former. The resulting trees were too large for inspection.

```
logP <= 4.005
|   mweight <= 111.77
|   |   'O[H]' <= 0.5 LM1 (80/49.7%)
|   |   'O[H]'  > 0.5 LM2 (22/50.7%)
|   mweight  > 111.77
|   |   'C=O' <= 0.5 LM3 (112/65.4%)
|   |   'C=O'  > 0.5
|   |   |   'CO' <= 1.5
|   |   |   |   'CN[H]' <= 1.5
|   |   |   |   |   'C[Cl]' <= 1.5 LM4 (7/0%)
|   |   |   |   |   'C[Cl]'  > 1.5 LM5 (2/6.68%)
|   |   |   |   'CN[H]'  > 1.5 LM6 (9/33.8%)
|   |   |   'CO'  > 1.5
|   |   |   |   'C[H]' <= 12.5
|   |   |   |   |   'N[H]' <= 0.5
|   |   |   |   |   |   'CO' <= 2.5 LM7 (5/0%)
|   |   |   |   |   |   'CO'  > 2.5 LM8 (10/46.1%)
|   |   |   |   |   'N[H]'  > 0.5 LM9 (5/16.3%)
|   |   |   |   'C[H]'  > 12.5
|   |   |   |   |   logP <= 2.26 LM10 (5/0%)
|   |   |   |   |   logP  > 2.26 LM11 (4/2.42%)
logP  > 4.005
|   logP <= 4.895 LM12 (27/53.9%)
|   logP  > 4.895
|   |   'C[H]' <= 15.5 LM13 (31/55%)
|   |   'C[H]'  > 15.5 LM14 (9/45.9%)

Linear models at the leaves:
  Unsmoothed (simple):
    LM1: class = 6.1 + 0.525'C[Cl]' + 0.618'CN' - 1.09'C=O' - 0.559'CN[H]'
    LM2: class = 4.71
    LM3: class = 7.38 - 0.00897mweight + 0.889'C[Br]' + 0.576'C[Cl]' + 0.522'CN' + 0.113'N=O'
    LM4: class = 6.04
    LM5: class = 6.7
    LM6: class = 9.83 - 1.8'N[H]'
    LM7: class = 4.56
    LM8: class = 5.6
    LM9: class = 6.15
    LM10: class = 6.04
    LM11: class = 6.52 - 0.252'O[H]'
    LM12: class = 6.77 + 0.182'C[Cl]' - 0.357'CO'
    LM13: class = 9.43 - 1.52'CN'
    LM14: class = 12.2 - 0.0157mweight
```

Fig. 2. Regression tree for predicting biodegradability induced by M5'.

Both a propositional (P1) and a relational representation (P1+R1) were used. Adding the relational information improves accuracy, the greatest jump being observed for classification accuracy of SRT-C. Using P2 in addition (P1+P2+R1) only improves the regression results marginally. SRT-C is better than SRC-R on accuracy, but worse on Accuracy (+/-1).

TILDE was used for both classification and regression, once using R1 and once using P1+R1. TILDE-C was used with default settings. TILDE-R was used with its `ftest` parameter set to 0.01, which causes maximal pre-pruning.

The use of P1 in addition to R1 does not change the performance of TILDE. Better performance is achieved with regression, not in terms of Accuracy but in terms of Accuracy (+/-1). Using P2 in addition (P1+P2+R1) yields worse regression results (r=0.58).

An example regression tree induced by TILDE-R from the entire dataset is given in Figure 3. This tree has actually been generated without using logP information. It was analysed and commented upon by the domain expert. The

```
activ(A,B)
carbon_5_ar_ring(A,C,D) ?
+--yes: [9.10211]  % Aromatic compounds are relatively slow to degrade
+--no:  aldehyde(A,E,F) ?
        +--yes: [4.93332] % Aldehydes are fast
        +--no:  atm(A,G,h,H,I) ? % If H not present should degrade slowly
                +--yes: mweight(A,J) , J =< 80 ?
                |       +--yes: [5.52184] % Low weight ones degrade faster
                |       +--no:  ester(A,K,L) ? % Esters degrade fast
                |               +--yes:mweight(A,M) , M =< 140 ?
                |               |      +--yes: [4.93332]
                |               |      +--no:  [5.88207]
                |               +--no: mweight(A,N) , N =< 340 ?
                |                       +--yes:carboxylic_acid(A,O,P) ? % Acids degrade fast
                |                       |      +--yes:[5.52288]
                |                       |      +--no: ar_halide(A,Q,R) ? % Halogenated - slow
                |                       |             +--yes: alkyl_halide(A,S,T) ?
                |                       |             |       +--yes: [11.2742]
                |                       |             |       +--no:  [7.81235]
                |                       |             +--no:  phenol(A,U,V) ?
                |                       |                     +--yes:mweight(A,W) , W =< 180 ?
                |                       |                     |      +--yes:[4.66378]
                |                       |                     |      +--no: [7.29547]
                |                       |                     +--no: [6.86852]
                |                       +--no: [8.28685]
                +--no:  mweight(A,X) , X =< 100 ?
                        +--yes: [6.04025]
                        +--no:  [8.55286]
```

Fig. 3. A regression tree for predicting biodegradability induced by TILDE.

fact that it does not use logP actually makes it easier for the influence of the functional groups on biodegradability to be identified. Namely, when logP is used, a large part of the tree uses logP only. Some of the expert comments are given in the tree itself.

5 Discussion

Overall, propositional systems applied to representation P2 yield best performance. M5' on this representation yields the highest overall accuracy, Accuracy (+/-1) and correlation. RIPPER follows with the second best classification accuracy and Accuracy (+/-1) matched only by TILDE-R. Of the relational learning systems, ICL performs best with highest classification accuracy and Accuracy (+/-1) comparable to that of SRT-R and TILDE-R.

Regression systems perform better than classification ones. This does not clearly show when one looks at accuracy alone, but it becomes clearer when one looks at Accuracy (+/-1). It thus seems that regression problems can best be handled by regression systems.

Using relational information in addition to the propositional formulation P1 does not bring drastic improvements. SRT and TILDE perform slightly better or the same on P1 + R1 as compared to P1. SRT and TILDE used for regression on P1 + R1 still perform (slightly) worse than M5' on P1. The reason for this might be the fact that M5' was using linear regression in the leaves, while SRT and TILDE were not.

Note that the propositional representations P1 and P2 contain structural features derived 1) directly from the functional group relations and 2) from

the atom and bond relations. These features count occurrences of substructures within compounds. P1 contains definitions of both small and larger groups (such as rings), while P2 mainly contains small structures (up to 4 atoms).

The biodegradation rates used in this study were expert estimates rather than measurements for the most part. We have thus been modeling expert opinions on biodegradation rates, and not biodegradation rates themselves. This means that we have also modeled expert estimation errors. To the authors' knowledge, only small datasets containing measured biodegradation rates for structurally related chemicals are publicly available at present.

6 Related work

Related work includes QSAR applications of machine learning and ILP, on one hand, and constructing QSAR models for biodegradability, on the other hand. On the ILP side, QSAR applications include drug design (e.g. [13]), mutagenicity prediction (e.g. [19]), and toxicity prediction [20]. The latter two are closely related to our application. In fact, we have used a similar representation and reused parts of the background knowledge developed for them.

On the biodegradability side, [11] is closest to our work. The last row of Table 1, marked BIODEG, gives the correlation between the actual values of the continuous class and predictions made by the BIODEG program [12]. The correlation is calculated for all 328 chemicals in our database, since the BIODEG program has been derived independently. This program estimates the probability of rapid aerobic biodegradation in the presence of mixed populations of environmental organisms. It uses a model derived by linear regression [11].

The best results of our experiments (correlation of 0.7) are considerably better than the BIODEG program predictions (correlation 0.6). Furthermore, while the reported performance results for the machine learning systems are for unseen cases, some of the 200 chemicals used in developing BIODEG also appear in our database. In [11], CAS numbers for 144 of the 200 chemicals used to derive BIODEG are provided; of these, 21 also appear in our database. The correlation of BIODEG predictions is thus probably even lower than 0.6 for unseen cases.

Work on applying machine learning to predict biodegradability includes [14], who compared several AI tools on the same domain and data and found these to yield better results than the classical statistical and probabilistic approaches, [23, 4] who applied neural nets, and [9] who applied several different approaches.

7 Conclusions and further work

Predicting biodegradability is a QSAR problem, similar to predicting mutagenicity or toxicity. Based on a handbook of biodegradation rates, we have developed a relational dataset including a structural representation of compounds and background knowledge on potentially relevant substructures. This dataset is suitable for both propositional and relational learning. Particular attention was paid to data quality issues: many datasets of this kind have surprisingly many errors,

such as incorrect SMILES codes, which essentially result in incorrect descriptions of the compounds and affect the resulting QSAR theories accordingly. The dataset itself is thus a contribution on its own.

We have applied a range of machine learning systems, including ILP systems, to several representations derived from the relational description of the compounds. Best performance was achieved on good propositionalisations derived by counting substructures. This is in agreement with, e.g., the predictive toxicology evaluation results (cf. this volume) where best results were achieved by propositional systems using relational features representing the presence/count of frequent substructures.

M5', which achieves the best results, outperforms an approach derived by biodegradability experts, implemented in the program BIODEG. The theories induced by the machine learning systems were easy to interpret (size permitting) and made sense to the domain expert. Given that the biodegradation rates that we used as values of the target variable are mostly estimates and not measured values, overall performance is satisfactory.

There is a variety of directions for further work. One possibility is to study overall degradation and biodegradation comparatively. Identifying chemicals for which degradation and biodegradation time differ is an important topic. Characterising such chemicals would be an interesting learning problem.

Another important issue is how performance is evaluated when only estimates of the target variable are provided. One could argue that if the learned theory predicts a value which is between the low and high estimate provided by an expert, its prediction is correct. In a sense, we may have applied a too strict evaluation criterion here, trying to fit the log mean half-life time, while providing a value in the provided interval may have been sufficient.

Predicting the logarithm of the mean of the low and high estimates of the degradation rate is close to predicting the logarithm of the high estimate. Predicting the (logarithm of the) low estimate and combining the two predictions might yield better results. This should also be investigated in further work.

Acknowledgements: This work was supported in part by the ESPRIT IV Project 20237 ILP2. Thanks are due to: Irena Cvitanič for help with preparing the dataset in computer-readable form; Christoph Helma for help in preparing the background knowledge and calculating logP; Ross King and Ashwin Srinivasan for providing some definitions of the functional group predicates.

References

1. Blockeel, H. and De Raedt, L. 1998. Top-down induction of first order logical decision trees. *Artificial Intelligence*, 101(1-2): 285–297.
2. Boethling, R.S., and Sabljic, A. 1989. Screening-level model for aerobic biodegradability based on a survey of expert knowledge. *Environ. Sci. Technol.* 23: 672–679.
3. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and Regression Trees.* Wadsworth, Belmont.

4. Cambon, B., and Devillers, J. 1993. New trends in structure-biodegradability relationships. *Quant. Struct. Act. Relat.* 12(1): 49–58.
5. Clark, P., and Boswell, R. 1991. Rule induction with CN2: some recent improvements. In *Proc. 5th European Working Session on Learning*, pages 151–163. Springer, Berlin.
6. Cohen W. 1995. Fast effective rule induction. In *Proc. 12th Intl. Conf. on Machine Learning*, pages 115–123. Morgan Kaufmann, San Mateo, CA.
7. Dehaspe, L., and Toivonen, H. 1999. Frequent query discovery: a unifying ILP approach to association rule mining. *Data Mining and Knowledge Discovery.*
8. De Raedt, L., and Van Laer, W. 1995. Inductive constraint logic. In *Proc. 6th Intl. Workshop on Algorithmic Learning Theory*, pages 80–94. Springer, Berlin.
9. Gamberger, D., Sekuak, S., and Sabljič, A. 1993. Modelling biodegradation by an example-based learning system. *Informatica* 17: 157–166.
10. Howard, P.H., Boethling, R.S., Jarvis, W.F., Meylan, W.M., and Michalenko, E.M. 1991. *Handbook of Environmental Degradation Rates.* Lewis Publishers.
11. Howard, P.H., Boethling, R.S., Stiteler, W.M., Meylan, W.M., Hueber, A.E., Beauman, J.A., and Larosche, M.E. 1992. Predictive model for aerobic biodegradability developed from a file of evaluated biodegradation data. *Environ. Toxicol. Chem.* 11: 593–603.
12. Howard, P. and Meylan, W. 1992. User's Guide for the Biodegradation Probability Program, Ver. 3. Syracuse Res. Corp., Chemical Hazard Assessment Division, Environmental Chemistry Center, Syracuse, NY 13210, USA.
13. King, R.D., Muggleton, S.H., Lewis, R.A., and Sternberg, M.J.E. 1992. Drug design by machine learning : the use of inductive logic programming to model the structure-activity relationship of trimethoprim analogues binding to dihydrofolate reductase. *Proc. National Academy of Sciences USA*, 89: 11322–11326.
14. Kompare, B. 1995. *The use of artificial intelligence in ecological modelling.* Ph.D. Thesis, Royal Danish School of Pharmacy, Copenhagen, Denmark.
15. Kramer, S. 1996. Structural regression trees. In *Proc. 13th Natl. Conf. on Artificial Intelligence*, pages 812–819. AAAI Press/The MIT Press.
16. Quinlan, J.R. 1993a. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA.
17. Quinlan, J.R. 1993b. Combining instance-based and model-based learning. In *Proc. 10th Intl. Conf. on Machine Learning*, pages 236–243. Morgan Kaufmann, San Mateo, CA.
18. Quinlan, J.R. 1996. Learning first-order definitions of functions. *Journal of Artificial Intelligence Research*, 5:139–161.
19. Srinivasan, A., Muggleton, S.H., Sternberg, M.J.E., and King, R.D. 1996. Theories for mutagenicity: A study in first-order and feature-based induction. *Artificial Intelligence* 85(1-2): 277-299.
20. Srinivasan, A., King, R.D., Muggleton, S.H., and Sternberg, M.J.E. 1997. The predictive toxicology evaluation challenge. In *Proc. 15th Intl. Joint Conf. on Artificial Intelligence*, pages 4–9. Morgan Kaufmann, San Mateo, CA.
21. Wang, Y., and Witten, I.H. 1997. Inducing model trees for continuous classes. In *Poster Papers - 9th European Conf. on Machine Learning*, pages 128–137. Prague, Czech Republic. URL: http://www.cs.waikato.ac.nz/~ ml/publications.html.
22. Weininger D. 1988. SMILES, a Chemical and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* 28(1): 31-6.
23. Zitko, V. 1991. Prediction of biodegradability of organic chemicals by an artificial neural network. *Chemosphere*, Vol. 23, No. 3: 305-312.