ELSEVIER

Journal of Hazardous Materials B84 (2001) 189–215

Journal of Hazardous Materials

www.elsevier.nl/locate/jhazmat

A review of structure-based biodegradation estimation methods

John W. Raymond [a], Tony N. Rogers [b,*], David R. Shonnard [b], Andrew A. Kline [b,1]

[a] *Department of Biophysics, University of Michigan, 930 N. University, Ann Arbor, Michigan, MI 48109-1055, USA*

[b] *Department of Chemical Engineering, Michigan Technological University (MTU), Houghton, Michigan, MI 49931-1295, USA*

Received 15 January 1999; received in revised form 15 March 2001; accepted 15 March 2001

Abstract

Biodegradation, being the principal abatement process in the environment, is the most important parameter influencing the toxicity, persistence, and ultimate fate in aquatic and terrestrial ecosystems. Biodegradation of an organic chemical in natural systems may be classified as primary (alteration of molecular integrity), ultimate (complete mineralization; i.e. conversion to inorganic compounds and/or normal metabolic processes), or acceptable (toxicity ameliorated). Most of the biodegradation correlations presented in the literature focus on the characterization of primary or ultimate, aerobic degradation.

The US Environmental Protection Agency (USEPA) is charged with determining the risks associated with the thousands of chemicals employed in commerce, an effort that is being facilitated through much research aimed at reliable structure-activity relationships (SAR) to predict biodegradation of chemicals in natural systems. To this end, models are needed to understand the mechanisms of biodegradation, to classify chemicals according to relative biodegradability, and to develop reliable biodegradation estimation methods for new chemicals. Frequently, published correlations associating molecular structure to biodegradation will attempt to quantify the degradability of a limited set of homologous chemicals. These correlations have been dubbed quantitative structure biodegradability relationships (QSBRs). More scarce and valuable to researchers are those models that predict the biodegradability of compounds possessing a wide variety of chemical structures. The latter may use any of several techniques and molecular descriptors to correlate biodegradability: QSBRs, pattern recognition, discriminant analysis, and principle component analysis (PCA), to name several. Generally, models either predict the propensity of a chemical to biodegrade using

*Corresponding author. Tel.: +1-906-487-2210; fax: +1-906-487-3213.

E-mail address: tnrogers@mtu.edu (T.N. Rogers).

[1] Present address : Department of Paper and Printing Science and Engineering, Western Michigan University, Kalamazoo, MI 49008, USA.

PII: S0304-3894(01)00207-2

Boolean-type logic (i.e. whether a chemical will "readily biodegrade" or not), or else they quantify the degree of biodegradation by providing information such as rate constants. Such quantitative predictions of biodegradability come in a diversity of parameters, including half-lives, various biodegradation rates and rates constants, theoretical oxygen demand (ThOD), biological oxygen demand (BOD), and others.

In this paper, after describing the advantages and disadvantages of the various biodegradation estimation methods found in the literature, the best models are compared to conclude which provide the greatest utility for determining the biodegradability of chemicals with widely varying structures. The group contribution technique presented by Boethling et al. [Environmen. Sci. Technol. 28 (1994) 459] appears to be the most advantageous for use in broad screening for tendency to biodegrade. The model is simple to use, calculating a probability of biodegrading ranging from 0 (none) to 1 (certain), and has proven to be accurate for a wide range of chemical structures, as established by the large, high-quality data set (BIODEG evaluated biodegradation database, Syracuse Research Corporation, Merrill Lane, Syracuse, NY 13210) used to develop this correlation. The authors, therefore, recommend the method of Boethling et al. [Environ. Sci. Technol. 28 (1994) 459] for the initial screening of chemicals to aid in determining whether additional information is necessary to establish relative biodegradability. For readers with applications requiring more quantitative results, such as biodegradation rate constants, enough model details are presented in this paper to allow the reader to pick a suitable correlation, although the reader is cautioned to consult the original, primary reference for the complete method description, equations, and limitations.

Keywords: Quantitative structure-activity relationship (QSAR); Quantitative structure biodegradability relationship (QSBR); Biodegradability estimation methods; Biodegradation; Mineralization; Biological oxidation; Biodegradation rate constant

1. Introduction

For most organic chemicals, biodegradation is the principal abatement process in the environment [2]; hence, biodegradation is the most important parameter influencing the behavior and associated toxicity in aquatic and terrestrial ecosystems [3]. The processes of biodegradation have been enumerated as follows [4]:

1.1. Primary biodegradation

Any biologically induced structural transformation in the parent compound that alters its molecular integrity.

1.2. Ultimate biodegradation

Biological conversion of an organic compound to inorganic compounds and the products associated with normal metabolic processes (mineralization).

1.3. Acceptable biodegradation

Biological degradation of an organic compound to the extent that toxicity or other undesirable characteristics are ameliorated.

Although acceptable biodegradation is the desired goal when determining the effects of a chemical released to environment, it is difficult to determine what is an acceptable level of biodegradation. Biodegradation is dependent upon many factors including temperature, population of microorganisms, degree of acclimation, accessibility of metabolic cofactors (i.e. O_2, nutrients, etc.), cellular transport properties, growth medium, chemical partitioning tendencies, etc. [3–6]. These variables are difficult or impossible to control, and the structure and toxicity of the resultant degradation products are often difficult to assess. Most of the biodegradation correlations presented in the literature focus on the characterization of primary or ultimate, aerobic degradation.

Each year the Environmental Protection Agency (EPA) must review thousands of chemicals in an attempt to determine the possible toxicological effects to the environment and to human health, and the premanufacture notices (PMN) submitted to the EPA for approval often do not contain information regarding the biodegradability of the compound in question leaving the reviewer with little information to render a satisfactory determination of the potential risks of exposure. In addition to the review of proposed chemicals, the EPA must also determine the risks associated with the multitude of chemicals presently employed in commerce [7]. This suggests the need for a method to reliably and conveniently ascertain a semi-quantitative judgment as to the biodegradability of a broad diversity of chemicals with little or no dependence on measured input.

Toward this end, much research has been performed to develop reliable structure-activity relationships (SAR) that can describe and predict the biodegradability of chemicals released to natural systems [8]. The published correlations associating structure and molecular activity to biodegradation typically quantify the degradability of a limited set of homologous chemicals. These correlations have been dubbed quantitative structure biodegradability relationships (QSBR). They commonly employ simple or multiple regression analyses on one or more molecular descriptors to characterize the biodegradability of a specific chemical. The predicted biodegradability is represented by a diversity of parameters, including half-lives, various biodegradation rates and constants, theoretical oxygen demand (ThOD), biological oxygen demand (BOD), etc. The purpose of QSBRs are enumerated as follows:

- Understand mechanisms of biodegradation.
- Classify chemicals according to relative biodegradability.
- Develop reliable biodegradation estimation methods for new compounds.

Published correlations that are able to predict the biodegradability of compounds displaying varying chemical structures are scarce in comparison to the profusion of QSBRs that quantitatively describe homologous series of chemicals. These heterologous models can be categorized into three groups [9]: (1) QSBRs, (2) pattern recognition methods, and (3) discriminant analysis. Pattern recognition models are the most complicated methods for biodegradability determinations and incorporate the use of "artificial intelligence" networks to decipher which chemical substructures contained within a compound may be responsible for biodegradability or persistence based on known microbial metabolic processes that have been programmed into the model. Discriminant analysis methods by contrast are essentially the statistical manipulation of suspected variables associated with biodegradability into groups to allow a discrimination function to ascertain the desired result. Within these three subgroups, models exist that either predict the propensity of a chemical to

biodegrade using Boolean-type logic (readily biodegrades or persists) or quantify the degree of biodegradation by providing information such as rate constants. The heterologous models (characterizing chemicals possessing multiple functional groups) typically include a greater number of molecular descriptors in the analysis than homologous correlations and are derived from larger training sets of chemicals.

The purpose of the following review is to introduce the various published QSBRs, provide an objective comparison of the utility of each correlation, and to describe each model with enough detail to allow the observer to discern how each QSBR is used. Following a description of the advantages and disadvantages of the various models, the models are compared to conclude which model(s) provide the greatest utility when used to determine the biodegradability of chemicals with widely varying structures.

2. Homologous models

Correlations for homologous models are typically represented by a simple linear or quadratic equation that includes one or more molecular descriptors. The molecular descriptors are selected based upon their ability to fit the measured data in the training set by accounting for specific reaction mechanisms, and the form of the equation is determined by means of regression or the method of least squares. The descriptor variables used in QSBRs have been categorized by structure and energetics/interactions of the system [2]. Table 1 represents a summary of the most frequently used QSBR descriptors and functions [2,8].

Table 2 represents an extensive listing of the available correlations for homologous series of chemicals. It is not suggested that the listing provides a comprehensive inven-

Table 1
QSBR descriptor summary

Descriptor	Definition	Comments
σ	Hammet substituent constant	Specifies the electron attracting or repelling effect of substituents. Regarded as an approximate measure of the relative electron density at the center of reaction
$\sigma^{\bullet}$	Taft sigma constant	Also describes electronic effects
K_{ow}	*n*-Octanol–water partition coefficient	Describes the hydrophobicity (i.e. transport through cell wall) of the compound
E_s	Taft steric constant	Classical descriptor for the steric effects of substituents on chemical and biological processes
pK_a	Acid dissociation constant	Describes acidic properties of a chemical
${}^{n}\chi_{m}$	Molecular connectivity indices	Identifies molecular topology. Order (n) is equal to number of C–C bonds. Subscripts (m) refer to the type of fragment (p: path, c: cluster, and pc: path/cluster)
k_{OH}	Alkaline hydrolysis constant	The rate constant for alkaline hydrolysis
γ	Van der Waals radius	Van der Waals radius measurement of a chemical
IR	Infrared peak frequencies and intensities	Measurements of infrared spectroscopy
	Other macroscopic descriptors	Includes molecular weight, abiotic reaction rate, retention times (Rt), etc.

Table 2
Correlations for homologous series of chemicals[a]

Substances	Equations	r (r^2)	n	F:s	Reference
Organic acids	$\%\mathrm{ThOD} = -286.999^4\chi_c + 86.069$	−0.935	10	55.1:16.46	[10]
	$\%\mathrm{ThOD} = -67.158^3\chi_c^v + 96.557$	−0.912	20	88.7:15.43	[11]
Linear and branched acids	$\%\mathrm{ThOD} = -252.507^4\chi_c - 22.048^3\chi_p + 122.303$	−0.980	10	85.6:9.80	[11]
Branched acids	$\%\mathrm{ThOD} = -194.107^4\chi_c + 64.651$	−0.946	10	67.9:10.03	[11]
Acids and alcohols	$\%\mathrm{ThOD} = -161.432^4\chi_c - 27.083^3\chi_p^v + 85.192$	−0.929	24	65.9:11.63	[11]
	$\%\mathrm{ThOD} = -148.734^4\chi_c + 56.678$	−0.850	24	57.2:16.15	[11]
Acyclic ketones	$\log(\%\mathrm{ThOD}) = -0.106\log(K_{ow})^2 + 0.241\log(K_{ow}) + 1.682$	0.99989	10	11.07:0.03	[6]
	$\log(\%\mathrm{ThOD}) = 0.698\log(K_{ow}) - 0.867\log(0.671K_{ow} + 1) + 1.870$	0.99996	10	18.58:0.02	[6]
Alcohols	$\%\mathrm{ThOD} = -34.451^2\chi + 122.765$	−0.871	14	37.6:15.59	[11]
	$\mathrm{BOD}_5 = 1.023 \times 10^3\Delta\delta_{C-O} + 1.504$	0.990	20	:2.54	[12]
	$\%\mathrm{ThOD} = -141.493^4\chi_c - 32.147^3\chi_p^v + 83.613$	−0.951	14	51.7:10.26	[11]
C8–C12 alcohols	$\log(\%\mathrm{ThOD}) = -0.192\log K_{ow} + 2.338$	−0.997	5	579:0.0129	[6]
Aldehydes	$\mathrm{BOD}_5 = 1.607 \times 10^3\Delta\delta_{C=O} - 4.231$	0.990	9	:2.175	[13]
Alkanes	$\mathrm{BOD}_5 = 0.0996\mathrm{ASA} + 0.055$	1.00	12	:0.270	[14]
Aromatic and aliphatic amines	$\mathrm{BOD}_5 = 1.004 \times 10^3\Delta\delta_{C-N} - 0.106$	0.999	15	:1.043	[15]
Anilines	$\log T_{50} = -0.48\mathrm{p}K_a + 2.67$	0.887	17	ND	[16]
Substituted aniline					
o	$\log(v) = -0.30\sigma_o + 1.24$	0.975	4	ND	[2]
m	$\log(v) = -1.53\sigma_m + 1.31$	0.970	3	ND	
p	$\log(v) = -0.78\sigma_p + 1.04$	0.942	5	ND	
Carbamates	$\log(\%\mathrm{D}) = -1.565^4\chi_{pc} + 3.768$	−0.985	7	167.2:0.112	[11]
	$\log(\%\mathrm{D}) = -2.145^4\chi_{pc}^v + 2.765$	−0.981	7	125.3:0.129	[11]
Carboxylic acids	$\mathrm{BOD}_5 = 0.996 \times 10^3\Delta\delta_{C-O} + 3.234$	0.987	40	:4.41	[17,18]
Chlorophenols	$\log T_{50} = -0.68\mathrm{p}K_a + 7.0$	0.977	5	ND	[19]
2,4-D esters	$\log \mathrm{Rc} = 0.816^2\chi^v - 11.928$	0.977	6	82.2:0.185	[11]
	$\log \mathrm{Rc} = 1.198^3\chi_p - 14.378$	0.974	6	73.0:0.195	[11]
Esters	$\mathrm{BOD}_5 = 1.001 \times 10^3\Delta\delta_{C-O} + 2.340$	0.981	19	:3.09	[17]
Ethers	$\log(\%\mathrm{ThOD}) = -0.517^2\chi^v + 2.597$	−0.987	6	149.3:0.076	[11]
	$\log(\%\mathrm{ThOD}) = -0.899^4\chi_{pc} + 1.186$	−0.977	6	84.8:0.100	[11]
	$\mathrm{BOD}_5 = 1.020 \times 10^3\Delta\delta_{C-O} + 1.486$	0.983	14	:2.72	[18]

Table 2 (*Continued*)

Substances	Equations	r (r^2)	n	F:s	Reference
Glycols	$BOD_5 = 0.993 \times 10^3 \Delta\delta_{C-O} + 1.309$	0.994	8	:2.74	[17]
R–X	$BOD_5 = 8.29L - 1.187$	0.976	9	:4.12	[15]
Ketones	$BOD_5 = 1.021 \times 10^3 \Delta\delta_{C=O} + 0.605$	0.989	7	:4.34	[18]
Phenols	$BOD_5 = 0.998 \times 10^3 \Delta\delta_{C-O} + 2.108$	0.983	11	:4.04	[18]
	$\log T_{50} = -0.21 pK_a + 2.0$	0.886	20	ND	[19]
Substituted phenol					
o	$\log(v) = -0.43\sigma_o + 1.70$	0.980	5	ND	[2]
m	$\log(v) = -0.62\sigma_m + 1.72$	0.940	4	ND	
p	$\log(v) = -0.32\sigma_p + 1.65$	0.990	4	ND	
Mean	$\log(v) = 0.32\Sigma\sigma + 1.43$	0.950	7	ND	
Phthalates	$\log(Kb) = -2.09 \log(Rt)^2 + 1.19 \log(Rt) - 1.15$	0.986	5	ND	[20]
	$Rc \times 10^3 = -24.31 \log(K_{ow}) + 394.84$	−0.931	12	65.1:37.48	[11]
	$Rc \times 10^3 = -37.156^2\chi + 547.519$	−0.969	12	151.9:25.52	[11]
	$Rc \times 10^3 = -37.312^2\chi^v + 436.429$	0.968	12	147.8:25.86	[11]
Phthalates and esters	$Rc \times 10^3 = -73.343^4\chi_p - 59.181^3\chi_c^v + 613.022$	−0.975	12	85.8:24.17	[11]
	$Rc \times 10^3 = -73.343^4\chi_p - 59.207^3\chi_c + 643.506$	−0.975	12	85.8:24.173	[11]
Phthalate esters	$Rc \times 10^3 = -0.977\,MW + 532.976$	−0.954	12	100.5:30.90	[11]
Propazamides and esters	$\log k = -2.74 - 1.22\sigma + 0.58\pi$	0.910	40	ND	[21]

[a] r: Residual; n: sample population; F: F-statistic; s: selectivity; R–X: halogenated hydrocarbons.

tory encompassing all available correlations, but the table is intended to provide a ready reference to expediently quantify the biodegradability of certain well defined chemicals when no experimental data are readily available.

3. Heterologous models

An extensive investigation into the existence of correlations for chemicals of varying structure has resulted in the following comprehensive review of all published heterologous models discovered during the preparation of this manuscript. The estimation models are presented by the authors/research group responsible for the publication since several researchers are responsible for methods of varying type. These methods include screening correlations to predict whether a specified chemical is biodegradable or not and relationships used to determine quantifiable rates of biodegradation.

3.1. Dearden and Nicholson: QSBR (atomic charge difference)

This research group also attempted to correlate biodegradability in the form of BOD with calculated parameters [22]. They obtained a screened set of 5-day BOD values for 240 compounds from the US EPA in Duluth, Minnesota. The measured BOD values were normalized by dividing the measure values by the calculated ThOD for each compound. Dearden and Nicholson [15,18,23] subdivided the chemicals contained in the data set into homologous series and attempted to correlate the normalized BOD values with various parameters, including molecular connectivities up to seventh order, K_{ow}, molecular volume, accessible molecular surface area, Sterimol steric parameters, and atomic charges. The authors were able to generate homologous correlations for halogenated hydrocarbons and a series of alkanes using the Sterimol length parameter (L) and accessible molecular surface area (ASA), respectively, but the most notable development in their research focused on the atomic charge parameter, d. Employing regression analysis, they were able to construct a correlation that accurately predicted normalized BOD values for a variety of chemicals based on the charge difference, ignoring sign convention, in the modulus charges ($\Delta\delta_{x-y}$) on the atoms of specified bonds (e.g. C–O, C=O, C–N, etc.) for each chemical structure. Eq. (1) encompasses amines, phenols, aldehydes, carboxylic acids, halogenated hydrocarbons, and amino-acids:

$$\mathrm{BOD} = (1.015 \times 10^3)\Delta\delta_{x-y} + 1.193; \quad n = 79,\ r = 0.993,\ s = 3.459 \tag{1}$$

Although it is possible to calculate the $\Delta\delta_{x-y}$ parameter, it is not readily obtainable. To acquire the value, the authors establish the chemical's XYZ molecular coordinates and then use the coordinates to determine the energy minimization using a molecular mechanics program. Once the energy minimization is obtained, the atomic charge across the key bond is calculated.

The authors further expanded this correlation on two separate occasions in 1987 [15,18] to include additional structural groups. Tables 3 and 4 list the functional groups and associated key bonds characterized by Dearden and Nicholson. Eq. (2) describes the relationship

Table 3
Structural groups and associated key bonds[a]

Structural group	Key bond	Frequency	Structural group	Key bond	Frequency
Alcohols	C–O	19	Glycols	C–O	8
Amines	C–N	15	R–X	C–X	9
Amino acids	C–O	–	Ketones	C=O	7
Aldehydes	C=O	9	Phenols	C–O	11
Carboxylic acids	C–O	40	Sugars		–
Esters	C–O	19	Sulphonates	S–O	20
Ethers	C–O	14			

[a] R–X: halogenated hydrocarbons; X: a halogen atom.

for alcohols, amines, amino acids, aldehydes, carboxylic acids, esters, ethers, glycols, halogenated hydrocarbons, ketones, phenols, sugars and sulphonates:

$$\mathrm{BOD} = (1.015 \times 10^3)\Delta\delta_{x-y} + 1.523; \quad n = 197,\ r = 0.991,\ s = 3.822 \tag{2}$$

This model is quite accurate when applied to chemicals that can be categorized in the fashion presented by the authors, but it is unclear how to approach modeling chemicals that can be characterized by more than one structural group [11]. This is the greatest obstacle limiting its utility as a predictive model for screening purposes, given the complexity of the chemicals submitted for review by the EPA. The atomic charge difference ($\Delta\delta_{x-y}$) is also a difficult parameter to calculate, limiting the model's functionality. These constraints aside, this method does provide a reasonably reliable mechanism to obtain quantitative aerobic biodegradation rates for a wide variety of chemicals.

3.2. Geating: discriminant analysis/group contribution

This model, having been developed and published by Geating [24], is a precursor to the correlations now available. This model was developed using biodegradation data published between 1974 and 1981. In the development of the model, three types of molecular descriptors were considered as potential variables. These included (1) molecular weight; (2) octanol–water partition coefficient; and (3) Wiswesser line notation (WLN)-based substructural keys. During development of the correlation, obtaining reliable octanol–water partition coefficient measurements in sufficient number proved to be too difficult, and the

Table 4
Geating model probabilities (2,4-dinitrophenol example)

Key	Degradable	Non-degradable
Constant	−6.340	−9.758
Molecular weight	0.023	0.028
Terminal nitro group	0.111	3.306
Hydroxyl group	3.173	0.143
Six-membered aromatic ring	4.259	6.614
Benzene ring	1.767	−2.172

final correlation was based solely on molecular weight and the presence of key substructural fragments. The model is essentially a combined discriminant analysis and group contribution method. An example determination for 2,4-dinitrophenol is as follows:

- Step 1. Sum values for each key from degradable column.
- Step 2. Sum values for each key from non-degradable column.
- Step 3. Insert values from steps 1 and 2 in the probability equation:

$$P = \frac{e^{\text{step 1}}}{e^{\text{step 1}} + e^{\text{step 2}}} \tag{3}$$

Probability values for the function range from 0 to 1. The greater the value for the resultant function, the greater the propensity of the compound to degrade. No numerical value was provided to differentiate between degradable and non-degradable compounds.

This method was used to correctly determine the biodegradability for 270 of 292 degradable compounds and 39 of 57 non-degradable compounds, but the model failed to classify 25 compounds from the test set. A more detailed description of this model is provided in the original publication. Although this model appears to offer some utility due to its simplicity, it is not as accurate as the other methods that were reviewed, and the reliability of the data used is uncertain.

3.3. Gombar, Enslein et al.: discriminant analysis

Gombar and Enslein, improving upon earlier models [25–27], employed a discriminant function analysis (DFA) to model the biodegradability of a chemical training set [15]. They used a two-group DFA, BF (biodegrades fast) and NBF (does not biodegrade fast). They further narrowed the study by creating two submodels separating aliphatic and aromatic compounds. The variables used in the model encompassed electronic, shape, connectivity, and substructure. The electronic descriptors included atomic charges, electron density, residual electronegativity, and polarizability. The shape features consisted of 14 kappa shape indices (m_k) for each molecule. The connectivity variables were represented by molecular connectivity indices of varying order and path type, and the substructural (fragment) descriptors were described by the MOLSTAC© system of Hdi [26] which defines over 3000 molecular fragments. To facilitate model development, the descriptor variables were statistically culled to help prevent chance classification. This resulted in 27 descriptor variables for the aromatic submodel and 22 for the aliphatic submodel. The two group DFA correlations used for both submodels are as follows:

$$d_A = c_A + w_{A1}x_1 + w_{A2}x_2 + \cdots + w_{Ap}x_p \tag{4}$$

$$d = c + wx + wx + \cdots + wx \tag{5}$$

where d_A is the discriminant score; w_A the discriminant weight; x_p the observed measurements on grouping variables; c_i the constant such that objects where $d_A > d_B$ belong in group A (biodegrades), and objects where $d_B > d_A$ belong in group B (persists). The numerical range immediately surrounding $d_A = d_B$ is defined as indeterminate, requiring further investigation.

Biodegradation data for 293 chemicals from the evaluated aerobic, biodegradation database, BIODEG [28], served as the initial training set for the model. The data was selected in a manner similar to the methodology used by Boethling [1,10], except only chemicals with more than three nonconflicting measurements were utilized. Chemicals coded BF in the BIODEG database were also labeled BF in the Gombar and Enslein model, and chemicals classified as BFA (biodegrades fast with acclimation), BS (biodegrades slowly), BSA (biodegrades slowly even with acclimation), and BST (biodegrades sometimes) were included in the NBF group. During model development, the authors subjected the data training set to rigid statistical constraints. Compounds lying "far outside" the distribution were considered to be outliers and were subsequently omitted from the training set. The authors also investigated the structures of the inaccurately predicted compounds for any trends and removed all compounds which exhibited those structural features regardless of whether they were classified correctly or not. It was assumed that the model did not contain adequate descriptors for these compounds. The resulting training set included 142 compounds in the aromatic submodel and 127 in the aliphatic submodel. The final discriminant model correctly predicted the biodegradability of the aromatic and aliphatic compounds in the training set with an overall accuracy of 91%. Table 5 illustrates the descriptor variables and the respective discriminant weights for use in Eqs. (4) and (5) for both the aromatic and aliphatic models.

Although Gombar and Enslein report that the discriminant model displayed a 91% accuracy for their carefully scrutinized training set, the wide-scale screening capability of the model has not been established. The rigorous constraining criteria used by the authors to regulate the distribution of the data in the training set corrupts its apparent utility as a full scale screening model. Even with the stringent criteria for the data, the model was no more accurate than other simpler models that involved much less data manipulation. Also the descriptor variables used in the model were selected more by statistical relevance than a detailed analysis of the aerobic biodegradation processes. Many of the descriptors are difficult to obtain, and their relevance to the process of biodegradation is not adequately interpreted by the authors. An additional limitation of the model is that the predictive capability of the model is based solely on the equation constant in instances where a given chemical cannot be identified by any of the descriptors. Since the constant associated with the classification NBF is greater than for BF, the chemical will be classified as not being readily biodegradable. This attribute does provide a safeguard mechanism, though, mandating additional review.

3.4. Howard, Boethling et al.: QSBR (AERUD)

This research group developed a rudimentary screening model for aerobic ultimate biodegradability (AERUD) in receiving waters using a compilation of topological indexes and macromolecular properties [7]. In developing the model, the group conducted a survey of 22 experts in the field of microbial degradation of xenobiotic chemicals. The experts were solicited to estimate the biodegradability of 50 organic chemicals of widely varying structures. The participants categorized the chemicals for ultimate aerobic biodegradation as either high, intermediate, or negligible and estimated the required time for the process to achieve completion on a scale of 1–4 signifying days, weeks, months, and longer, re-

Table 5
Descriptor variables and associated coefficients (Gombar and Enslein)

Aromatic model			Aliphatic model		
Descriptor variable	Coefficient (NBF)	Coefficient (BF)	Descriptor variable	Coefficient (NBF)	Coefficient (BF)
Equation constant	−4.122	−12.092	Equation constant	−1.539	−10.297
Benzene ring; w/1 or more –OH substituent	0.849	5.042	Saturated alcohol; no *tert* C in molecule	2.072	19.195
Fused or multiple benzene rings w/–OH groups	2.293	19.920	Carboxylic acid; max. eight consecutive $–CH_2$ units	1.949	19.015
O=C–O–CH fragment; max. four consecutive C atoms	3.918	22.426	O=C–O=CH fragment; max. four consecutive C atoms	2.401	20.468
Difference of valence and skeleton chain-type connectivity indices of order 5	3.508	29.472	Ketone (CH–CO–CH)	2.329	15.042
Charge on aliphatic C bound to singly bonded N	0.482	11.092	Saturated amide	2.320	20.085
O atom bound to C of unfused aromatic ring; 1 count	1.340	6.591	SO_3 fragment	1.480	17.571
Alkyl benzene; may have –OH or $–NH_2$ substituents	0.097	8.684	Sec-acyclic amine	2.734	18.510
Valence cluster-type connectivity index of order 3	15.885	0.373	Aldehyde fragment	2.841	16.437
One electron-releasing group on single benzene	4.120	11.069	$N–CH_2CH_2–OH$	−0.640	−17.734
Sum of charges on aliphatic C bound to –OH group	7.833	32.091	Non-methyl-*tert*-amine	1.111	−20.071
Carbonyl group bound to C of benzene ring; 1 count; only C, H and O atoms in molecule	1.160	9.504	*Tert*-amine	1.684	19.735
Ethyl group bound to a heteroatom	−0.442	−15.727	Hydrocarbon; <14 C atoms	0.685	17.931
Pyridine system; aliphatic C bound at position 2	2.088	12.066	Substituted *iso*-butyl group; no –OH in molecule	1.627	−9.645
Aryl amino fragment	2.483	9.130	*n*-Nonyl fragment	1.143	−6.816
Para-substituted phenol; no methoxy substitutions	0.809	7.440	Chain of 12 C atoms	0.319	10.807
Valence cluster-type connectivity index of order 5	–34.998	10.865	Tetra methylene fragment in cyclic molecule	3.820	−7.584
Cl or Br bound to C of benzene; no N in molecule	1.609	−0.482	Saturated sec-amine	0.627	5.553
Amino phenol system	1.474	−7.351	Saturated alkyl halide fragment	3.110	−2.499
Ortho or *meta* substituted aniline	1.621	−3.765	*Tert*-butyl group	1.416	−0.930
Aldehyde group	4.300	−4.330	Mono carboxylic acid	0.919	5.011
One electron-releasing and one -withdrawing group *para* substituted on benzene	0.063	−4.353	Sec-alcohol	0.922	−2.905
A-CCHCHC-A fragment in benzene ring; A: aliphatic non-cyclic atom	−1.251	1.864			
Pyridine w/1 or more $–NH_2$ substitutions	4.382	−2.627			
N doubly bonded to O	4.164	1.726			
Two electron-withdrawing groups *para* to each other on benzene; no -releasing groups	1.305	6.533			

spectively. Since the survey responses were based on estimations, the authors removed 10% of the observations for each chemical from the tail of each distribution to achieve a better-behaved data set.

During the development of the AERUD model, Boethling et al. performed a regression analysis with a variety of potential descriptors, including molecular connectivity indexes (χ), octonol–water partition coefficient (K_{ow}), molecular weight, and the number of covalently bonded chlorine atoms. During this analysis, the authors encountered difficulty calculating the molecular connectivity indexes for four of the surveyed compounds. These compounds were subsequently removed from the data set. The final relationship that was derived through the regression included the second order/valence connectivity index ($^2\chi^v$), fourth order/path-cluster connectivity index ($^4\chi_{pc}$), number of chlorine atoms (n_{Cl}), and molecular weight (M_w). The equation for the sample population of 46 chemicals with a residual of 0.868 is represented as follows:

$$\text{AERUD} = 0.6\ln[^2\chi^v] + \frac{57.25n_{Cl}}{M_w} + \frac{17.56[^4\chi_{pc}]}{M_w} + 1.45 \tag{6}$$

The authors examined the results of this equation and compared them to the survey estimations. They developed a crude classification of the survey chemicals to investigate any relationship between the residuals and chemical structures. The chemicals were classified with respect to the presence of esters, amides, anhydrides, unbranched alkyl groups with greater than four carbons, heterocyclic nitrogen, and whether the chemical contains an oxygen bound to a carbon atom. The residuals for each compound associated with the group classifications were added to the value predicted by Eq. (6) to produce a "corrected" prediction. Boethling et al. then performed a regression between the original and modified AERUD values to achieve the following relationship:

$$\text{AERUD}^{obsd} = 0.946\text{AERUD}^{corr} + 0.137 \tag{7}$$

This correlation resulted in a variance of 88.8%.

The predictive capabilities of the model were tested by comparing the model results to two separate validation sets. The validation sets were obtained through a literature search and from the BIODEG biodegradation database. The chemicals selected from the BIODEG database possessed a summary rating code of BF (biodegrades at a fast rate) and BSA (biodegrades slowly even with acclimation) to facilitate comparison. The values calculated by AERUD were assigned a demarcation of 2.5. All values above 2.5 were assumed to be minimally biodegradable, and all values below 2.5 were presumed to biodegrade readily. The two validation sets encompassing a total of 40 chemicals were compared to the predicted results from the AERUD model. The model correctly classified 36 of the 40 chemicals (90%) as being readily biodegradable or persistent.

Although this model performed satisfactorily with the carefully selected validation set presented by the authors, the utility of this model for predictive use is dubious. The governing correlation was generated using estimated values for biodegradability on a limited number of chemicals that were statistically pruned to exhibit an acceptable distribution. Its accuracy when applied to a large set of greatly diverse compounds was not examined.

3.5. QSBR (group contribution)

Howard et al. presented a predictive model to assess aerobic biodegradability based on group contribution method for a wide variety of chemical structures in 1991 [13]. They chose 34 substructural fragments based upon certain "rules of thumb" regarding the effects of chemical structure on biodegradability and knowledge of common biodegradation pathways. The fragments selected were those known or suspected to have a material impact on biodegradability. From the BIODEG database of over 700 chemicals, the researchers assembled a listing of chemicals that were classified as either BF (biodegrading at a fast rate) or BS, BSA, and BSS (biodegrading slowly). Only chemicals that contained more than one source of nonconflicting biodegradation rates were used. The authors state that no chemicals were arbitrarily omitted from the training set solely because they were determined to be poorly fitted. The final inventory of 229 chemicals was used in a regression analysis to determine the coefficients associated with the 34 substructural fragments in the following equation:

$$Y_j = a_0 + a_1 f_{11} + a_2 f_{12} + \cdots + a_i f_{ij} + e_j \tag{8}$$

where f_{ij} is the number of ith substructures in jth chemical; a_0 the equation intercept; a_i the regression coefficient for ith substructure; e_j the error term; mean value is zero.

The variable Y was defined as a binary indicator. A value of 1 signifies the threshold for rapid biodegradation, and zero for slow biodegradation. The value 0.5 served as the point separating rapid and slow biodegradation. This model accurately classified 96% of the rapidly biodegrading chemicals and 78% of the slowly degrading chemicals with a ratio of 92% (211 out of 229) overall in the training set.

In 1992, Howard et al. [10] amended the previous model to include 35 substructural fragments and modified the previous regression analysis to include a set 264 test chemicals. In this version, three of the original 34 substructural fragments were omitted and were replaced by four other fragments. An analysis identical to the prior model was used to calculate the revised coefficients for Eq. (8). In addition, the authors developed a nonlinear model to determine the probability of aerobic biodegradation to compare to the linear correlation. It is represented as follows:

$$Y_j = \frac{e^{(a_0 + a_1 f_{11} + \cdots + a_i f_{ij})}}{1 + e^{(a_0 + a_1 f_{11} + \cdots + a_i f_{ij})}} \tag{9}$$

The accuracy of both models proved to be comparable for the 264 chemical training set with overall accuracies of 90.5 and 89.8% for the linear and nonlinear correlations, respectively. The nonlinear model proved to be slightly more accurate on the validation set of 27 chemicals with accuracies of 81.5 and 88.8% for the linear and nonlinear models, respectively. Although the overall accuracy for both data sets are similar, the nonlinear model predicted the condition of slow biodegradation more accurately in both instances.

This group contribution model was further improved in 1994 [1], the training set of chemicals was increased to 295, and the list of substructural fragments was modified to include 36 substructures and molecular weight. Table 6 presents the inventory of structural fragment along with their frequency of occurrence in the training set and model coefficients.

Table 6
Group contribution structural fragments and coefficients [10]

Structural fragment of compound	Biodegradation database			Biodegradation survey results		
	Frequency	Linear	Nonlinear	Frequency	Linear coefficient	Nonlinear
Equation constant		0.748	3.01		3.848	3.199
Molecular weight	295	−0.000476	−0.0142	200	−0.00144	−0.00221
Unsub. aromatic (≤rings)	2	0.319	7.191	1	−0.343	−0.586
Phosphate ester	5	0.314	44.09	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
Unsub. 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_2–CH_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
Tiazine 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 hydroC (≥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 four 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

The two regression models (linear and nonlinear) achieved an accuracy 89.5% (264/295) and 93.2% (275/295).

In addition to the biodegradation probability predictions, the authors extended the utility of the aforementioned correlations to include a crude estimate of the duration to achieve primary and ultimate biodegradation. They conducted a survey similar to the one performed for the AERUD model with 17 biodegradation experts evaluating 200 chemicals. The linear correlation was used to determine regression coefficients for the surveyed chemicals for both primary and ultimate biodegradation. Using the regression coefficients presented in Table 6, the model correctly predicted the time required for biodegradation according to the survey with an accuracy of 82.5 and 83.5% for primary and ultimate biodegradation, respectively.

The linear and nonlinear models are attractive because both correlations are relatively simple to implement, and their predictive capability is among the most accurate of the heterologous models. The model parameters are based on scientific observation and conjecture associated with microbial degradation metabolism rather than pure statistical manipulation, and the training set used to determine the substructural fragments is sufficiently large to minimize statistical anomalies and describe a wide variety of chemicals. Substructural fragments that vary in sign for the various correlations are considered to be ambiguous, and the biodegradation predictions for chemicals that are heavily influenced by these substructural fragments should be evaluated with discretion. The authors have suggested that predictions between 0.4 and 0.6 be considered "indeterminate" and should be examined with greater scrutiny.

3.6. Klopman, Balthasar et al.: pattern recognition (computer-aided structure evaluation (CASE))

The models developed for chemical screening have typically been founded on regression or discriminant analysis, but Klopman and coworkers [9,14,16] have fashioned an approach based on pattern recognition and discriminant analysis, CASE. The CASE program is able to recognize molecular structure from a linear inscription of the chemical formula (KLN code) [20]. The program automatically identifies, tabulates, and statistically analyzes substructures that are presumed to be responsible for the biological activity or inactivity of groups of molecules, biophores and biophobes. In a development published in 1993, Klopman used a data base of 283 aliphatic and aromatic chemicals from BIODEG, similar to the one used by Howard et al. [10]. It contained 119 readily biodegradable and 164 persistent chemicals.

The chemical data was entered into the CASE program in KLN code and were labeled either active (biodegrades readily) or inactive (persistent). The program was then able to generate all known possible fragments resulting from cleaving the molecules into subunits from 2–10 heavy atoms with attached hydrogens. These fragments were then labeled as active or inactive depending upon whether the molecule of origin was active or inactive. This set of substructural biophores and biophobes was then statistically condensed to include only those fragments that had a statistical significance. The resultant substructural database consisted of 26 biophores and 11 biophobes. Only 18 of the biophores and six of the biophobes occurred more than once in the database.

To validate the CASE constructed set of substructural descriptors, a validation set of 27 chemicals, identical to that used by Howard et al. [10] was used to test the accuracy of

the CASE model. The CASE program identified and tabulated the targeted substructural fragments occurring in the validation set from the CASE database to make a prediction of biodegradation probability. The CASE program correctly predicted the probability of biodegradability for the validation set at 74%.

Evident from the published results, this model is not as accurate as the other available methods, and its use as a screening device in this form is not likely. Although the accuracy of the model could conceivably be improved by incorporating a larger database, the greatest attribute of this method is its potential for further development. The "artificial intelligence" coding used to select the biophores and biophobes could be further augmented with a greater knowledge base of microbial metabolic pathways and mechanisms to improve its predictive capability. This program also shows promise for the determination of the products of biodegradation. The same type of selection process used to select substructural fragments from parent compounds could be repeated on the substructural fragments to determine the metabolic byproducts of biodegradation and ascertain the definitive goal of the study of biodegradation, acceptable biodegradation. All other models only address either primary or ultimate biodegradation which does not sufficiently characterize the risk associated with a chemical once it is released to the environment. This possibility is addressed in an additional model development by Klopman et al., called META.

3.7. Pattern recognition (META)

Klopman et al. [21] used the CASE programming structure to generate a computer program called META that predicts the metabolic products formed during the aerobic biodegradation of parent compounds. This model is unique in that its primary use is not a simple Boolean analysis of a compound's propensity to biodegrade, but it incorporates a hierarchial logic to predict the most probable metabolites formed from the aerobic transformation of chemicals during biodegradation. The CASE/artificial intelligence system recognizes molecular fragments, biophores, in a compound that are potential sites for microbial attack. Through its dictionary of transformation rules and associated metabolites, the META program then deduces the possible degradation products of the biophore structures. The bio-transformation dictionary used in the program was established by an extensive literary search, and all records in the transformation inventory are known or presumed microbial metabolism mechanisms. The 13 biophores used in the model were obtained using the CASE convention on a database of 385 chemicals, 172 biologically active and 213 persistent chemicals.

Each transform consists of a target fragment and an associated product fragment. The target fragment embodies a group of 2–11 connected heavy atoms along with information regarding its hybridization state and attached hydrogen atoms. Once the target fragments are located in the chemical structure using the CASE methodology, META performs a prioritization algorithm base on the results of experimental data to determine the most likely transformation products. The program can also provide all potential degradation products if requested. The META transformation dictionary also includes degradation pathways for structures other than those in the biophore inventory. Table 7 summarizes some of the more important functionalities and transforms used in the model. The program also accounts for the presence of spontaneous reactions. These are reactions initiated by advantageous

Table 7
Biotransforms for some functionalities[a]

Reaction type	Transforms
Monooxygenation of ketones to form esters	F: CH_n–CO–CH_2; CH_n–$COCH_2$(2–O–3) ($n = 1, 2, 3$); F: CH_n–CO–C=CH–(3–CH=); R: CH_n–CO–C=CH–(3–CH=)(1–O–2) ($n = 1, 2, 3$)
Hydroxylation of methyl/cyclic ketones and further breakdown	F: CH_3–CO–C=; R: CH_3–CO–C=(1–OH); F: OH–CH_2–CO–C=; R: OH–COHCO–C=(3–OH)
Monooxygenation of methoxyl group on aromatic ring	F: CH_3–O–C=; R: CH_2–O–C=(1–OH)
Loss of ammonia from amino acids	F: NH_2–CH–CH_2–(2–CO); R: NH_3CH=CH–(2–CO)
Hydrolysis of alkyl sulfates	F: CH_n–O–SO_2–OH; R: CH_n–$OHSO_2$–OH(3–OH) ($n = 1$ or 2)
Cleavage of imine	F: N=CH–CH_n–; R: NH_2COH–CH_n–; F: NH=CH–CH_n–; R: NH_3COH–CH_n– ($n = 0, 1, 2$, or 3)
Azo reduction	F: CH_n–N=N–; R: CH_n–NH_2NH_2– ($n = 0, 1, 2$, or 3)

[a] "F" represents "find"; "R" represents "replace with".

free energy changes which occur without catalysis following the formation of unstable transitional compounds by a prior transformation. As of 1995, the META program included over 100 distinct transformations. To account for the observed toxicity to microorganisms by certain chemicals, META includes a database of toxicophores and issues a warning statement to the user whenever toxicophore structure is detected by the program. These deal primarily with mono-halo-substituted chemicals.

This model was executed using the same validation set of 27 chemicals used in the Boethling group contribution method [10] and the CASE model development. The authors postulated that the 13 chemicals that were readily biodegradable should be metabolized using a least one of the transformation mechanisms in the META program. The results of the trial indicated that the META program did identify at least one degradation transformation for each readily biodegradable compound in the training set. The META program also flagged 12 of the 14 persistent chemicals as containing toxicophores, thus inhibiting degradation. Of the remaining two compounds, META found no transformation processes for one, and the other (3-methylcholanthrene) was predicted to undergo oxygenolytic elimination. The authors speculate that the observed inactivity of this compound is due to its insolubility in water.

This program appears to be promising as a predictive tool to determine the possible metabolic byproducts of biodegradation. The authors state that the model may have a propensity to "overpredict" metabolites for the degradation of a given compound but feel that this allows the user to establish all possible products so that the occurrence of any toxic metabolites is not neglected. It is stated that the META program may occasionally predict degradation byproducts for stable, persistent chemicals when no experimental

transformations have been observed, but this only provides the possible metabolites that may occur once microorganisms have adapted to the chemical and acquire the ability to degrade it. This program also has the advantage that additional transformation pathways and toxicophores can be added to the database to allow for the characterization of increased numbers of chemical compounds, including anaerobic degradation. The program could also be modified to include substructural fragments that are noted to be toxic to humans as well as microorganisms as part of a risk screening analysis.

3.8. Niemi et al.: discriminant analysis

In 1987, Niemi et al. [11] presented a combined multivariate statistical and a heuristic model in order to predict the biodegradability of compounds on a screening level. Niemi et al. prepared a database of BOD measurements of approximately 1200 tests for about 400 chemicals, essentially the same database used by Dearden and Nicholson [15]. The list was reduce to 287 chemicals by selecting only tests of 5-day duration unless degradation was completed sooner or tests of longer duration did not show degradation. An attempt was made to include only acclimated tests, but often, this information was not provided. If conflicting values for a chemical were reported, the authors used the highest value in their analyses. To normalize the varying BOD measurements, the researchers divided the measurements by the chemicals respective theoretical oxygen demand if the chemical were to totally degrade. The resultant percent ThOD values were then standardized by means of estimated half-lives if the measured BOD were of some duration other than 5 days. To facilitate their analysis, the authors declared the distinction between persistent and degradable chemicals as those possessing a half-life of 15 days (16% ThOD).

Prior to developing the multivariate DFA, Niemi et al. calculated a total of 54 molecular connectivity indices based on the order and term of the index. Additional descriptors included molecular weight, K_{ow}, molar volume, molar refraction, and parachor. In an attempt to reduce the dimensionality of the molecular connectivity indices, the authors conducted a principal components analysis (PCA) [29]. They performed the PCA for 45 molecular connectivity indices for 16,121 of the chemicals listed in the TSCA inventory. The authors chose the TSCA list because they felt that it provided a better representation of the "universe" of manufactured chemicals than their limited BOD training set. The authors discovered that eight principal components described more than 94% of the variation for the 16,121 chemicals. In general, principal component 1 (PC 1) was related to the relative size of the molecule, PC 2 was related to the degree of molecular branching and PC 3 was generally associated with cyclic compounds. PC 4 through PC 8 were more elusive, based on subtle variations within the molecule.

Niemi et al. grouped the chemicals in the training set using *K*-means clustering with the eight principal components and the value of K_{ow} as the clustering variables. They then used the connectivity indices and physicochemical variables as discriminators to distinguish between degradable and persistent chemicals within a cluster. They limited their analysis to a maximum of nine variables and eight clusters. During the testing of the model, it was observed that the results of the *K*-means clustering was significantly influenced by the presence of outliers. To remedy this situation, the authors divided the training set data base into two groups, the "outer space" (54 chemicals) and the "inner space" (233 chemicals). The

"outer space" was defined as chemicals with at least one of the eight principal components that was more than two standard deviations from the mean. This redefined model was then analyzed using different combinations of clusters and variables. Distinct combinations of variables were important as discriminators within each of the clusters. Overall, 94% of the persistent chemicals and 85% of the degradable chemicals were properly categorized using the best combinations in the "outer and inner space".

Niemi et al. then used the results of the multivariate model in the development of the heuristic model. They observed the results of the multivariate model to identify structural features that were consistently found among chemicals within a specific cluster. For the heuristic model, they designated five chemical groups: (1) unbranched, noncyclic chemicals; (2) branched, noncyclic chemicals; (3) aliphatic cyclic chemicals; (4) aromatic cyclic chemicals and (5) mixed aliphatic and aromatic cyclic chemicals. The authors found that many of the degradable and persistent chemicals were consistent across several of the groups. The groups were eventually modified so that 12 structural groups represented degradable chemicals, and 16 structural groups described persistent chemicals. Table 8 presents a summary of the groups along with the range of half-lives ascribed to each structural feature. This heuristic model was used by the authors to correctly predict the biodegradability of 91% of the degradable chemicals and 96% of the persistent chemicals.

Although the multivariate and the heuristic models are based on a substantive set of data, they are not conducive to predictive screening. The multivariate model is limited in that it has been excessively tailored to fit a single training set. The delimiting variables that specify which cluster in which to place a chemical were chosen purely through statistical manipulation, and it is unclear how to characterize a new compound that may not "fit" the data set without first performing another discriminant analysis incorporating the chemical in question. There appears to be a minimum of scientific theory or conjecture in the selection of the discriminating variables, and the authors offer no explanation as to why certain variables may be more significant than others in the prediction of biodegradability.

The heuristic model, although premised on observed and conjectured mechanisms of biodegradation, is overly simplified. The authors present a range of half-lives for each designated structural descriptor, but it is unclear how to characterize a compound containing several of the structural features, especially if the range of half-lives differ greatly. They provide no mechanism in the literature for apportioning the effects of each structural group for a chemical containing more than one feature. This model appears to be a basis for the development of a more complicated group contribution correlation rather than as a separate viable model.

3.9. Tabak, Govind et al.: QSBR (group contribution)

In 1990, Desai et al. [17] introduced a group contribution model to quantitatively predict aerobic, first order biodegradation constants for widely varying organic compounds. This research group characterized the first order rate constant in a manner similar to Boethling et al. and Gombar and Enslein [25,26] by the following relationship:

$$\ln(k) = \sum_{j=1}^{L} N_j \alpha_j \tag{10}$$

Table 8
Structural features associated with heuristic model[a]

Degradable			Persistent		
No.	Descriptor	Half-life (day)	No.	Descriptor	Half-life (day)
1	One halogen subs. on an unbranched chemical	<12	1	*Tert*-butyl terminal branch	>15
2	One cyano subs. on an unbranched chemical	<10	2	Epoxides	>20
3	Aldehydes	2–11	3	Aliphatic chemicals with fused rings and no branches	>35
4	Hydrocarbons	3–17	4	Two terminal isopropyl subgroups on noncyclic chemical	>35
5	Alcohols, esters, amines	2–16	5	Aliphatic cyclic chemicals without branches	>40
6	Acids	3–12	6	Halogen subs. on a branched, noncyclic or cyclic chemical	>5
7	Amino acids	2–5	7	Isopropyl or dimethyl amine subs. without other "degradable" subs.	>25
8	Sulfonates	2–17	8	Two halogen subs. on an unbranched, cyclic chemical	>15
9	Subs. benzene ring ($K_{ow} < 2.18$)	2–16	9	More than two hydroxy subs. on an aromatic ring	>15
10	Biphenyl and two or less hydroxy-subs. polyaromatics	<15	10	Two or more rings	>20
11	Cyclic chemicals consisting only of C, O, N, and H	2–15	11	Two terminal diamino groups on a noncyclic chemical	>35
12	Two aromatic rings (e.g. naphthalene and amino-napth.)	<15	12	More than one amino branch on ring with Nas ring member	>100
			13	Two terminal double-bonded C on an unbranched chemical	>100
			14	Benzene ring with >2 subs. (non-hydroxy) and $K_{ow} > 2.18$	>100
			15	Cyano group on a chain of >8 atoms	>100
			16	Highly branched chemicals	>100

[a] All degradable descriptors assume that other subgroups associated with persistence are not present.

Table 9
Groups and contribution values (first order rate constant)

Structural group	α_j
Methyl (CH_3)	−1.367
Methylene (CH_2)	−0.0438
Hydroxy (OH)	−1.709
Acid (COOH)	−1.313
Ketone (CO)	−0.507
Amine (NH_2)	−1.465
Aromatic CH (ACH)	−0.502
Aromatic carbon (AC)	1.066

where N_j is the number of groups of type j in compound; α_j the contribution of group of type j; L the total number of groups in compound.

In order to determine the contribution weights, α_j, the series of linear relationships described by Eq. (10) were solved using the method of least squares. The authors state that this linear correlation is adequate for the purpose of determining first order approximations but will degenerate if interactions between structural groups becomes significant. They state that these effects could possibly be considered by incorporating second order or higher terms into the equation.

Since the nature of the result is strictly quantitative, the amount of available experimentally relevant measurements is much less than that for the Boolean-type screening models. The authors used biodegradation rates obtained through a literature search. The total number of chemicals employed in the estimation of group contribution values was 18, and the number of structural groups was eight. The authors made sure that each structural group occurred in at least five compounds to minimize chance correlations, according to standard univariate statistical analyses. Table 9 shows the structural groups used in the analyses along with their respective weight values.

To validate the model, the authors experimentally determined the first order rate constant for 11 compounds and then compared the measurements to the model predictions. The results are presented in Table 10.

Table 10
Comparison of actual and predicted rate constants

Compound	Experimental ln (k)	Predicted ln (k)	Error (%)
o-Cresol	−2.688	−2.950	9.75
m-Cresol	−2.369	−2.950	24.5
p-Cresol	−2.465	−2.950	19.7
Phenol	−3.001	−3.151	5.00
2,4-Dimethylphenol	−2.846	−2.744	−3.39
2-Butanone	−3.133	−2.940	4.84
Acetone	−3.116	−3.241	4.00
Butylbenzene	−3.129	−2.940	−6.02
1-Phenylhexane	−3.397	−3.028	−10.87
Aniline	−3.124	−2.907	−6.92
Benzoic acid	−2.163	−2.755	27.39

Table 11
Structural groups and weight constants (Monod rate constants)

Structural groups	α_j Values		
	K_s	Y	μ_{max}
Aromatic carbon (AC)	−0.033	0.19	−0.01
Aromatic CH (ACH)	0.048	0.95	0.06
Methyl (CH_3)	0.045	0.92	0.06
Methylene (CH_2)	−0.028	0.51	0.01
Methelene (CH)	−0.107	2.41	−0.03
Hydroxy (OH)	0.173	0.80	0.07
Ester (COO)	0.057	−0.11	0.00
Ketone (CO)	0.182	2.87	0.33
Chlorine (Cl)	−0.023	−0.29	0.09
Nitro (NO_2)	−0.025	−0.13	0.09

In 1992, Tabak et al. [30] expounded on this model to include the determination of Monod rate constants, μ_{max} and K_s. They used approximately 28 chemicals to calculate equation weights for 10 structural groups. The structural groups and associated weight constants are presented in Table 11. The Monod constant model was validated with a test set of 14 chemicals. The authors state that the experimental values agree within 25% to the predicted values.

These models, though simple in construction, do not appear to be accurate enough to produce reliable values for rate constants for a large array of compounds. The authors state confidence in the model to generate results within an order of magnitude. This is probably satisfactory for most instances where degradation constants are required, but the results of the model seem to be more useful for semi-quantitative comparison than for actual numerical rate constants. The inventory of structural groups contained in the model is very limited, and the meager training set of chemicals used to calculate the weight parameters detracts confidence in the model's ability to predict rate constants for the widely varying structures of industrially significant compounds. Reliability of this model could be improved by increasing the number of chemicals in the training set and the quantity of fragments in the substructural inventory, similar to the process used by the Boethling research group.

3.10. Group contribution/neural network

In 1993, Tabak and Govind [31] devised a predictive model to calculate first order kinetic biodegradation constants using a multi-layered neural network model. This model was developed using the same structural fragments and training set as that used for the previous first order rate constant model. This method was designed in an attempt to include the effects of interactions between the various groups. In this model, each node has several inputs and calculates a single output. The input values exhibit known activation and weight values. The output from each node is then determined nonlinearly by Eq. (11):

$$O_{pj} = \frac{1}{1 + \exp(-\Sigma W_{ji} O_{pi} + \Theta_j)} \tag{11}$$

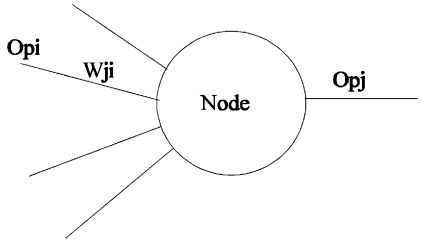

Fig. 1. Single node processing element in a neural network.

where O_{pj} is the output value of node j; O_{pi} the output value of node i; W_{ji} the connection weight between the ith and jth nodes; Θ_j is the bias of the jth node.

Fig. 1 illustrates a single node processing element.

The neural network constructed for this model consisted of three layers with eight input nodes and eight intermediate layer nodes. The single output node provides the predicted value of the first order rate constant. Each input node corresponds with one of the eight chemical groups used in the previous linear model to determine first order rate constants (Table 9). The inputs consist of the number of each structural group in a given chemical and its associated weight value. The weight values were established by using a gradient search technique to minimize the mean square difference. A more detailed description of the neural network methodology is presented by Bhagat [19]. The calculated output of each node is normally within the range of 0.0–0.1 unless the node corresponds to the group of the current input in which case the output typically ranges between 0.9 and 1.0.

To test the accuracy of the neural network model, the results from the linear model and the neural network model were compared to the experimental values for the 18 chemicals contained in the training set and the eight chemicals in the validation set. The results are shown in Table 12. From these results, it is evident that the neural network model which incorporates interstructural activities is more accurate than the linear model in most instances, especially with respect to the independent validation set.

This capability of this model for the prediction of biodegradation constants for large numbers of chemicals with widely varying structures has not been confirmed. It was constructed using only 18 chemicals in training set and eight in the validation set. Although it performed well in predicting the biodegradation constants for these chemicals, it has not been tested on enough compounds of with varying biodegradation rates. These compounds all have degradation constants within an order of magnitude, and it does not appear that any decisively persistent chemicals were include in the model development. The neural network methodology does initially appear promising. It would be interesting to modify the model to incorporate the structural groups presented by Boethling and increase the number of chemicals in the training set substantially. Since the Boethling substructural groups are founded on a considerable review of biodegradation literature and have proven their accuracy in the Boethling group contribution model, it seems that the accuracy of this model would be greatly improved if neural networks were used in the analysis. Since the neural network model incorporates inter-structural relationships, the effects of various substructures on the metabolism of other substructures could be considered.

Table 12
Neural network and linear model comparison

Compound	Experimental ln (*k*)	Neural network		Linear method	
		−ln (*k*)	Error (%)	−ln (*k*)	Error (%)
Training set					
Ethyl alcohol	3.02	3.01	0.33	2.97	1.43
Butyl alcohol	3.19	3.16	0.94	3.24	1.30
Ethylene glycol	3.49	3.45	1.15	3.39	2.87
Acetic acid	2.66	2.68	0.75	2.49	6.66
Propanoic acid	2.81	2.81	0.06	2.65	5.84
n-Butyric acid	2.87	2.83	1.39	2.75	4.17
n-Valeric acid	2.65	2.70	1.89	2.88	8.86
Adipic acid	2.96	2.93	1.01	2.94	0.55
Methyl ethyl ketone	3.58	3.63	1.4	3.31	11.90
Hexamethylenemine	4.43	4.22	4.74	3.96	10.43
n-Hexylamine	2.96	2.97	0.33	3.52	19.11
Monoethanolamine	3.35	3.38	0.90	3.41	1.80
Acetamide	3.03	3.01	0.66	3.48	15.19
Benzene	2.92	2.94	0.68	2.87	1.62
Benzyl alcohol	2.96	2.94	0.68	3.12	5.57
Toluene	2.73	2.70	1.10	2.70	1.10
Acetophenone	3.34	3.31	0.90	3.33	0.38
Aminophenol	3.27	3.29	0.61	3.13	4.26
Validation set					
o-Cresol	2.69	2.62	2.02	2.87	6.59
Phenol	3.00	2.91	3.17	2.99	0.29
2,4-Dimethylphenol	2.85	2.58	9.29	2.74	3.82
Butylbenzene	3.13	3.18	1.53	3.10	5.84
Acetone	3.12	3.14	0.60	3.15	0.96
1-Phenylhexane	3.40	3.67	7.90	4.76	40.0
Aniline	3.12	3.00	3.85	4.03	29.0
Benzoic acid	2.16	2.31	6.95	3.64	68.5

4. Summary and conclusions

It is evident that each of the QSBR models that were surveyed has its own advantages and disadvantages. In order to expedite the comparison process, Table 13 was developed to summarize the utility of each model for predictive use in determining the biodegradability of various chemical compounds. The table ranks each model according to its complexity or reproducibility, accuracy, effective range of chemical structures, reliability of data set, and the size of the data set used to develop the model. Rankings are given on a scale from 1 to 10 with 10 being the highest score achievable. The following rankings have been provided solely for initial comparison purposes only, and their subjective nature warrants judicious interpretation.

The table helps illustrate that although each method has its benefits and limitations, the group contribution technique presented by Howard et al. [10] appears to be the most advantageous for use in predictive screening. The model is simple in structure and has

Table 13
Model comparisons (heterologous biodegradability correlations)[a]

	Author	Method	Complexity	Accuracy	Range	Reliability of data	Size of data set
1	Deardon et al.	Atomic charge	❹	❽	❻	❻	❼
2	Geating	DA/GC	❾	❼	❼	❺–❻	❽
3	Gombar et al.	DA	❸	❽	❽	❼	❽
4	Howard et al.	AERUD	❼	❼	❺	❸	❷
5	Howard et al.	GC	❿:	❽	❼	❼	❽
6	Klopman et al.	PR (CASE)	❶:	❺	❼	❼	❽
7	Klopman et al.	PR (META)	❶:	❼	❼	❼	❿:
8	Niemi et al.	DA	❷	❾	❻	❻	❽
9	Tabak et al.	GC	❿:	❻	❹	❼	❶:
10	Tabak et al.	Neural net/GC	❼	❼	❹	❼	❶:

[a] DA: discriminant analysis; GC: group contribution; PR: pattern recognition; ❶:: lowest scorer; ❿:: highest score.

proven to be reliably accurate for a wide range of chemical structures, established by the large data set, and the quality of the data set used to develop this correlation was obtained from the BIODEG evaluated biodegradation database which is recognized as a source of the some of the most reliable biodegradation data available. Therefore, it is recommended that the Howard et al. [10] group contribution method be used for the initial screening of chemical compounds to aid in determining whether additional biodegradation information is necessary to reliably establish relative biodegradability.

Acknowledgements

This review of biodegradation estimation methods was sponsored by the US EPA Environmental Research Laboratory in Athens, GA, under Assistance Agreement Number CR823226-01-0, "Prediction of Chemical Parameters by Computer." The authors express their appreciation to the EPA Project Officer, Dr. Samuel Karickhoff, for supporting and encouraging our efforts. Despite generous financial support for this work by the US EPA, the calculations and comparisons herein have not been critically evaluated by the Agency, and no official endorsement is implied.

References

[1] R.S. Boethling, P.H. Howard, W. Meylan, W. Stiterler, H. Beauman, N. Tirado, Group contribution method for predicting probability and rate of aerobic biodegradation, Environ. Sci. Technol. 28 (1994) 459–465.
[2] W.J.G.M. Peijnenburg, Structure-activity relationships for biodegradation: a critical review, Pure Appl. Chem. 66 (1994) 1931–1941.
[3] P. Kuenemann, P. Vasseur, J. Devillers, Structure biodegradability relationships, in: W. Karcher, J. Devillers (Eds.), Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology, Kluwer Academic Publishers, Dordrecht, 1990.

[4] K.M. Scow, Rate of biodegradation, in: W.J. Lyman, W.F. Rosenblatt, (Eds.) Handbook of Chemical Property Estimation Methods, McGraw-Hill, New York, 1983.
[5] A.M. Chakrabarty, Biodegradation and Detoxification of Environmental Pollutants, CRC Press, Boca Raton, FL, 1982.
[6] J.R. Parsons, H.A.J. Govers, Quantitative structure relationships for biodegradation, Ecotoxicol. Environ. Safety 19 (1990) 212–227.
[7] R.S. Boethling, A. Sablijic, Screening-level model for aerobic biodegradability based on a survey of expert knowledge, Environ. Sci. Technol. 23 (1989) 672–679.
[8] S.A. Moore, J.D. Pope, J.T. Barnett, L.A. Suarez, Structure-Activity Relationships and Estimation Techniques, US Environmental Protection Agency, Athens, GA, 1989.
[9] G. Klopman, Artificial intelligence approach to structure-activity studies: computer automated structure evaluation of biological activity of organic molecules, J. Am. Chem. Soc. 106 (1984) 7315–7321.
[10] P.H. Howard, R.S. Boethling, W.M. Stiteler, W.M. Meylan, A.E. Hueber, H.A. Beauman, M.E. Larosche, Predictive model for aerobic biodegradability developed from a file of evaluated biodegradation data, Environ. Toxicol. Chem. 11 (1992) 593–603.
[11] G.J. Niemi, G.D. Veith, R.R. Regal, D.D. Vaishnav, Structural features associated with degradable and persistent chemicals, Environ. Toxicol. Chem. 6 (1987) 515–527.
[12] R.S. Boethling, B. Gregg, F.R. Gabel, N.W. Campbell, A. Sablijic, Expert systems survey on biodegradation of xenobiotic chemicals, Ecotoxicol. Environ. Safety 18 (1989) 252–267.
[13] P.H. Howard, R.S. Boethling, W. Stiteler, W. Meylan, J. Beauman, Development of a predictive model for biodegradability based on BIODEG, the evaluated biodegradation database, in: J.L.M. Hermens, A. Opperhuizen (Eds.), QSAR in Environmental Toxicology, Vol. IV, Elsevier, New York, 1991.
[14] G. Klopman, MULTICASE: a hierarchical computer automated structure evaluation program, Quantitative Struct. Activity Relationships 11 (1992) 176–184.
[15] J.C. Dearden, R.M. Nicholson, Correlation of biodegradability with atomic charge difference and superdelocalizability, in: K.L.E. Kaiser (Ed.), QSAR in Environmental Toxicology, Reidel, Dordrecht, 1987.
[16] G. Klopman, D.M. Balthasar, H.S. Rosendranz, Application of the computer-automated structure evaluation (CASE) program to the study of the structure-biodegradation relationships of miscellaneous chemicals, Environ. Toxicol. Chem. 12 (1993) 231–240.
[17] S.M. Desai, R. Govind, H.H. Tabak, Development of quantitative structure-activity relationships for predicting biodegradation kinetics, Environ. Toxicol. Chem. 9 (1990) 473–477.
[18] J.C. Dearden, R.M. Nicholson, QSAR study of the biodegradability of environmental pollutants, in: D. Hadzi, B.J. Blanzic (Eds.), QSAR in Drug Design and Toxicology, Elsevier, Amsterdam, 1987.
[19] P. Bhagat, An introduction to neural nets, Chem. Eng. Prog. (1990) 55–60.
[20] G. Klopman, M.J. McGonigal, Computer simulation of physical-chemical properties of organic molecules. 1. Molecular system identification, J. Chem. Information Comput. Sci. 21 (1981) 48–52.
[21] G. Klopman, M. Dimayuga, J. Talafous, META: 1. A program for the prediction of metabolic transformation of chemicals, J. Chem. Information Comput. Sci. 34 (1994) 1320–1325.
[22] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.
[23] J.C. Dearden, R.M. Nicholson, The prediction of biodegradability by the use of quantitative structure-activity relationships: correlation of biological oxygen demand with atomic charge difference, Pesticide Sci. 17 (1986) 305–310.
[24] J. Geating, Project Summary, Literature Study of the Biodegradability of Chemicals in Water, Vols. 1 and 2, US Environmental Protection Agency, EPA-600/S2-172/176, 1981.
[25] V.K. Gombar, K. Enslein, Quantitative Struct. Activity Relationships 9 (1990) 321.
[26] V.K. Gombar, K. Enslein, A structure-biodegradability relationship model by discriminant analysis, in: J. Devillers, W. Karcher (Eds.), Applied Multivariate Analysis in SAR and Environmental Studies, Kluwer Academic Publishers, Dordrecht, 1991.
[27] K. Enslein, M.E. Tomb, T.R. Lander, Structure-activity models of biological oxygen demand, in: K.L.E. Kaiser (Ed.), QSAR in Environmental Toxicology, Reidel, Dordrecht, 1984.
[28] P.H. Howard, BIODEGã, Syracuse Research Corporation's Environmental Fate Database, Copyright 1987.

[29] G.J. Niemi, R.R. Regal, D.D. Vaishnav, G.D. Vieth, A preliminary model to predict biodegradability from chemical structure, in: A.W. Bourquin, P.H. Pritchard, W.W. Walker, R. Parrish (Eds.), Proceedings of the Biodegradation Kinetics, Navarrre Beach, FL, EPA/600/9-85/018, US Environmental Protection Agency Office of Research and Development, Gulf Breeze, FL, 1983.
[30] H.H. Tabak, C. Gao, S. Desai, R. Govind, Development of predictive structure-biodegradation relationship models with the use of respirometrically generated biokinetic data, Water Sci. Technol. 26 (1992) 763–772.
[31] H.H. Tabak, R. Govind, Prediction of biodegradation kinetics using a nonlinear group contribution method, Environ. Technol. Chem. 12 (1993) 251–260.