Biodegradation **11:** 37–47, 2000.
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

Investigating the specificity of regulators of degradation of hydrocarbons and hydrocarbon-based compounds using structure-activity relationships

Jacob G. Bundy[1,2], David G. Durham[3], Graeme I. Paton[1] & Colin D. Campbell[2]
[1]*Department of Plant and Soil Science, University of Aberdeen, Aberdeen AB24 3UU, UK;* [2]*Soil Science Group, Macaulay Land Use Research Institute, Craigiebuckler, Aberdeen AB15 8QH, UK;* [3]*School of Pharmacy, Robert Gordon University, Schoolhill, Aberdeen AB10 1FE, UK*

Accepted 25 April 2000

Key words: bioluminescence, biosensor, hydrocarbon degradation, QSAR, specificity

Abstract

Microbial biosensors which have genes for bioluminescence coupled to genes that control hydrocarbon degradation pathways can be used as reporters on the specificity of regulation of those pathways. Structure-activity relationships can be used to discover what governs that specificity, and can also be used to separate compounds into different groups depending on mode of action. Published data for four different bioluminescent biosensors, reporting on toluene (two separate biosensors), isopropylbenzene, and octane, were analyzed to develop structure-activity relationships between biological response and physical/chemical properties. Good QSARs (quantitative structure-activity relationships) were developed for three out of the four biosensors, with between 88 and 100 per cent of the variance explained. Parameters found to be important in controlling regulator specificity were hydrophobicity, lowest unoccupied molecular orbital energies, and molar volume. For one of the biosensors, it was possible to show that the biological response to chemicals tested fell into three separate classes (non-hydrocarbons, aliphatic hydrocarbons, and aromatic hydrocarbons). A statistically significant QSAR based on hydrophobicity was developed for the fourth biosensor, but was poor in comparison to the other three (44 per cent variance explained).

Introduction

Bioluminescence is widely accepted to be an excellent reporter mechanism for microbial biosensors (Atlas et al. 1992; Meighen 1988). By coupling the genes for bioluminescence to genes for specific promoters, biosensors can be produced that are activated by a particular biological response or activity (Barkay et al. 1995). Biosensors of this type have been constructed to report on heavy metal resistance, N and P starvation, nitrate and nitrite, and on a range of stresses, such as heat shock (Tescione and Belfort 1993; Selifonova et al. 1993; Kragelund et al. 1995; Prest et al. 1997; Belkin et al. 1997). The largest number, however, report on hydrocarbon degradation pathways: biosensors exist for naphthalene and salicylate, toluene and other monocyclic aromatic hydrocarbons, isopropylbenzene, octane, and polychlorinated biphenyls, using both prokaryotic and eukaryotic luciferases (King et al. 1990; Applegate et al. 1998; Ikariyama et al. 1997; Willardson et al. 1998; Selifonova and Eaton 1996; Layton et al. 1998; Sticher et al. 1997). These catabolic sensors use promoters from hydrocarbon degradation pathways for *lux* or *luc* genes, producing microbes that respond in a sensitive, rapid, and quantitative way to hydrocarbon micropollutants.

Bacterial enzymes for hydrocarbon degradation typically have a low specificity, allowing microbes to deal with a large number of different, related compounds that they might potentially encounter (Zylstra and Gibson 1997). In practice, this means that a microbial catabolic sensor will respond to a range of structurally related compounds. There can even be a response to chemicals surprisingly different to the archetypal substrate compound – e.g., Selifonova and Eaton (1996) found that an isopropylbenzene biosensor was induced by a range of compounds including halogenated chemicals, both aliphatic and aromatic, and even heterocycles. It follows that *lux*-marked biosensors can be used as tools to study the regulation of hydrocarbon degradation pathways. Data can be produced rapidly and easily – hence they are ideal for developing quantitative structure-activity relationships (QSARs). Previous studies on the specificity of regulation have not attempted to establish QSARs to explain this specificity (Abril et al. 1989; Marqués and Ramos 1993). QSARs are often considered solely as models that are required to predict unknown biological data, for which a robust and well-validated model is necessary, together with a full understanding of which chemicals are appropriate for the model. However, QSARs can also be used not just as predictive tools, but to gain more understanding about a biological system and what affects its response to different chemicals (Hermens 1995). A meta-analysis of published data from four different studies using catabolic hydrocarbon biosensors (toluene, isopropylbenzene, and octane sensors) was carried out, with the aim of deriving QSARs that would further understanding of the specificity of regulation of hydrocarbon degradation pathways.

Materials and methods

Luminescence data were taken from four different published studies. The archetypal response compounds were, respectively, toluene, toluene, isopropylbenzene, and octane.

Biosensor A: *Escherichia coli* DH5α (pGLTUR) (Willardson et al. 1998)
Based on the TOL plasmid and *luc* (eukaryotic luciferase) reporter gene. The *luc* genes were placed under the control of P_u, the promoter of the upper pathway genes, which is regulated by the hydrocarbon-binding XylR protein. Biosensor A thus reports on the specificity of the XylR protein.

Biosensor B: *E. coli* HB101 (pTSN316) (Ikariyama et al. 1997)
This is similarly based on the TOL plasmid and *luc*, and has the *luc* genes ultimately regulated by the XylR protein.

Biosensor C: *E. coli* HMS174 (pOS25) (Selifonova and Eaton 1996)
Based on the *ipb* operon, which degrades isopropylbenzene, and *lux* (prokaryotic) genes. The *luxCDABE* genes, which encode for luciferase and for enzymes generating the aldehyde substrate of bacterial luciferase, were placed under the control of the regulatory elements *ipbR* and *ipbO/P* of the *ipb* operon.

Biosensor D: *E. coli* DH5α (pGEc74, pJAMA7)
Based on the OCT plasmid, which encodes genes for the degradation of octane, and *lux* genes. The *luxAB* genes (coding only for the luciferase) were placed under the control of the *alkB* promoter, which is regulated by the AlkS protein product of the *alkST* region (van Beilen et al. 1994).

Full details of the luminescence assay procedures are given in the original studies and will not be repeated here. Briefly, luminescence assays were performed by the original authors in all cases by growing the cells to a predetermined optical density, exposing the cells to known concentrations of the compounds for a certain length of time, and then reading the luminescence measurements on a luminometer or liquid scintillation counter. Biosensors B and D were stored as frozen stocks before use. Biosensors A and B are based on eukaryotic *luc* genes, and therefore also required lysis of the cells and addition of luciferin as a substrate for the luciferase enzymes before luminescence was measured.

It is usually necessary to transform the biological response to develop a QSAR. The data for biosensors A, B and C were log-transformed; this was found not to be necessary for biosensor D. The data for biosensor C were also normalized by dividing by concentration (μM) because they were reported at different concentrations. The biological response measurement is hence referred to as log(*ind*) or *ind*: *ind* refers to induction, i.e., the ratio of light output in arbitrary luminescence units of cells exposed to compounds causing luminescence (induced cells) to the light output of cells not so exposed (uninduced cells). (This measurement is not normalized with respect to cell concentration.) Biosensor A is an exception: data were reported not as raw induction values, but as saturation constants $K_{1/2}$: $K_{1/2}$ is defined as the concentration of chemical (μM) which elicited a level of induction that was 50% of the maximum level that could be obtained.

The energy levels of the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO) were calculated for geometry-optimized molecules using the CNDO option of HyperChem (Hypercube, Florida, USA), and are given in electronvolts. Log P is $\log K_{\mathrm{ow}}$, the logarithm of the octanol/water partition coefficient, a widely-used parameter that is a measure of hydrophobicity. All $\log P$ values were taken from Hansch et al. (1995) and Eastcott et al. (1988). All values for *vol*, molar volume (used in Equation (9)) are taken from Eastcott et al. (1988).

Stepwise multiple linear regression was used to select the combination of statistically significant ($p < 0.05$) parameters that maximized the percentage of variance explained (adjusted R^2 value) by the regression. The regression models obtained were validated by calculating Q^2 values. Q^2 is obtained from "leave one out" testing, also known as cross-validation. A data point is removed from the set, and the regression recalculated; the predicted value for that point is then compared to its actual value. This is repeated until each datum has been omitted once; the sum of squares of these deletion residuals can then be used to calculate Q^2, an equivalent statistic to R^2. Q^2 values can be considered a measure of the predictive power of a regression equation: whereas R^2 can always be increased artificially by adding more parameters, Q^2 decreases if a model is overparameterized (Eriksson et al. 1997), and is therefore a more meaningful summary statistic for QSAR. The regressions were also tested by visual examination of plots of fitted values against actual values, and plots of fitted values against residuals. All statistical analysis was carried out using Genstat for Windows 3.2 (NAG Ltd., Oxford, UK).

Results

All physical/chemical data used in the regression equations are given in Table 1. All of the biological data, in the form in which they were used for the regression equations, are given in Table 2.

Biosensor A: Willardson et al. (1998)
Data were reported as saturation constants $K_{1/2}$. A good relationship was obtained against $\log P$ alone:

$$\begin{aligned}\log K_{1/2} &= -0.95 \log P + 4.93\\ n &= 13, Q^2 : 0.86, \%V : 89 \qquad (1)\end{aligned}$$

n is the number of data used in regression, $\%V$, the percentage of variance explained; adjusted R^2 value.

However, multiple linear regression showed that addition of a LUMO term improved the model slightly (Figure 1):

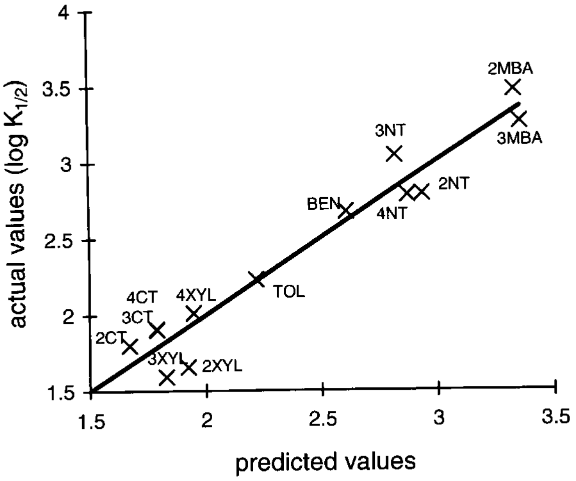

Figure 1. Model for biosensor A (Equation (2)): plot of actual values against predicted values. Line shows $x = y$, i.e., ideally fitted model. Full compound names listed in Tables 1 and 2.

$$\begin{aligned}\log K_{1/2} &= -0.912 \log P - 0.137 LUMO + 525\\ n &= 13, Q^2 : 0.89, \%V : 92. \qquad (2)\end{aligned}$$

The extra term is statistically significant ($p = 0.031$).

Biosensor B: Ikariyama et al. (1997)
Data were reported as induction levels, with a background level of 1.0 and a maximum light level of 11.5 (for toluene at the aqueous saturation limit). Data were given for 16 compounds dissolved at their aqueous saturation limit. The concentrations of these chemicals were not measured, and consequently it was decided not to use these luminescence data for regression analysis. A subset of 12 compounds had also been tested at a known concentration of 0.1 mM and these data were analyzed.

A weak relationship was found based on $\log P$ and $(\log P)^2$ (Figure 2):

$$\begin{aligned}\log ind &= 3.06 \log P - 0.534 (\log P)^2 - 2.57\\ n &= 12, \%V : 44. \qquad (3)\end{aligned}$$

The two $\log P$ terms were statistically significant, but not to a high degree ($p < 0.05$). The constant was not significantly different from zero ($p = 0.181$), but forcing the regression through zero reduced the percentage of variance explained (38%). The fit of the regression is poor, with a low percentage of the variance explained. Examination of a plot of fitted values against residuals shows clustering of adjacent residuals.

Table 1. Physical/chemical data used for deriving regression equations.

	log P	$(\log P)^2$	LUMO (eV)	HOMO (eV)	Molar volume ($\mathrm{cm}^3\ \mathrm{mol}^{-1}$)	Equation[a]
1,2,4-Trimethylbenzene (124TMB)	3.70	13.69	3.45	−11.95		4, 5, 7, 8
1,2,4,5-Tetramethylbenzene (1245TMB)	4.00	16.00	3.41	−11.61		4
Naphthalene (NAP)	3.30	10.89	1.93	−11.11		4, 5, 7, 8
1-Methylnaphthalene (1MNAP)	3.87	14.98	1.87	−10.79		4, 5, 7, 8
2-Methylnaphthalene (2MNAP)	3.86	14.90	1.90	−10.91		4, 5, 7, 8
1-Ethylnaphthalene (1ENAP)	4.39	19.27	1.92	−10.72		4, 5, 7, 8
Hexane (C6)	3.90	15.21	6.63	−14.18		4, 5
Cyclohexane (CYC)	3.44	11.83	6.49	−13.68		4, 5
Methylcyclohexane (MCYC)	3.61	13.03	6.45	−13.52		4, 5
Ethylbenzene (EBEN)	3.15	9.92	3.72	−12.80		4, 5, 7, 8
n-butylbenzene (BBEN)	4.38	19.18	3.74	−12.80		4, 5, 7, 8
Isopropylbenzene (IPB)	3.66	13.40	3.73	−12.80		4, 5, 7
Decalin (DEC)	4.83	23.33	6.27	−12.55		4, 5
1,2-diethylbenzene (12DEB)	3.72	13.84	3.80	−12.31		4, 5
1,2,3,4-Tetramethylbenzene (1234TMB)	3.98	15.84	3.41	−11.87		4, 5
Ethylcyclohexane (ECYC)	4.21	17.72	6.44	−13.37		4, 5
Cyclohexene (CYCE)	2.86	8.18	4.43	−12.92		4, 5
Fluorobenzene (FBEN)	2.27	5.15	3.56	−13.30		4–6
Chlorobenzene (CBEN)	2.89	8.35	3.54	−12.78		4–6
Aniline (ANI)	0.90	0.81	4.11	−11.06		4–6
Pyridine (PYR)	0.63	0.40	3.99	−12.46		4–6
Benzaldehyde (BZA)	1.48	2.19	1.83	−12.87		4–6
Benzyl alcohol (BA)	1.05	1.10	3.62	−12.82		4–6
Phenol (PHE)	0.65	0.42	3.87	−12.41		4–6
Trichloroethylene (TCE)	2.61	6.81	2.12	−12.91		4–6
Tetrachloroethylene (TeCE)	3.40	11.56	1.62	−12.72		4–6
Pentachloroethane (PCEt)	3.22	10.37	0.66	−14.18		4–6
1,2-Dichloroethane (12DCEt)	1.48	2.19	2.25	−14.06		4, 6
Trifluorotoluene (TFT)	3.01	9.06	2.58	−14.42		4–6
2-Bromotoluene (2BT)	3.62	13.10	3.41	−11.85		4–6
Benzothiophene (BTP)	3.12	9.73	3.03	−10.54		4–6
1-Indanone (1IN)	1.88	3.53	2.01	−12.44		4–6
Indole (IND)	2.14	4.58	3.35	−10.54		4–6
4-Tolualdehyde (4TA)	2.09	4.37	1.81	−12.42		4–6
Tribromoethylene (TBE)	3.20	10.24	2.58	−11.70		4–6
1,1,2,2-Tetrachloroethane (1122TCEt)	2.90	8.41	1.19	−13.98		4–6
Toluene (TOL)	2.73	7.45	3.74	−12.89		1–5, 7, 8
Benzene (BEN)	2.13	4.54	4.07	−13.89		1–5, 7, 8
2-Xylene (2XYL)	3.12	9.73	3.61	−12.46		1–3
3-Xylene (3XYL)	3.20	10.24	3.66	−12.55		1–3
4-Xylene (4XYL)	3.15	9.92	3.54	−12.14		1–5, 7, 8
2-Chlorotoluene (2CT)	3.42	11.70	3.34	−12.49		1–6
3-Chlorotoluene (3CT)	3.28	10.76	3.35	−12.59		1–3
4-Chlorotoluene (4CT)	3.33	11.09	3.36	−12.24		1–3
2-Methylbenzyl alcohol (2MBA)	1.65	2.72	3.50	−12.45		1, 2
3-Methylbenzyl alcohol (3MBA)	1.60	2.56	3.56	−12.56		1, 2

Table 1. Continued

	log P	$(\log P)^2$	LUMO (eV)	HOMO (eV)	Molar volume ($\mathrm{cm^3\ mol^{-1}}$)	Equation[a]
2-Nitrotoluene (2NT)	2.30	5.29	1.46	−12.86		1, 2
3-Nitrotoluene (3NT)	2.42	5.86	1.43	−12.90		1, 2
4-Nitrotoluene (4NT)	2.37	5.62	1.42	−12.87		1, 2
2-Ethyltoluene (2ET)	3.53	12.46	3.69	−12.40		3
3-Ethyltoluene (3ET)	3.88	15.05	3.73	−12.47		3
4-Ethyltoluene (4ET)	3.90	15.21	3.66	−12.06		3
Pentane (C5)					115.30	9
Hexane (C6)					126.70	9
Heptane (C7)					146.50	9
Octane (C8)					162.50	9
Nonane (C9)					207.20	9
Decane (C10)					229.40	9

[a] “Equation" column lists which regression equations a chemical has been used in.

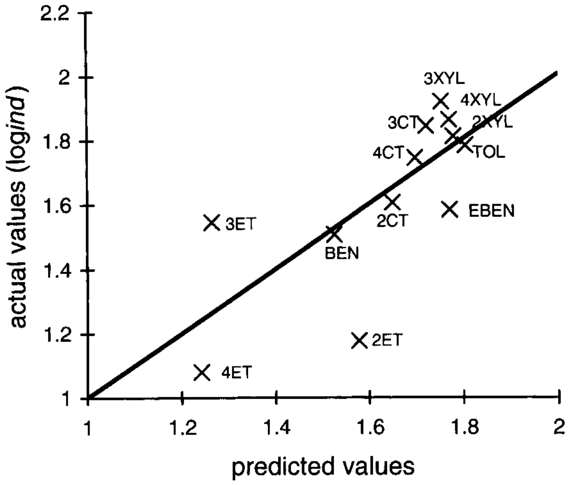

Figure 2. Model for biosensor B (Equation (2)): plot of actual values against predicted values. Line shows $x = y$, i.e., ideally fitted model. Full compound names listed in Tables 1 and 2.

Removing the three chlorinated non-hydrogen compounds from the regression left nine remaining compounds (all aromatic hydrocarbons); however reanalysis of the reduced set showed no improvement.

Biosensor C: Selifonova and Eaton (1996)

This dataset is the largest and hence most suitable for QSAR analysis. Results were given for minimal and maximal induction after 100 and 250 minutes; only the results for maximal induction after 250 minutes have been analyzed, as induction after 100 minutes was very low in comparison. Data were normalized before analysis by dividing the reported induction value by the concentration of the chemical (μM), and then scaled by taking the logarithm to base ten.

The data were initially analyzed as a complete set:

$$\begin{aligned} \log(ind) &= 4\log P - 0.645(\log P)^2 - 5.29 \\ n &= 42, \%V : 59. \end{aligned} \tag{4}$$

The points corresponding to 1,2,4,5-tetramethylbenzene and 1,2-dichloroethane were outliers. The regression was reanalyzed without these points:

$$\begin{aligned} \log(ind) &= 3.62\log P - 0.57(\log P)^2 \\ &\quad -0.1595 LUMO \\ n &= 40, \%V : 72. \end{aligned} \tag{5}$$

The terms in log P and $(\log P)^2$ are highly significant ($p < 0.001$), and LUMO energy is also significant ($p = 0.044$). The regression would usually be considered acceptable with over 70 per cent of variance explained; however, a plot of predicted against actual values indicated that a few points have undue leverage. The residuals plot showed marked clustering of adjacent residuals, and large residuals associated with high fitted values. This indicates that Equation (5) is not a good model of the dataset.

Consequently, the data were divided into two sets, hydrocarbons and non-hydrocarbons, and reanalyzed. (In this paper “hydrocarbon” is defined strictly as a compound containing only carbon and hydrogen atoms, and a “non-hydrocarbon” is considered to be any compound containing any other elements, even if based on a hydrocarbon molecule such as e.g., a

Table 2. Biological data used in regression equations. (All data taken from previously published sources.)

Biosensor:	A	B	C	D
	$\log(K_{1/2})$[a]	log(*ind*/μM)[b]	log (*ind*/μM)[b]	100*ind/ind*$_{\text{oct}}$[c]
1,2,4-Trimethylbenzene (124TMB)			0.68	
1,2,4,5-Tetramethylbenzene (1245TMB)			−2.02	
Naphthalene (NAP)			2.23	
1-Methylnaphthalene (1MNAP)			1.08	
2-Methylnaphthalene (2MNAP)			1.04	
1-Ethylnaphthalene (1ENAP)			−0.01	
Cyclohexane (CYC)			0.00	
Methylcyclohexane (MCYC)			−0.16	
Ethylbenzene (EBEN)		1.58	1.68	
n-Butylbenzene (BBEN)			−1.30	
Isopropylbenzene (IPB)			1.69	
Decalin (DEC)			−0.30	
1,2-Diethylbenzene (12DEB)			0.40	
1,2,3,4-Tetramethylbenzene (1234TMB)			0.15	
Ethylcyclohexane (ECYC)			−0.33	
Cyclohexene (CYCE)			−0.17	
Fluorobenzene (FBEN)			0.00	
Chlorobenzene (CBEN)			1.08	
Aniline (ANI)			−2.10	
Pyridine (PYR)			−2.70	
Benzaldehyde (BZA)			−0.59	
Benzyl alcohol (BA)			−1.96	
Phenol (PHE)			−2.40	
Trichloroethylene (TCE)			0.46	
Tetrachloroethylene (TeCE)			1.57	
Pentachloroethane (PCEt)			0.54	
1,2-Dichloroethane (12DCEt)			−2.40	
Iodobenzene (IB)			0.01	
Trifluorotoluene (TFT)			1.71	
2-Bromotoluene (2BT)			1.30	
2-Iodotoluene (2IT)			1.95	
Benzothiophene (BTP)			0.83	
1-Indanone (1IN)			0.75	
Indole (IND)			−0.35	
4-Tolualdehyde (4TA)			−0.52	
Tribromoethylene (TBE)			1.92	
1,1,2,2-Tetrachloroethane (1122TCEt)			0.40	
Toluene (TOL)	2.23	1.78	1.36	
Benzene (BEN)	2.67	1.51	0.61	
2-Xylene (2XYL)	1.66	1.81		
3-Xylene (3XYL)	1.59	1.91		
4-Xylene (4XYL)	2.01	1.86	1.30	
2-Chlorotoluene (2CT)	1.80	1.60	1.95	
3-Chlorotoluene (3CT)	1.90	1.84		
4-Chlorotoluene (4CT)	1.91	1.74		
2-Methylbenzyl alcohol (2MBA)	3.47			
3-Methylbenyzl alcohol (3MBA)	3.26			
2-Nitrotoluene (2NT)	2.79			
3-Nitrotoluene (3NT)	3.04			

Table 2. Continued

Biosensor:	A $\log(K_{1/2})^a$	B $\log(ind/\mu\text{M})^b$	C $\log\ (ind/\mu\text{M})^b$	D $100 ind/ind_{\text{oct}}{}^c$
4-Nitrotoluene (4NT)	2.78			
2-Ethyltoluene (2ET)		1.18		
3-Ethyltoluene (3ET)		1.54		
4-Ethyltoluene (4ET)		1.08		
Pentane (C5)				13
Hexane (C6)			−0.23	44
Heptane (C7)				81
Octane (C8)				100
Nonane (C9)				100
Decane (C10)				69
3-Methylheptane (3MH)				36

[a] Data are log-transformed saturation constants, $K_{1/2}$.

[b] Data are log-transformed induction values, *ind*, where *ind* represents increase in signal relative to control, normalized by concentration (μM).

[c] Data are induction values normalized to response against octane.

substituted benzene.) Non-hydrocarbon compounds showed a strong relationship with $\log P$:

$$\begin{aligned} \log(ind) &= 1.42 \log P - 3.43 \\ n &= 21,\ Q^2 : 0.86, \%V : 88. \qquad (6) \end{aligned}$$

This is a good model, with a high proportion of the variance explained. Plotting log (*ind*) against $\log P$ showed an excellent fit (Figure 3); the residuals plot was acceptably random. The fit for these compounds has clearly been improved by modelling them separately from hydrocarbon compounds.

In contrast, no significant relationship was found for the set of all (aliphatic + aromatic) hydrocarbons. To attempt to improve the relationship, the data were remodelled with the points for the six aliphatic hydrocarbons and 1,2,4,5-tetramethylbenzene removed:

$$\begin{aligned} \log(ind) &= 7.17 \log P - 1.23(\log P)^2 \\ &\quad -0.443 LUMO - 7.3 \\ n &= 13,\ Q^2 : 0.78, \%V : 84. \qquad (7) \end{aligned}$$

Examination of the residuals plot shows that isopropylbenzene is an outlier, its predicted value being too low. Recalculating the data without isopropylbenzene (Figure 4):

$$\begin{aligned} \log(ind) &= 6.20 \log P - 1.09(\log P)^2 \\ &\quad -0.544 LUMO - 5.34 \\ n &= 12,\ Q^2 : 0.90, \%V : 95. \qquad (8) \end{aligned}$$

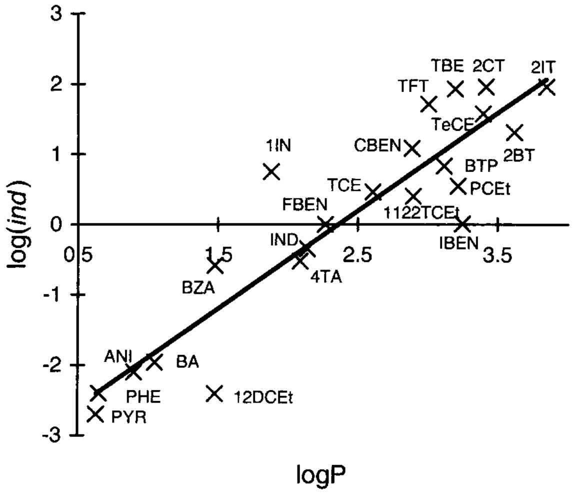

Figure 3. Model for biosensor C: non-hydrocarbon compounds only (Equation (6)). Log(*ind*) plotted against hydrophobicity ($\log P$). Full compound names listed in Tables 1 and 2.

All the parameters are highly significant ($p < 0.001$, except for the constant: $p = 0.007$). A plot of fitted values against residuals shows slight clustering of adjacent points; however examination of the plot of actual against predicted values does not show any problem with the regression.

Biosensor D: Sticher et al. (1997)

Induction data for 24 compounds (23 hydrocarbons and dicyclopropylketone) are reported, normalized to the response for octane (100%). However only

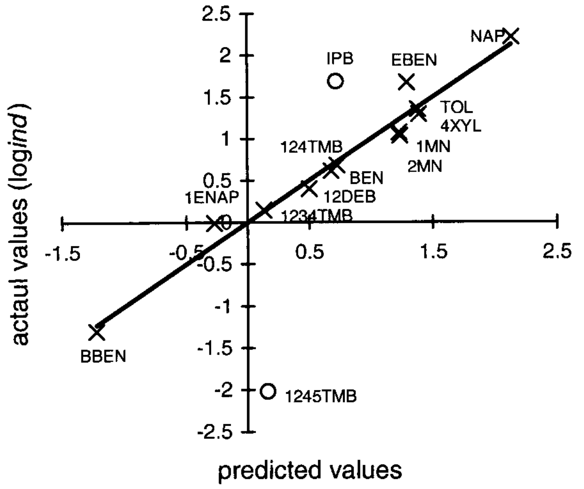

Figure 4. Model for biosensor C: aromatic hydrocarbons only, data for isopropylbenzene and 1,2,4,5-tetramethylbenzene omitted in calculating regression (Equation (8). Plot of actual values against predicted values; line shows $x = y$, i.e., ideally fitted model. Circles represent omitted data: isopropylbenzene and 1,2,4,5-tetramethylbenzene. Full compound names listed in Tables 1 and 2.

seven of these compounds (n-alkanes from pentane to decane, and 3-methylheptane) showed induction significantly greater than background. This would usually be too few data points to attempt to derive a QSAR, but in this case the biological response shows a trivial dependence on carbon chain length, with maximum induction reached at a chain length of eight to nine and a rapid decline in luminescence thereafter. A slightly better fit was obtained by modelling the data (without 3-methylheptane) against molar volume, rather than simple chain length (Figure 5):

$$\begin{aligned} ind &= 7.46vol - 0.0202vol^2 - 578 \\ n &= 6,\ Q^2 : 1.00,\ \%V : 100 \end{aligned} \qquad (9)$$

vol is the molar volume in $\mathrm{cm}^3\ \mathrm{mol}^{-1}$.

Discussion

Biosensor C: The removal of 1,2,4,5-tetramethylbenzene and 1,2-dichloroethane from equation 5 as outliers has a scientific basis. Inspection of the data reveals that the light levels for both are very low (−2.02 and −2.40 respectively), and much lower than for similar compounds, e.g., the isomer 1,2,3,4-tetramethylbenzene has a light level of 0.15. As these values are of the logarithm of normalized induction it shows that induction is more than two orders of magnitude lower than that observed for apparently similar hydrocarbon compounds. This implies that the compounds are not inducers of the *ipb* pathway: they are not part of the homologous series of compounds on which the model is based. It is fundamental to the development of QSARs that they are based on a genuinely homologous series (i.e., the chemicals elicit a biological response through the same mode of action), so it is important to eliminate chemicals that do not form part of the series. Homology is usually assumed from structural similarity, but in this instance is implicit in the fact that there is only one source of bioluminescence in the system. A probable explanation for the lack of response of 1,2,4,5-tetramethylbenzene and 1,2-dichloroethane is that they are the wrong shape to be recognized by the regulatory protein. The physical/chemical data used in the regression equations do not contain any steric or shape information, and hence the QSAR is unable to cope with these compounds.

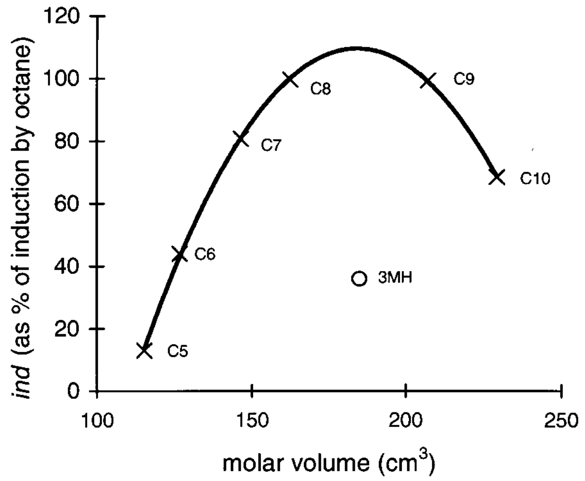

Figure 5. Model for biosensor D: *ind* plotted against molar volume (Equation (9). Circle represents 3-methylheptane (point omitted in calculating regression). Full compound names listed in Tables 1 and 2.

It is clear that Equation (7) is improved by removal of isopropylbenzene, which is a slight outlier, giving Equation (8). It is unexpected that this should be the case: isopropylbenzene is the archetypal degradation substrate for this pathway, and hence should definitely be part of the series of compounds which this biosensor responds to. (It should be noted that the model containing isopropylbenzene, Equation (7), is still highly significant, with good adjusted R^2 and Q^2 values.) The reason for this point being an outlier is not known.

Biosensor D: 3-methylheptane clearly does not fall into the same set as the *n*-alkanes. The value predicted (111) using Equation (9) is very different from the actual value of 36. Further testing with a wider range of branched alkanes would be necessary to determine if, like *n*-alkanes, they formed a QSAR that was dependent on molar volume.

General

Abril et al. (1989) used *lacZ* reporter gene biosensors to study the specificity of the regulatory proteins of the TOL pathway. (The specificity of the degradative enzymes was also studied.) They found that the XylR protein had a broad effector specificity, responding to a range of mono-, di-, and trisubstituted alkyl and chlorobenzenes. The results correspond well with those for biosensors A and B, with the exception that Abril et al. did not observe induction by benzene. Abril et al. concluded that substitution of the benzene ring is necessary for activation of XylR; further chloro- or alkyl substituents also led to activation; and that benzaldehydes did not cause activation except for 4-chlorobenzaldehyde.

A QSAR approach offers a more systematic and powerful way of examining this category of data than simple inspection. QSARs are not always generated with the intention of building a model that is necessary to make predictions for unknown chemicals – they can be used to provide information on a biological system (Hermens 1995). Layton et al. (1999) used a QSAR approach to validate the response of a *lux*-marked toxicity biosensor: they demonstrated that toxicity of non-polar narcotic compounds could be predicted by hydrophobicity. This relationship is widely established in ecotoxicity testing (Cronin and Dearden 1995) and showed that the toxicity biosensor responded in accordance with typical aquatic test organisms. Similarly, the regression equations developed in this study can give clues to the underlying factors controlling specificity. Hydrophobicity is shown to be the most important descriptor for the two TOL plasmid-derived biosensors, A and B. Hydrophobicity could conceivably affect the strength of effector binding to the XylR protein; it might also affect specificity by controlling the rate at which compounds could pass across the cell membrane (i.e., bioavailability). Equation 2 also contains a weakly significant (p = 0.031) term in LUMO energy. LUMO energy can be considered a measure of a compound's electron affinity, or alternatively susceptibility to nucleophilic attack (Lynam et al. 1998): this implies that the interaction of the XylR protein with the chemicals is important, not just membrane transport. It is not surprising that the response can be predicted using a single chemical descriptor: the chemicals tested were chosen because they were known inducers of the P_u promoter, and hence form a homologous set. The linear free energy relationship hypothesis states that small changes in the physical/chemical properties of chemicals within a set cause linear changes in the free energy of a reaction (such as binding to a protein), enabling a linear relationship to be derived (Okey and Stensel 1996). Because the chemicals form a known series, trends within that series are governed by very simple properties; conversely, a set including chemicals from outwith that series might require additional descriptors and more complicated models to predict the response.

A direct comparison of the regression equations derived for the two toluene biosensors A and B shows that there are differences between the results. As stated above hydrophobicity (log P) is the main term for both A and B; however the relationship is much stronger for biosensor A and can be improved further by the addition of a term for LUMO energy. The reason for the difference is not definitely known, but it can be suggested that one possibility is that the biological response of the two biosensors was not the same. The responses for biosensor A cover a range from 1.59 to 3.47 on a logarithmic scale – almost two orders of magnitude. However, the responses for biosensor B only cover a range of 1.08 to 1.91, i.e., less than one order of magnitude. As a result a weaker relationship would be expected for biosensor B. A second possible explanation is that the data for biosensor B were only measured at a single concentration (0.1 mM), whereas the data for biosensor A were measured over a range of concentrations and the summary statistic $K_{1/2}$ was used. The response of the biosensor is only likely to be linear over a narrow concentration range; as a result the biological data for A may be more suitable for response modelling than the data for B.

The response of the TOL plasmid-based biosensors can be compared to the response of the *ipb*-based biosensor C. Equation (8) shows that hydrophobicity and LUMO energy are both highly significant for the set of aromatic hydrocarbons. This is a very similar result to the TOL biosensor A (Equation (2)): similar factors appear to regulate specificity. This is reasonable: hydrocarbon degradation genes are both evolutionarily and functionally conserved, and *ipb* genes show homology with toluene dioxygenase genes (Williams and

Sayers 1994; Aoki et al. 1996). It is interesting that the aromatic hydrocarbons have a quadratic relationship with log P, whereas the non-hydrocarbons have a linear relationship: however, this could be because the log P values for the hydrocarbons are mostly higher than those of the non-hydrocarbons. Compounds with low hydrophobicity may be expected to diffuse slowly across cell membranes: hence it is possible that the response of low log P compounds is dominated by bioavailability, resulting in a positive relationship with log P.

QSARs can also be used to examine the biological behaviour of individual chemicals, or even chemical classes. Outliers can indicate the limits of applicability of a QSAR (Lipnick 1991): 3-methylheptane is shown to fall into a different set from the n-alkanes for induction of *alkB* (Equation (9)). 1,2,3,5-tetramethylbenzene and 1,2-dichloroethane stimulate too little light production to be considered part of the set of compounds recognized by the regulatory elements of the *ipb* pathway (Equation (5)); in contrast, the archetypal substrate isopropylbenzene stimulates too much (Equation (8)). Comparison of Equations (4)-(8) clearly shows that three differently-acting *groups* of chemicals can be distinguished, apart from the outliers: non-hydrocarbons are dependent on log P alone, but aromatic hydrocarbons require a term in LUMO as well as in (log $P)^2$. Aliphatic hydrocarbons are further shown to have a clearly different biological response from aromatic hydrocarbons.

An important point is to what extent these results relate to actual biodegradation potential of the compounds concerned: could the regression equations obtained be considered QSBRs, quantitative structure-biodegradation relationships? If so, this approach is of great potential: the existing models could be improved (by testing a wider range of substrates, and possibly by including a wider range of physical/chemical parameters in the analysis), and construction of new *lux* or *luc* fusions would allow many different degradation pathways and groups of compounds to be tested. This is of particular importance because a lack of reproducible quantitative data on biodegradation is one of the factors that limits the development of QSBRs (Degner et al. 1991).

The sensors have genes for bioluminescence fused to genes that code for regulatory proteins which control the level of expression of the hydrocarbon degradation pathways. Hence the biosensors will respond to gratuitous inducers – compounds which induce gene expression but are not actually substrates of the initial enzyme. Thus the QSARs derived are not QSBRs. For example, biosensors A and B both gave a response to benzene, which is not metabolized by the TOL pathway. Conversely biosensor D, based on genes from the OCT plasmid, showed an extremely narrow specificity: responding only to n-alkanes from C5 to C10, and to 3-methylheptane. The OCT genes are known to hydroxylate n-alkanes up to C12, and very similar gene systems can hydroxylate a wide range of compounds including cyclic aliphatics and alkylbenzenes (van Beilen et al. 1994), so there is a discrepancy between the known degradative ability of the pathway and the *lux* response. In this case the regulation of the pathway appears more specific than the biodegradation capability.

The biosensors described in this study have regulatory elements which recognise the initial, undegraded hydrocarbons or related compounds. However it is possible to produce biosensors that are based on systems in which the regulatory element recognizes an intermediate of degradation – e.g., King et al. (1990) describe a *lux*-marked biosensor based on the NAH pathway. The regulatory protein in this pathway responds to the intermediate, salicylate, rather than the parent, naphthalene. Hence this biosensor is genuinely measuring biotransformation, as the parent molecule must have been oxidized in order to induce bioluminescence. Biosensors of this type could be used to produce structure-activity relationships that do actually predict biodegradation.

Conclusion

Structure-activity relationships can be generated that explain the specificity of regulatory proteins for hydrocarbon degradation pathways, with high adjusted R^2 and Q^2 values. Hydrophobicity is the most important parameter in these relationships (possibly because it governs initial uptake by the cell); LUMO energy is also significant in several of the models that have been generated. These relationships do not predict biodegradation, because the set of inducer compounds for a pathway is not identical to the set of possible substrates; however, they nonetheless increase understanding of the specificity of regulation and hence the mechanisms of hydrocarbon degradation.

Microbial biosensors (such as *lux* or *luc* marked bacteria) are well suited to producing data for simple modelling: the assays are quick and reproducible, allowing large numbers of chemicals to be tested. This is ideal for the requirements of QSAR modelling. This study demonstrates that the biosensors described, or biosensors based on similar principles, could be used to generate data intended specifically for QSARs explaining or predicting aspects of hydrocarbon degradation pathways. If biosensors were used that respond to the intermediates of degradation, structure-biodegradation models could be developed that could be used to predict the biodegradation potential of novel or untested compounds.

References

Abril MA, Michan C, Timmis KN & Ramos JL (1989) Regulator and enzyme specificities of the TOL plasmid-encoded upper pathway for degradation of aromatic hydrocarbons and expansion of the substrate range of the pathway. J. Bacteriol. 171: 6782–6790

Aoki H, Kimura T, Habe H, Yamane H, Kodama T & Omori T (1996) Cloning, nucleotide sequence, and characterization of the genes encoding enzymes involved in the degradation of cumene to 2-hydroxy-6-oxo-7-methylocta-2,4-dienoic acid in *Pseudomonas fluorescens* IP01. Journal Of Fermentation and Bioengineering 81: 187–196

Applegate BM, Kehrmeyer SR & Sayler GS (1998) A chromosomally-based *tod-luxCDABE* whole-cell reporter for benzene, toluene, ethylbenzene, and xylene (BTEX) sensing. Appl. Environ. Microbiol. 64: 2730–2735

Atlas RM, Sayler GS, Burlage R & Bej AK (1992) Molecular approaches for environmental monitoring of microorganisms. Biotechniques 12: 706–717

Barkay T, Nazaret S & Jeffrey W (1995) Degradative genes in the environment. In: Young LY & Cerniglia CE (Eds) Microbial Transformation and Degradation of Toxic Organic Chemicals (pp. 545–577). Wiley-Liss, New York

Belkin S, Smulski DR, Dadon S, Vollmer AC, Van Dyk TK & LaRossa RA (1997) A panel of stress-responsive luminous bacteria for the detection of selected classes of toxicants. Wat. Res. 31: 3009–3016

Cronin MTD & Dearden JC (1995) QSAR in toxicology. 1. Prediction of aquatic toxicity. Quant. Struct. Act. Relat. 14: 1–7.

Degner P, Nendza M & Klein W (1991) Predictive QSAR models for estimating biodegradation of aromatic compounds. Sci. Total Environ. 109: 253–259

Eastcott L, Shiu WY & Mackay D (1988) Environmentally relevant physical-chemical properties of hydrocarbons: a review of data and development of simple correlations. Oil & Chemical Pollution 4: 191–216

Eriksson L, Johansson E & Wold S (1997) Quantitative structure-activity relationship model validation. In: Chen F & Schüürman G (Eds) Quantitative Structure-Activity Relationships in Environmental Sciences (pp 381–397). SETAC, Pensacola Hansch C, Leo A & Hoekman D (1995) Exploring QSAR: Hydrophobic, Electronic, and Steric Constants. ACS, Washington DC

Hermens, JLM (1995) Prediction of environmental toxicity based on structure-activity relationships using mechanistic information. Sci. Total Environ. 171: 235–242.

Ikariyama Y, Nishiguchi S, Koyama T, Kobatake E & Aisawa M (1997) Fiber-optic-based biomonitoring of benzene derivatives by recombinant *E. coli* bearing luciferase gene-fused TOL-plasmid immobilized on the fiber-optic end. Anal. Chem. 69: 2600–2605

King JMH, DiGrazia PM, Applegate B, Burlage R, Sanseverino J, Dunbar P, Larimer, F & Sayler GS (1990) Rapid, sensitive bioluminescence reporter technology for naphthalene exposure and biodegradation. Science 249: 778–781

Kragelund L, Christoffersen B, Nybroe O & de Bruijn FJ (1995) Isolation of *lux* reporter gene fusions in *Pseudomonas fluorescens* DF57 inducible by nitrogen or phosphorus starvation. FEMS Microbiol. Ecol. 17: 95–106

Layton AC, Gregory B, Schultz TW & Sayler GS (1999) Validation of genetically engineered bioluminescent surfactant resistant bacteria as toxicity assessment tools. Ecotoxicol. Environ. Safety 43: 222–228

Layton AC, Muccini M, Ghosh MM & Sayler GS (1998) Construction of a bioluminescent reporter strain to detect polychlorinated biphenyls. Appl. Environ. Microbiol. 64: 5023–5026

Lipnick RL (1991) Outliers: their origin and use in the classification of molecular mechanisms of toxicity. Sci. Total Environ. 109/110: 131–153

Lynam MM, Kuty M, Damborsky J, Koca J & Adriaens P (1998) Molecular orbital calculations to describe microbial reductive dechlorination of polychlorinated dioxins. Environ. Toxicol. Chem. 17:988–997

Marqués S & Ramos JL (1993) Transcriptional control of the *Pseudomonas putida* TOL plasmid catabolic pathways. Molecular Microbiol. 9: 923–929

Meighen EA (1988) Enzymes and genes from the *lux* operons of bioluminescent bacteria. Ann. Rev. Microbiol. 42: 151–176

Okey RW & Stensel HD (1996) A QSAR-based biodegradability model: a QSBR. Wat. Res. 30: 2206–2214

Prest AG, Winson MK, Hammond JRM & Stewart GSAB (1997) Construction and application of a *lux*-based nitrate biosensor. Lett. Appl. Microbiol. 24: 355–360

Selifonova O, Burlage R & Barkay T (1993) Bioluminescent sensors for detection of bioavailable Hg(II) in the environment. Appl. Environ. Microbiol. 59: 3083–3090

Selifonova O & Eaton RW (1996) Use of an *ipb-lux* fusion to study regulation of the isopropylbenzene catabolism operon of *Pseudomonas putida* RE204 and to detect hydrophobic pollutants in the environment. Appl. Environ. Microbiol. 62: 778–783

Sticher P, Jaspers MCM, Stemmler K, Harms H, Zehnder AJB & Van der Meer JR (1997) Development and characterization of a whole-cell bioluminescent sensor for bioavailable middle-chain alkanes in contaminated groundwater samples. Appl. Environ. Microbiol. 63: 4053–4060

Tescione L & Belfort G (1993) Construction and evaluation of a metal-ion biosensor. Biotechnol. Bioeng. 42: 945–952

van Beilen JB, Wubbolts MG & Witholt B (1994) Genetics of alkane oxidation by *Pseudomonas oleovorans*. Biodegradation 5: 161–174

Willardson BM, Wilkins JF, Rand TA, Schupp JM, Hill KK, Keim P & Jackson PJ (1998) Development and testing of a bacterial biosensor for toluene-based environmental contaminants. Appl. Environ. Microbiol. 64: 1006–1012

Williams PA & Sayers JR (1994) The evolution of pathways for aromatic hydrocarbon oxidation in *Pseudomonas*. Biodegradation 5: 195–217

Zylstra GJ & Gibson DT (1991) Aromatic hydrocarbon degradation: a molecular approach. Genetic Engineering 13: 183–203