Environ. Sci. Technol. 1994, 28, 459–465

Group Contribution Method for Predicting Probability and Rate of Aerobic Biodegradation

Robert S. Boethling,† Philip H. Howard,*,‡ William Meylan,‡ William Stiteler,‡ Julie Beauman,‡ and Nestor Tirado†

Office of Pollution Prevention and Toxics 7406, U.S. Environmental Protection Agency, 401 M Street SW, Washington, D.C. 20460, and Syracuse Research Corporation, Merrill Lane, Syracuse, New York 13210

Two independent training sets were used to develop four mathematical models for predicting aerobic biodegradability from chemical structure. All four of the models are based on multiple regressions against counts of 36 preselected chemical substructures plus molecular weight. Two of the models, based on linear and nonlinear regressions, calculate the probability of rapid biodegradation and can be used to classify chemicals as rapidly or not rapidly biodegradable. The training set for these models consisted of qualitative summary evaluations of all available experimental data on biodegradability for 295 chemicals. The other two models allow semi-quantitative prediction of primary and ultimate biodegradation rates using multiple linear regression. The training set for these models consisted of estimates of primary and ultimate biodegradation rates for 200 chemicals, gathered in a survey of 17 biodegradation experts. The two probability models correctly classified 90% of the chemicals in their training set, whereas the two survey models calculated biodegradation rates for the survey chemicals with $R^2 \cong 0.7$. These four models are intended for use in chemical screening and in setting priorities for further review.

Introduction

Chemical scoring systems for identifying substances of priority concern have proliferated in concert with environmental legislation. Much of the early impetus for developing such systems derived from the need to review Premanufacture Notifications (PMNs) under Section 5 of the Toxic Substances Control Act (TSCA) and by the TSCA-mandated screening of existing chemicals by the U.S. Interagency Testing Committee. However, virtually every EPA program is now involved in chemical scoring in one way or another. Examples include reportable quantity (RQ) adjustment methodology under the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA; Superfund); the Superfund Hazard Ranking System; the Office of Pesticide Programs' Inerts Ranking Program for "inert" components of pesticide formulations; and methodologies for listing chemicals on the Toxics Release Inventory (TRI). This list is by no means exhaustive.

The characteristics of ranking systems vary, but the majority include explicit consideration of persistence, bioconcentration potential, and aquatic and human toxicity. Persistence is primarily a function of biodegradability for the majority of organic chemicals released to soil and water. This creates a problem for priority-setting exercises, because experimental biodegradation data are typically either lacking entirely or do not exist in a form that can be easily incorporated into automated screening methods.

In responding to this need, we first developed a weight-of-evidence procedure (*1*) for collecting and evaluating available data, due to the considerable variability of biodegradation data. The data and evaluations were then made available on-line in BIODEG, a component of the Environmental Fate Data Base (*2*, *3*) that now contains information on more than 800 discrete organics. The records for each chemical in the BIODEG file constitute a comprehensive assessment of experimental mixed-culture biodegradation data that exist for the chemical; pure culture data are not included because they offer little insight into environmental biodegradation rates. Each test result, whether it reflects biochemical oxygen demand (BOD), CO_2 production, loss of parent, or something else, is assigned a qualitative descriptor code such as BR (biodegrades rapidly) or BSA (biodegrades slowly even with acclimation). Aspects of biodegradation such as acclimation, microbial toxicity and temperature are considered in the evaluation process. A reliability code (3, one test available; 2, two tests available; and 1, three or more consistent tests available), which reflects the amount and consistency of the available data, is also assigned for each biodegradation summary code. There are summary evaluation codes for overall aerobic biodegradation, aerobic biodegradation in screening tests, biological treatment simulations, grab sample tests with soil or water, and field studies.

Subsequently, the summary evaluation codes for overall aerobic biodegradation were used to develop two models for predicting aerobic biodegradability from chemical substructures (*4*). These models, based on multiple linear and nonlinear regression against counts of 35 preselected substructures, calculated the probability of rapid biodegradation and successfully classified as rapidly or not rapidly biodegradable 90% of 264 chemicals in the training set and 27 chemicals in an independent validation set. Klopman et al. (*5*) have also used this data set to develop a predictive model based on computer-automated structure evaluation (CASE) methodology, and Gombar and Enslein (*6*) have described models for subsets (aliphatic and aromatic chemicals) of the BIODEG data.

Although the above models are based on carefully evaluated experimental data, their capabilities are limited to classification. To provide a consistent set of data for quantitative modeling and to determine the feasibility of a biodegradation expert system, we conducted a survey in 1986 in which 22 biodegradation experts were asked to estimate rates and products of degradation for 50 organic chemicals (*7*). A screening-level model for predicting aerobic biodegradability was developed from the survey data (*8*), but the usefulness of that data set was limited by its small size.

In this paper, we describe four new screening-level biodegradability models. Two of these represent enhancements to our previously described (*4*) linear and nonlinear BIODEG models; i.e., those based on experi-

† U. S. Environmental Protection Agency.
‡ Syracuse Research Corp.

0013-936X/94/0928-0459$04.50/0

mental data. The other two models are based on data from a new and greatly expanded survey, in which a panel of 17 experts estimated rates of primary (loss of parent chemical identity) and ultimate (essentially, conversion to CO2 and water) degradation under aerobic conditions in aquatic environments for 200 chemicals. These models permit semi-quantitative prediction of aquatic biodegradation rates. The independent variables for all four models are a revised set of 36 chemical substructures from the original linear and nonlinear models (*4*) plus molecular weight. With the addition of molecular weight, predictions are possible for all chemicals even if they do not contain any of the 36 structural fragments. More importantly, the successful fitting of a single set of chemical substructures to both the evaluated (BIODEG) and the survey data sets affirms the importance of these substructures in estimating biodegradability.

Methods

Biodegradation Database. Summary evaluation codes for overall aerobic biodegradation, used in the development of the linear and nonlinear BIODEG models, were retrieved from BIODEG, the Evaluated Biodegradation Database. The design and development of this file were described in detail in a previous publication (*1*) and briefly in the Introduction. BIODEG is a component of the Environmental Fate Data Base (EFDB), available on-line or in a PC-compatible format from Syracuse Research Corp. (contact P. H. Howard for details).

Linear and Nonlinear Models. The basic approach in the development of the linear and nonlinear BIODEG models has been described (*4*). For this exercise, the training set and the independent validation set from the earlier work were combined to yield a new training set of 295 chemicals. Because the approach had already been validated, it was not considered necessary to keep a separate validation set. This data set consisted of 186 chemicals that received summary evaluations of "biodegrades rapidly" and 109 chemicals designated "does not biodegrade rapidly". An indicator variable was formed with chemicals in the rapid biodegradation category being assigned a value of 1 and chemicals in the slow biodegradation category being assigned a value of 0. The indicator variable was then used as the dependent variable in multiple linear and nonlinear regressions against 37 independent variables. With this definition of the dependent variable, a regression model estimates the probability that a chemical is in the "biodegrades rapidly" group.

In our previously described (*4*) linear and nonlinear models, counts of 35 structural fragments (i.e., the number of times a substructure occurs in the molecule) constituted the independent variables. For this study, several changes were made in the set of independent variables used in the regression analyses. Two new fragments (-CF_3 and unsubstituted phenyl group, -C_6H_5) were added, and three fragments were redefined. The latter include the quaternary carbon and the tertiary alcohol fragments of the earlier models (*4*), now eliminated and replaced with a single fragment (carbon with four single bonds and no hydrogens), and the unsubstituted linear alkyl chain ≥C4, which now can be used only if it is terminal (i.e., -CH_2-$CH_2CH_2CH_3$). Finally, molecular weight was added as a continuous variable, since it is well-known that as molecular size increases, biodegradability generally decreases (*9*, *10*). In general, atoms were used only once; that is, if an atom is part of one fragment, it cannot be part of another. Table 1 lists these fragments and their regression-derived coefficients.

The linear model was defined as

$$Y_j = a_0 + a_1 f_1 + a_2 f_2 + \ldots + a_{36} f_{36} + a_m M_w + e_j \quad (1)$$

where Y_j is the probability that chemical j will biodegrade fast, or for the survey models, the primary or ultimate biodegradation rate, f_n is the number of nth substructure in jth chemical, a_0 is the intercept, a_n is the regression coefficient for nth substructure, M_w is the molecular weight, a_m is the regression coefficient for M_w, and e_j is the error term (mean value is zero). Regression coefficients were estimated by the method of least squares, using the REG procedure of the PC version of the Statistical Analysis System (SAS Institute, Cary, NC). Although the assumption of homogeneous variance does not hold whenever the dependent variable is defined as above, the least-squares method still results in unbiased estimates.

The logistic equation was used as the basis for the nonlinear model. This model

$$Y_j = \frac{\exp(a_0 + a_1 f_1 + a_2 f_2 + \ldots a_{36} f_{36} + a_m M_w)}{1 + \exp(a_0 + a_1 f_1 + a_2 f_2 + \ldots a_{36} f_{36} + a_m M_w)} \quad (2)$$

estimates the probabilities near 0.0 whenever the linear combination in the exponent takes large negative values; near 0.5 whenever that linear combination is near 0.0; and close to 1.0 whenever the linear combination takes a large positive value. The maximum likelihood method was used for estimating the coefficients for this model rather than the method of least squares, because the model is not a linear function of the unknown coefficients. The estimates were obtained by using the CATMOD procedure of PC-SAS.

For each of the estimated regression coefficients, a standard error was computed as well as a test statistic for evaluating the hypothesis that the true population value is 0.0. The test statistic followed an asymptotic χ^2 distribution in the case of the maximum likelihood estimates (nonlinear model) and an F distribution for the least-squares estimates (linear model). A p value was also calculated for each of the test statistics. These p values and statistics are not included in Table 1 to preserve clarity, but are available from the authors.

The standard errors and the test statistics (or their p values) were used only as an approximate indication of the contribution of a particular fragment rather than as a basis for eliminating the fragment from the model. We took this approach because the objective was not to determine the most parsimonious subset of fragments for predicting biodegradation status but to keep the model as broadly applicable as possible. As a result, there are collinearities among some of the fragments that could affect the accuracy of some of the p values computed for the test statistics.

Biodegradation Survey. Information relating to the purpose, design and implementation of an earlier survey of expert knowledge has been published (*7*). For this study, we developed a larger database of biodegradability estimates by conducting a second survey in which 17 experts evaluated 200 organic chemicals (there were 50 chemicals in the first survey). Each expert rated the primary and

Table 1. Structural Fragments and Coefficients

fragment or parameter	BIODEG models freq[a]	BIODEG models linear coeff	BIODEG models nonlinear coeff	survey models freq[a]	survey models primary coeff	survey models ultimate coeff
equation constant		0.748	3.01		3.848	3.199
M_w	295	−0.000476	−0.0142	200	−0.00144	−0.00221
unsubstituted aromatic (≤3 rings)	2	0.319	7.191	1	−0.343	−0.586
phosphate ester	5	0.314	44.409	6	0.465	0.154
cyanide/nitrile (C≡N)	5	0.307	4.644	11	−0.065	−0.082
aldehyde (CHO)	4	0.285	7.180	5	0.197	0.022
amide (C(=O)N or C(=S)N)	9	0.210	2.691	13	0.205	−0.054
aromatic (C(=O)OH)	24	0.177	2.422	6	0.0078	0.088
ester (C(=O)OC)	23	0.174	4.080	25	0.229	0.140
aliphatic OH	34	0.159	1.118	18	0.129	0.160
aliphatic NH_2 or NH	13	0.154	1.110	7	0.043	0.024
aromatic ether	11	0.132	2.248	11	0.077	−0.058
unsubstituted phenyl group (C_6H_5)	25	0.128	1.799	22	0.0049	0.022
aromatic OH	46	0.116	0.909	21	0.040	0.056
linear C4 terminal alkyl ($CH_2CH_2CH_2CH_3$)	44	0.108	1.844	26	0.269	0.298
aliphatic sulfonic acid or salt	4	0.108	6.833	4	0.177	0.193
carbamate	4	0.080	1.009	6	0.194	−0.047
aliphatic (C(=O)OH)	33	0.073	0.643	10	0.386	0.365
alkyl substituent on aromatic ring	36	0.055	0.577	36	−0.069	−0.075
triazine ring	5	0.0095	−5.725	4	−0.058	−0.246
ketone (CC(=O)C)	12	0.0068	−0.453	10	−0.022	−0.023
aromatic F	1	−0.810	−10.532	1	0.135	−0.407
aromatic I	2	−0.759	−10.003	2	−0.127	−0.045
polycyclic aromatic hydrocarbon (≥4 rings)	6	−0.657	−10.164	2	−0.702	−0.799
N-nitroso (NN=O)	4	−0.525	−3.259	1	0.019	−0.385
trifluoromethyl (CF_3)	1	−0.520	−5.670	2	−0.274	−0.513
aliphatic ether	11	−0.347	−3.429	16	−0.0097	−0.0087
aromatic NO_2	14	−0.305	−2.509	13	−0.108	−0.170
azo group (N=N)	2	−0.242	−8.219	3	−0.053	−0.300
aromatic NH_2 or NH	32	−0.234	−1.907	23	−0.108	−0.135
aromatic sulfonic acid or salt	11	−0.224	−1.028	8	0.022	0.142
tertiary amine	10	−0.205	−2.223	10	−0.288	−0.255
carbon with 4 single bonds and no H	9	−0.184	−1.723	32	−0.153	−0.212
aromatic Cl	40	−0.182	−2.016	27	−0.165	−0.207
pyridine ring	18	−0.155	−1.638	8	−0.019	−0.214
aliphatic Cl	12	−0.111	−1.853	14	−0.101	−0.173
aromatic Br	5	−0.110	−1.678	4	−0.154	−0.136
aliphatic Br	5	−0.046	−4.443	2	0.035	0.029

[a] Number of compounds in the training set containing the fragment.

ultimate biodegradability of each chemical on a semi-quantitative scale, which used the terms hours, days, weeks, months, and longer than months to indicate the approximate time they thought would be required for the process to proceed to completion. As the measure of central tendency, we calculated an arithmetic mean score for each chemical after assigning numerical scores to the individual responses as follows: 5 = hours; 4 = days; 3 = weeks; 2 = months; 1 = longer. The total number of responses for each chemical often exceeded 17, since many experts indicated a range of time by marking more than one term.

The 200 survey chemicals covered a very wide range of structure and molecular weight, and the majority were multifunctional. In general, chemicals were selected to be included in the survey for the specific purpose of testing hypotheses regarding the effects of certain substructures on estimated biodegradability. Some examples follow. To explore postulated negative influences on estimated biodegradability, 50 of the 200 chemicals were halogenated, 17 had nitro groups, 18 had quaternary carbon atoms (defined as four single bonds to non-hydrogen atoms), 20 had three or more fused rings, and 35 had nitrogen-containing heterocycles of various types. With respect to expected positive influences on estimated biodegradability, 56 chemicals were biologically hydrolyzable or postulated to be so, and 15 chemicals had unsubstituted linear alkyl chains of C4 or larger. Of the 200 chemicals in the survey and 295 in the experimental (BIODEG) data set, only 20 were common to both sets.

Survey Models. Multiple linear regressions were performed using the mean scores for primary and ultimate biodegradation as dependent variables. The independent variables were the same as those used in the linear and nonlinear BIODEG models just described; i.e., counts of 36 structural fragments plus molecular weight. The fragments and regression-derived coefficients are listed in Table 1. Regression coefficients were estimated by the method of least squares, using the REG procedure of PC-SAS. Primary or ultimate biodegradability is calculated for any chemical by summing, for all the fragments present in the chemical, the number of times (if any) each fragment occurs times its coefficient, and then adding the summation to a constant that was determined for the entire training set, plus the product of the chemical's molecular weight and the M_w coefficient. A file that lists the 200 survey chemicals, the predicted primary and ultimate biodegradation scores, and the chemicals' CAS registry numbers is available from the authors.

Results

Biodegradation Survey. Table 2 contains summary statistics for the 200 survey chemicals and the experts' responses. Ethylene glycol diacetate was judged to be the

Table 2. Summary Statistics for Survey Data

		minimum		maximum	
parameter	mean	score	chemical	score	chemical
primary	3.52	2.37	pentabromoethylbenzene	4.57	ethylene glycol diacetate
SD_{pri}	0.84	0.51	ethylene glycol diacetate	1.28	Vat Blue 4; ethylenediaminetetrakis(methylphosphonic acid)
ultimate	2.60	1.44	pentabromoethylbenzene	3.89	ethylene glycol diacetate
SD_{ult}	0.83	0.58	ethylene glycol diacetate; picloram; pentabromoethylbenzene	1.15	maleic hydrazide
pri-ult[a]	0.92	0.43	ϵ-caprolactone	1.75	dacthal
M_w	228.6	53.1	acrylonitrile	697.6	tris-2,3-dibromopropylphosphate

[a] Primary degradation score minus ultimate degradation score for the same chemical.

most easily degraded chemical and pentabromoethylbenzene the least degradable for both primary and ultimate degradation. The highest and lowest possible mean scores are 5 and 1, respectively, but no such value was observed for any of the 200 chemicals. This would have required the unanimous judgment of all 17 experts that biodegradation would occur either in hours (=5, by definition) or longer than months (=1, by definition). Using the standard deviations of the responses for a given chemical as a measure of agreement or disagreement, unanimity of judgment was also greatest for these two chemicals. In contrast, the largest standard deviations were observed for Vat Blue 4 and ethylenediaminetetrakis(methylphosphonic acid) (primary) and maleic hydrazide (ultimate). On the average, scores for primary and ultimate degradation for each chemical differed by almost 1 unit (0.92) relative to the ordinal scale used to assign scores to the experts' estimates. The largest difference (1.75) was observed for dacthal, a tetrachlorinated herbicide with two ester functions that were considered to be relatively easily hydrolyzed.

Biodegradation Models. Coefficients fitted by the regressions for all four models are listed in Table 1. With the BIODEG models, the probability of rapid biodegradation can be predicted by using the linear or nonlinear coefficients from Table 1 and either eq 1 or eq 2, respectively. With the survey models, biodegradability can be predicted using the coefficients for primary or ultimate biodegradation in Table 1 and eq 1. To illustrate a typical estimation of biodegradability, we will explain the calculations necessary for predicting the ultimate biodegradability of *o*-phenylphenol (M_w = 170) using the survey model. Using Table 1 and eq 1, we have

$$Y_j = \text{equation constant} + (-0.00221)\,(M_w) + (0.022)\,(\text{one unsubstituted phenyl group, } C_6H_5) + (0.056)\,(\text{one aromatic OH group}) = 3.199 + (-0.3757) + 0.022 + 0.056 = 2.90$$

The mean score for this chemical from the biodegradation experts was 3.08. The integer 3 corresponded to "weeks" in the tabulation of the individual survey responses.

Performance of the BIODEG models in classifying chemicals in their training set is summarized in Table 3. Each model classified chemicals in the training set with about 90% accuracy overall, but results were slightly better for the nonlinear model. With either model, rapidly degraded chemicals were classified more accurately than slowly degraded chemicals. The distributions of residuals from the survey models are shown in Figure 1, along with the R^2 values and percentages of residuals ≤ ±0.1, ±0.3, and ±0.5 (absolute value). The mean residuals (absolute value) for primary and ultimate degradation for all 200 chemicals were 0.173 and 0.206, respectively. There were eight chemicals with residuals ≥0.6 in absolute value, and these are listed in Table 4.

Table 3. Performance of Biodegradability Models in Classifying Chemicals in Their Respective Training Sets

	BIODEG models		survey models	
parameter	linear[a]	nonlinear[a]	primary[b]	ultimate[c]
total correct	264/295	275/295	165/200	167/200
% correct total	89.5	93.2	82.5	83.5
% correct, fast biodegradation	97.3 (181/186)	97.3 (181/186)	84.9 (101/119)	93.5 (101/108)
% correct, slow biodegradation	76.1 (83/109)	86.2 (94/109)	79.0 (64/81)	71.7 (66/92)

[a] Fast biodegradation is defined as a predicted probability >0.5 for being classified as a BR (*Biodegrades Rapidly*). [b] Fast biodegradation is defined as a biodegradability score ≥3.5 [c] Fast biodegradation is defined as a biodegradability score >2.5.

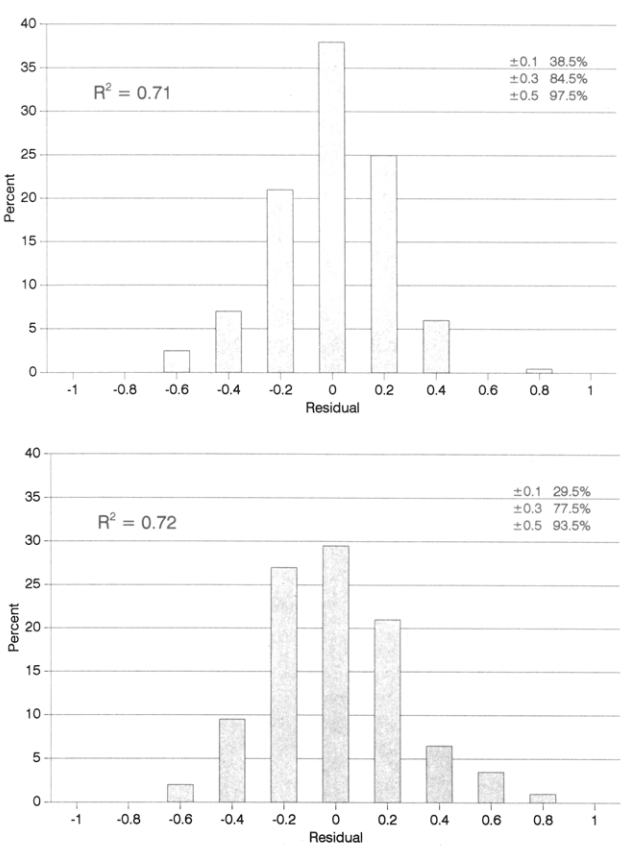

Figure 1. Distribution of residuals from biodegradation survey models. (A, top) Primary degradation model. (B, bottom) Ultimate degradation model.

Table 4. Poorly Predicted Survey Chemicals[a]

chemical	predicted: experts	predicted: model	residual
primary degradation			
silvex	2.82	3.43	−0.60
2,2,4,4,6,8,8-heptamethylnonane	2.43	3.06	−0.63
di-*tert*-butyldicarbonate	4.05	3.23	0.82
ultimate degradation			
ϵ-caprolactone	3.70	3.09	0.61
n-decanal	3.80	3.17	0.63
11-cyanoundecanoic acid	3.68	3.01	0.67
hexachlorophene	1.77	1.10	0.67
ethylene glycol diacetate	3.89	3.16	0.73
di-*tert*-butyldicarbonate	3.18	2.29	0.89

[a] All survey chemicals with residuals (experts predicted minus model predicted) ≥ |±0.6|.

The primary and ultimate survey models calculate a biodegradability score rather than a probability of rapid biodegradation, with integers corresponding to the descriptors (hours, days, weeks, etc.) used in the survey. This makes direct comparison of performance for the two types of models (BIODEG vs survey) difficult. One way to enable such a comparison is to evaluate performance of the survey models in classifying chemicals in the survey training set. To accomplish this, we defined rapid primary degradation as a biodegradability score of ≥3.5, corresponding to the descriptor days–weeks. For ultimate degradation, we defined rapid biodegradation as a biodegradability score of >2.5, which corresponds to weeks–months. Using these criteria, performance of the primary and ultimate survey models as classifiers (Table 3) was somewhat below that observed for the BIODEG models, with slightly more than 80% of the survey chemicals classified correctly by each model. As was true for the BIODEG models, rapidly degraded chemicals were more accurately classified than slowly degraded chemicals.

Accuracy of Experts' Estimates. To assess directly the accuracy of the experts' biodegradability estimates, we retrieved and reviewed experimental data for all survey chemicals that also had water grab sample data in the BIODEG database (*11–34*). Our assessments of the literature data for these 13 chemicals with respect to the approximate length of time required for complete degradation (defined here as six half-lives) are summarized in Table 5. For comparison, the mean survey scores, our interpretation of them relative to the biodegradability descriptors used in the survey, and the calculated (model) values are also presented. It is evident that the experts' estimates of biodegradability in aquatic environments were generally consistent with existing experimental data. Biodegradability scores calculated using the survey models (last column of Table 5) also tracked well with the experts' estimates, with mean residuals (absolute value) for these chemicals of 0.16 for primary degradation ($n = 6$) and 0.20 for ultimate degradation ($n = 7$).

Discussion

Our results demonstrate that a single set of chemical substructures and molecular weight allow an acceptably accurate prediction of both experimentally determined biodegradability, as reflected in the BIODEG evaluation codes, and experts' estimates of primary and ultimate biodegradation rates. This finding lends credence to the notion that these factors are important determinants of biodegradability. It also validates expert judgment, as reflected in the survey data and the models based on it. The models thus derived have been encoded in an IBM-compatible PC program (Biodegradation Probability Program, available from Syracuse Research Corp.) that predicts the probability of rapid biodegradation and the time required for primary and ultimate degradation. Only the chemical's SMILES notation (*35*) or CAS registry number is required as input.

Expert judgment is also validated by direct comparison of survey scores to grab sample biodegradation data (Table 5). Chlorothalonil seems to be an exception, because the experts predicted that primary degradation would occur in weeks to months, whereas the experimental data suggest days to weeks. But this is an unusual situation since, according to Davies (*36*), primary degradation is much faster than anticipated because the nitriles in chlorothalonil direct nucleophilic attack to the 4 and 6 positions on the ring. A lesson to be learned from this is that even the collective wisdom of experts may be in error when applied to specific chemical structures and should not be considered a substitute for adequate testing.

Our previous models (*4*) included a library of 35 structural fragments in order to ensure that the models be as broadly applicable as possible. However, no predictions could be made for chemicals that did not contain any of these substructures. With the inclusion of the molecular weight parameter no structures are excluded, although the reliability of predictions based on molecular weight alone is probably fairly low except for chemicals with very low or very high molecular weights. Among the five new or redefined substructures listed in Table 1, at least two also have clear mechanistic significance, since unsubstituted terminal alkyl groups (represented by the linear C4 fragment) and unsubstituted phenyl groups both provide sites for the initiation of well-known biodegradation pathways (*20, 37*).

The signs of the coefficients for the fragments and parameters listed in Table 1 are generally consistent with commonly accepted generalizations regarding effects of chemical structure on biodegradability. For example, ester, alcohol, and carboxylic acid groups usually enhance biodegradability (*9, 10*), and all have positive signs in all four models. On the other hand, halogens, nitro groups, and quaternary carbons are assumed to make a chemical more resistant to degradation, and all have negative signs.

However, there are also a number of fragments for which the signs are not the same in the four models. In some cases the coefficients are small for all four models, which suggests that the fragment may not be very important in determining biodegradability. An example is the ketone fragment. For other fragments, it may be observed that the signs of the coefficients are often inconsistent where the BIODEG and survey training sets contained only a few chemicals with that fragment. Confidence in those coefficients is therefore low, but could be raised by additional testing. Examples of fragments for which few data are available include the aromatic F, N-nitroso, and aliphatic Br fragments.

Another phenomenon is that the primary and ultimate coefficients (survey models) are sometimes quite different in magnitude. This is to be expected. In the case of the aldehyde, amide, and carbamate fragments, for example, this suggests that these fragments are considered by experts to be likely sites of initial attack, but without major

Table 5. Comparison of Survey Data to Measured Biodegradability for Survey Chemicals with Water Grab Sample Data

chemical	survey		literature				
	score[a]	int[b]	int[c]	n^d	U or P[e]	ref	model[f]
eicosane	4.19	≤d	d–wk; <wk; <mo	3	P	11–13	3.98
dimethylformamide	4.09	≤d	d	1	P	14	3.94
cumene	3.68	d–wk	wk	1	P	15	3.61
propanil	3.61	d–wk	wk; wk; wk; >mo; >mo	5	P	16–18	3.40
Acid Orange 6	3.45	d–wk	>d	1	P	19	3.47
chlorothalonil	2.39	wk–mo	d; d–wk; d–wk; wk; wk	5	P	20	2.68
acrylonitrile	3.27	d–wk	wk; wk	2	U	21	3.00
o-phenylphenol	3.08	d–wk	d–wk	1	U	22	2.90
diphenyl ether	2.79	wk–mo	mo	1	U	23	2.81
di-2-cyanoethyl ether	2.79	wk–mo	wk–mo; mo	2	U	21, 24	2.76
tert-butylbenzene	2.62	wk–mo	mo	1	U	23	2.72
hexachlorophene	1.77	≥mo	>wk; >mo	2	U	25	1.10
benzanthracene	1.76	≥mo	>d; >d; wk; >wk; >wk; wk–mo; mo; >mo; >mo; >mo; >mo; >mo; >mo	13	U	26–34	1.89

[a] Observed biodegradability score from survey; value given is for either ultimate or primary degradation, depending on the type of literature data. [b] Interpretation of the survey score according to the following scheme (d = days; wk = weeks; mo = months): ≥4 = ≤d; <4 ≥ 3 = d–wk; <3 ≥ 2 = wk–mo; <2 = ≥mo. [c] Interpretation of each study in terms of the approximate time required for complete degradation, defined as six half-lives for primary degradation and 60–70% of theoretical for ultimate degradation, in natural water grab samples. [d] Number of studies. [e] U = ultimate; P = primary. [f] Predicted primary or ultimate degradation using the appropriate survey model.

influence on rates of ultimate degradation. Conversely, triazine rings, azo bonds, and pyridine rings, for example, seem to be viewed as negative for ultimate but not necessarily primary degradation.

Close inspection of the residuals from the survey models suggests several ways in which these models could be improved. In some cases the solution is obvious. For example, the experts assumed that di-*tert*-butyldicarbonate (Table 4) would be readily hydrolyzed, but our models lack a carbonate fragment. Alkyl chains represent a more subtle problem. Thirteen compounds in the survey had linear alkyl chains of C9 or greater, and 10 of these had positive residuals for both primary and ultimate degradation. This suggests that long alkyl chains were viewed by the experts as having a positive impact on biodegradability, but that the linear C4 terminal alkyl fragment does not adequately account for this effect. On the other hand, compounds with cycloalkane rings (six survey compounds) and aromatic rings with two nitrogens (i.e., pyrazines, pyrimidines, and pyridazines; six survey compounds) generally had negative residuals, suggesting that these groups were considered to increase resistance to biodegradation. It should be noted that both single- and three-nitrogen aromatics (i.e., pyridines and triazines) are already represented by fragments in our models, and all but one of the coefficients are negative. In spite of this, we did not add new fragments for cycloalkane or two-nitrogen heteroaromatic rings, because our approach was to require that such chemicals also be adequately represented in the experimental data (BIODEG) training set, and they were not. Additional testing will probably be required to establish an adequate database of measured values. This kind of analysis shows how the models may be used to identify chemical classes in need of testing.

There is no doubt that the fragment constant approach to biodegradability modeling that we have taken is somewhat simplistic and does not, for example, take into account the possible interactions among fragments in multifunctional molecules. Nevertheless, the models described above meet our goal of providing quantitative or semi-quantitative estimates of biodegradation rate for use in chemical ranking schemes, in addition to estimates of probability of rapid biodegradation.

Literature Cited

(1) Howard, P. H.; Hueber, A. E.; Boethling, R. S. *Environ. Toxicol. Chem.* **1987**, *6*, 1.

(2) Howard, P. H.; Sage, G. W.; LaMacchia, A.; Colb, A. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 38.

(3) Howard, P. H.; Hueber, A. E.; Mulesky, B. C.; Crisman, J. S.; Meylan, W.; Crosbie, E.; Gray, D. A.; Sage, G. W.; Howard, K. P.; LaMacchia, A.; Boethling, R.; Troast, R. *Environ. Toxicol. Chem.* **1986**, *5*, 977.

(4) Howard, P. H.; Boethling, R. S.; Stiteler, W. M.; Meylan, W. M.; Hueber, A. E.; Beauman, J. A.; Larosche, M. E. *Environ. Toxicol. Chem.* **1992**, *11*, 593.

(5) Klopman, G.; Balthasar, D. M.; Rosenkranz, H. S. *Environ. Toxicol. Chem.* **1993**, *12*, 231.

(6) Gombar, V. K.; Enslein, K. In *Applied Multivariate Analysis in SAR and Environmental Studies*; Devillers, J., Karcher, W., Eds.; Kluwer: Boston, MA, 1991; pp 377–414.

(7) Boethling, R. S.; Gregg, B.; Frederick, R.; Gabel, N. W.; Campbell, S. E.; Sabljic, A. *Ecotoxicol. Environ. Saf.* **1989**, *18*, 252.

(8) Boethling, R. S.; Sabljic, A. *Environ. Sci. Technol.* **1989**, *23*, 672.

(9) Alexander, M. *Biotechnol. Bioeng.* **1973**, *15*, 611.

(10) Scow, K. M. In *Handbook of Chemical Property Estimation Methods*; Lyman, W.J., Reehl, W. F., Rosenblatt, D. H., Eds.; McGraw-Hill: New York, 1982; pp 9-1–9-85.

(11) Bertrand, J. C.; Esteves, J. L.; Mulyono, M.; Mille, G. *Chemosphere* **1986**, *15*, 205.

(12) Matsumoto, G. *Water Res.* **1983**, *17*, 1803.

(13) Walker, J. D.; Calomiris, J. J.; Herbert, T. L.; Colwell, R. R. *Mar. Biol.* **1976**, *34*, 1.

(14) Dojlido, J. R. *Investigations of Biodegradability and Toxicity of Organic Compounds, Final Report 1975–1979*; Environmental Protection Agency: Cincinnati, OH, 1979; EPA 600/2-79-163.

(15) Walker, J. D.; Colwell, R. R. *Prog. Water Technol.* **1975**, *7*, 783.

(16) Call, D. J.; Brooke; L. T.; Kent, R. J.; Knuth, M. C.; Anderson, C.; Moriarty, C. *Arch. Environ. Contam. Toxicol.* **1983**, *12*, 175.

(17) El-Dib, M. A.; Aly, O. A. *Water Res.* **1976**, *10*, 1055.

(18) Paris, D. F.; Rogers, J. E. *Appl. Environ. Microbiol.* **1986**, *51*, 221.

(19) Michaels, G. B; Lewis, D. L. *Environ. Toxicol. Chem.* **1986**, *5*, 161.

(20) Gibson, D. T.; Subramanian, V. In *Microbial Degradation of Organic Compounds*; Gibson, D. T., Ed.; Dekker: New York, 1984; pp 181–252.

(21) Ludzack, F. J.; Schaffer, R. B.; Bloomhuff, R. N.; Ettinger, M. B. *Proc. 13th Ind. Waste Conf., Eng. Ext. Bull., Purdue Univ., Engr. Ext. Ser. PP* **1958**, *13th*, 297.
(22) Gonsior, S. J.; Bailey, R. E.; Rhinehart, W. L.; Spence, M. W. *J. Agric. Food Chem.* **1984**, *32*, 593.
(23) Ludzack, F. J.; Ettinger, M. B. *Eng. Ext. Ser. (Purdue Univ.)* **1963**, *no. 115*, 278.
(24) Cherry, A. B.; Gabaccia, A. J.; Senn, H. W. *Sewage Ind. Wastes* **1956**, *28*, 1137.
(25) Lee, R. F.; Ryan, C. In *Microbial Degradation of Pollutants in Marine Environments*; Bourquin, A. W., Pritchard, P. H., Eds.; Environmental Protection Agency: Gulf Breeze, FL, 1979; EPA-600/9-79-012; pp 443–450.
(26) Gardner, W. S.; Lee, R. F.; Tenore, K. R.; Smith, L. W. *Water, Air, Soil Pollut.* **1978**, *11*, 339.
(27) Herbes, S. E.; Schwall, L. R. *Appl. Environ. Microbiol.* **1978**, *35*, 306.
(28) Herbes, S. E.; Southworth, G. R.; Schaeffer, D. L.; Griest, W. G.; Maskarinec, M. P. In *The Scientific Basis of Toxicity Assessment*; Witschi, H., Ed.; Elsevier/North-Holland: New York, 1980; pp 113–128.
(29) Herbes, S. E. *Appl. Environ Microbiol.* **1981**, *41*, 20.
(30) Hinga, K. R.; Pilson, M. E. Q.; Lee, R. F.; Farrington, J. W.; Tjessem, K.; Davis, A.C. *Environ. Sci. Technol.* **1980**, *14*, 1136.
(31) Lee, R. F. *Proc. Oil Spill Conf. (API Publ.)* **1977**, *4284*, 611.
(32) Lee, R. F.; Gardner, W. S.; Anderson, J. W.; Blaylock, J. W.; Barwell-Clarke, J. *Environ. Sci. Technol.* **1978**, *12*, 832.
(33) Lee, R. F.; Ryan, C. *Can. J. Fish Aquat. Sci.* **1983**, *40*, 86.
(34) Roubal, G.; Atlas, R. M. *Appl. Environ. Microbiol.* **1978**, *35*, 897.
(35) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(36) Davies, P. E. *Bull. Environ. Contam. Toxicol.* **1988**, *40*, 405.
(37) Britton, L. N. In *Microbial Degradation of Organic Compounds*; Gibson, D.T., Ed.; Dekker: New York, 1984; pp 89–129.

Received for review June 3, 1993. Revised manuscript received October 25, 1993. Accepted November 1, 1993.[⊛]

[⊛] Abstract published in *Advance ACS Abstracts*, December 15, 1993.