Quantitative Analysis of Substrate Specificity of Haloalkane Dehalogenase LinB from *Sphingomonas paucimobilis* UT26†

Jan Kmunícek,‡,§ Kamila Hynková,‡,§ Tomás Jedlicka,‡ Yuji Nagata,|| Ana Negri,⊥ Federico Gago,⊥ Rebecca C. Wade,@ and Jirí Damborský*,‡

National Centre for Biomolecular Research, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic, Department of Environmental Life Sciences, Graduate School of Life Sciences, Tohoku University, 2-1-1 Katahira, Sendai 980-8577, Japan, Department of Pharmacology, University of Alcalá, E-28871 Alcalá de Henares, Madrid, Spain, and EML Research, Villa Bosch, Schloss-Wolfsbrunnenweg 33, D-69118 Heidelberg, Germany

Received September 27, 2004; Revised Manuscript Received December 16, 2004

ABSTRACT: Haloalkane dehalogenases are microbial enzymes that cleave a carbon−halogen bond in halogenated compounds. The haloalkane dehalogenase LinB, isolated from *Sphingomonas paucimobilis* UT26, is a broad-specificity enzyme. Fifty-five halogenated aliphatic and cyclic hydrocarbons were tested for dehalogenation with the LinB enzyme. The compounds for testing were systematically selected using a statistical experimental design. Steady-state kinetic constants K_m and k_{cat} were determined for 25 substrates that showed detectable cleavage by the enzyme and low abiotic hydrolysis. Classical quantitative structure−activity relationships (QSARs) were used to correlate the kinetic constants with molecular descriptors and resulted in a model that explained 94% of the experimental data variability. The binding affinity of the tested substrates for this haloalkane dehalogenase correlated with hydrophobicity, molecular surface, dipole moment, and volume:surface ratio. Binding of the substrate molecules in the active site pocket of LinB depends nonlinearly on the size of the molecules. Binding affinity increases with increasing substrate size up to a chain length of six carbon atoms and then decreases. Comparative binding energy (COMBINE) analysis was then used to identify amino acid residues in LinB that modulate its substrate specificity. A model with three statistically significant principal components explained 95% of the experimental data variability. van der Waals interactions between substrate molecules and the enzyme dominated the COMBINE model, in agreement with the importance of substrate size in the classical QSAR model. Only a limited number of protein residues (6−8%) contribute significantly to the explanation of variability in binding affinities. The amino acid residues important for explaining variability in binding affinities are as follows: (i) first-shell residues Asn38, Asp108, Trp109, Glu132, Ile134, Phe143, Phe151, Phe169, Val173, Trp207, Pro208, Ile211, Leu248, and His272, (ii) tunnel residues Pro144, Asp147, Leu177, and Ala247, and (iii) second-shell residues Pro39 and Phe273. The tunnel and the second-shell residues represent the best targets for modulating specificity since their replacement does not lead to loss of functionality by disruption of the active site architecture. The mechanism of molecular adaptation toward a different specificity is discussed on the basis of quantitative comparison of models derived for two protein family members.

Haloalkane dehalogenases (EC 3.8.1.5) are microbial enzymes that cleave a carbon−halogen bond in halogenated alkanes, cycloalkanes, alkenes, selected ethers, and alcohols (*1*). Haloalkane dehalogenases act by a hydrolytic mechanism involving use of a water molecule as the only cosubstrate. The enzymes can be used for the protection of the environment, e.g., in bioremediation of contaminated areas (*2*), in removal of intermediates of chemical syntheses (*3*), and in biosensors. However, since haloalkane dehalogenases show low or no activity toward some industrially important substrates, they are also targets for protein engineering studies.

Structurally, haloalkane dehalogenases belong to the α/β-hydrolase superfamily (*4*). The core of each enzyme is similar and consists of two different domains: the α/β-fold (main) domain, which is conserved in all α/β-hydrolases, and the so-called cap domain. The main domain is composed of a β-sheet made up of eight β-strands surrounded by six α-helices. The cap domain is composed of an additional bundle of five α-helices connected by loops. The active site is located between these two domains in an internal, predominantly hydrophobic cavity and can be reached from the solvent through a tunnel. At least three different groups of haloalkane dehalogenases can be distinguished according

† This work was supported by a grant from the Czech Ministry of Education (J07/98:14310005 to J.D.) and North Atlantic Treaty Organization Grant EST.CLG.980504 (F.G., R.C.W., and J.D.). R.C.W. gratefully acknowledges the support of the Klaus Tschira Foundation. J.D. is a recipient of the EMBO/HHMI Scientist Fellowship.

* To whom correspondence should be addressed. Fax: +420-5-49492556. E-mail: jiri@chemi.muni.cz.

‡ Masaryk University.

§ These authors contributed equally.

|| Tohoku University.

⊥ University of Alcalá.

@ EML Research.

10.1021/bi047912o CCC: $30.25
Published on Web 02/11/2005

to their different substrate specificities (*5*). Each of these categories has its own representative with known three-dimensional structure: DhlA[1] from *Xanthobacter autotrophicus* GJ10 (*6*), DhaA from *Rhodococcus rhodochrous* NCIMB 13064 (*7*), and LinB from *Sphingomonas paucimobilis* UT26 (*8*). The ratio of active site volumes for these three representatives (DhlA:DhaA:LinB) was determined to be 1:2:2.5 (*5*). The distinct substrate specificities for the three classes of haloalkane dehalogenases are mainly due to differences in (i) the composition and geometry of the active site, (ii) the halide-stabilizing residues, and (iii) the entrance tunnel connecting the active site with the protein surface.

Quantitative structure−activity relationship (QSAR) approaches relate experimental or calculated structural properties of organic molecules to their biological activities. QSAR studies identify critical relationships between properties and the geometric and chemical characteristics of a molecular system. A number of models have been developed for enzymatic and microbial catalysis (for reviews, see refs *9*−*12*). Most of the published models are only qualitative, as a major obstacle to the development of quantitative models is the lack of uniformly measured data for large numbers of compounds. This prerequisite is essential if a reliable statistical analysis is to be attempted to validate the resulting model.

Comparative binding energy (COMBINE) analysis (*13*) is a technique for deriving QSARs from a set of three-dimensional structures of enzyme−ligand complexes. COMBINE analysis was originally applied to enzyme−inhibitor interactions in the drug design field, whereas its applicability to studying enzyme−substrate binding and to protein design has been tested more recently (*14*−*17*). Thus, a COMBINE model was constructed for DhlA (*14*) that quantitatively accounted for 91% (73% cross-validated) of the variance in the apparent dissociation constants of 18 substrates and identified the residues contributing most significantly to the substrate specificity of this haloalkane dehalogenase. Later, this model was further improved by the use of automated molecular docking techniques and quantum mechanical calculations in the construction of the enzyme−substrate complexes (*16*). In this report, we simultaneously apply classical QSAR techniques and COMBINE analysis to quantitatively analyze substrate specificity in the haloalkane dehalogenase LinB.

METHODS

A chemometric strategy consisting of the following steps was applied for optimal selection of the substrates for testing and for construction of robust quantitative models: (i) formulation of the class of similar compounds, (ii) multivariate characterization and definition of design variables, (iii) selection of a training set of representative compounds, (iv) experimental determination of kinetic constants, (v) classical QSAR derivation, and (vi) construction of COMBINE models.

Selection of Compounds for Testing. The starting class of halogenated compounds consisted of 196 chlorinated, brominated, iodinated, and fluorinated hydrocarbons compiled in our in-house database (http://loschmidt.chemi.muni.cz/peg/). Substances with physicochemical properties preventing reliable determination of kinetic constants under laboratory conditions or containing a substructure known to resist catalysis by the haloalkane dehalogenases were excluded: (i) compounds in the gas state under laboratory conditions, (ii) compounds for which the logarithm of the octanol−water partition coefficient is greater than 4, and (iii) fluorinated compounds, (iv) compounds with more than one halogen bound to a single carbon atom, and (v) compounds bearing a halogen substituent on an sp^2 carbon atom. The set of compounds entering the experimental design after this initial preselection comprised 116 compounds.

Multivariate Characterization and Definition of Design Variables. The structures of the halogenated substrates were built using the molecular modeling package Insight II, version 95 (Accelrys), and then prerefined by molecular mechanics optimization. Full energy minimization of the structures was achieved by the BFGS algorithm, as implemented in the semiempirical quantum mechanical program MOPAC (*18*) using the AM1 Hamiltonian and PRECISE stopping criteria. Molecular descriptors for multivariate characterization were calculated with TSAR coupled with VAMP, version 3.1 (Oxford Molecular). The set of 24 calculated descriptors was complemented by four physicochemical properties compiled from the Sigma-Aldrich handbook (Table 1).

Selection of a Training Set of Representative Compounds. Experimental design was used for selection of the training set. Principal component analysis (PCA) was applied to a data matrix containing 116 halogenated compounds (objects) and 28 physicochemical descriptors (independent variables). The data were centered and scaled to unit variance prior to PCA. Ten compounds detected as outliers in PCA were excluded from the data set to improve data homogeneity. Upon recalculation of the model, four latent variables (scores) that summarize the original variables in the data matrix were constructed and used as principal properties for a 2^4 factorial design.

Determination of Kinetic Constants and Reaction Products. Kinetic experiments were conducted with LinB purified to homogeneity as described previously (*19*, *20*). Michaelis−Menten kinetic constants were estimated by initial-velocity measurements. Gas chromatography was used for determination of substrate and product concentrations. A dehalogenation reaction was performed in 25 mL Reacti-Flasks closed by Mininert Valves. Ten milliliters of glycine buffer (pH 8.6) was mixed with seven different substrate concentrations. The highest concentration of substrate in the glycine buffer also served as an abiotic control. The reaction mixture was equilibrated for 30 min in a shaking water bath at 37 °C prior to initiation of the reaction. The enzymatic reaction

[1] Abbreviations: AM1, Austin model 1; COMBINE, comparative binding energy; *D*, density; DhaA, haloalkane dehalogenase from *Rhodococcus*; DhlA, haloalkane dehalogenase from *X. autotrophicus* GJ10; DIP, dipole moment; EST, sum of E-state indices; GC, gas chromatography; GC−MS, gas chromatography−mass spectrometry; HF, heat of formation; LinB, haloalkane dehalogenase from *S. paucimobilis* UT26; logP, logarithm of the octanol−water partition coefficient; LUMO, energy of the lowest occupied molecular orbital; MM, molecular mass; MR, molar refractivity; MV, molecular volume; MW, molecular weight; *n*, refractive index; PC, principal component; PCA, principal component analysis; PLS, partial least-squares projection to latent structures; POL, polarizability; QSAR, quantitative structure−activity relationship; Q^2, cross-validated correlation coefficient; *R*, molecular volume:surface ratio; R^2, correlation coefficient; SA, surface area; TE, total energy; VIP, variable importance in projection.

Table 1: Molecular Descriptors Used for Experimental Design and Model Construction

abbreviation	descriptor name	units	source
MW	molecular weight		handbook
bp	boiling point	°C	handbook
n	refractive index		handbook
D	density	g/mL	handbook
MM	molecular mass	g/mol	TSAR 3.1
MV1	molecular volume (tsar)	Å^3	TSAR 3.1
M1−3s	moments of inertia 1−3 (size)	$\times 10^{-39}$ g/cm^2	TSAR 3.1
M1−3l	principal axes of inertia 1−3 (length)	Å	TSAR 3.1
EV	ellipsoidal volume	Å^3	TSAR 3.1
logP1	octanol−water partition coefficient		TSAR 3.1
LIP	total lipole		TSAR 3.1
MR	molar refractivity		TSAR 3.1
PHI	shape flexibility index		TSAR 3.1
RAN	Randic topological index		TSAR 3.1
BAL	Balaban topological index		TSAR 3.1
WIE	Wiener topological index		TSAR 3.1
EST	sum of E-state indices		TSAR 3.1
SA	surface area	Å^2	TSAR 3.1
POL1	polarizability (tsar)	Å^3	TSAR 3.1
TE	total energy	eV	TSAR 3.1
HF	heat of formation	kcal/mol	TSAR 3.1
LUMO	energy of LUMO	eV	TSAR 3.1
HOMO	energy of HOMO	eV	TSAR 3.1
DIP	total dipole	debye	TSAR 3.1
MV2	molecular volume (volsurf)	Å^3	VOLSURF 2.0
S	molecular surface	Å^2	VOLSURF 2.0
R	molecular volume:surface ratio	Å	VOLSURF 2.0
G	molecular globularity		VOLSURF 2.0
W1−8	hydrophilic regions energy level 1−8	Å	VOLSURF 2.0
Iw1−8	integy moments 1−8	Å	VOLSURF 2.0
Cw1−8	capacity factors 1−8	Å	VOLSURF 2.0
Emin1−3	local interaction energy minima 1−3	kcal/mol	VOLSURF 2.0
D12,13,23	local interaction energy minimum distance 12,13,23	Å	VOLSURF 2.0
D1−8	hydrophobic regions at energy level 1−8	Å^3	VOLSURF 2.0
ID1−8	hydrophobic integy moments 1−8	Å	VOLSURF 2.0
HL1−2	hydrophilic−lipophilic balances 1 and 2		VOLSURF 2.0
A	amphiphilic moment	Å	VOLSURF 2.0
CP	critical packing parameter		VOLSURF 2.0
POL2	polarizability (volsurf)	Å^3	VOLSURF 2.0
CME	conformation minimum energy	kcal/mol	MOPAC 6.0
EA	electron affinity	eV	MOPAC 6.0
SE	steric energy	kcal/mol	MOPAC 6.0
BL	bond length of a C−X bond	Å	MOPAC 6.0
BO	bond order of a C−X bond		MOPAC 6.0
BS	bond strain of a C−X bond	kcal/mol	MOPAC 6.0
$Q_{x,c}$	partial charge on atom X and C	au	MOPAC 6.0
$HOMO_{x,c}$	HOMO density on atom X and C		MOPAC 6.0
$LUMO_{x,c}$	LUMO density on atom X and C		MOPAC 6.0
$EFD_{x,c}$	electrophilic frontier density on atom X and C		MOPAC 6.0
$NFD_{x,c}$	nucleophilic frontier density on atom X and C		MOPAC 6.0
$RFD_{x,c}$	radical frontier density on atom X and C		MOPAC 6.0
$ESD_{x,c}$	electrophilic superdelocalizability on atom X and C	eV^{-1}	MOPAC 6.0
$NSD_{x,c}$	nucleophilic superdelocalizability on atom X and C	eV^{-1}	MOPAC 6.0
$RSD_{x,c}$	radical superdelocalizability on atom X and C	eV^{-1}	MOPAC 6.0
logP2	octanol−water partition coefficient (sar)		SAR 3.0
Sw1	water solubility (based on logP)	mol/L	SAR 3.0
Sw2	water solubility (based on logP and mp)	mol/L	SAR 3.0

was initiated by adding 100 μL of the enzyme preparation. The reaction progress was followed by withdrawing 0.5 mL samples at 0, 20, and 30 min periods. The samples were mixed with 0.5 mL of methanol to terminate the reaction and directly applied to a gas chromatography apparatus equipped with a flame ionization detector (Hewlett-Packard 6890). The DB-FFAP capillary column [30 m $\times$ 0.25 mm $\times$ 0.25 μm (J&W Scientific)] was used for separation. Samples were injected by using a split technique (split ratio of 50:1). The temperature program was isothermal and was dependent on the character of the analyzed compound. K_m and k_{cat} values with their standard deviations were calculated by the method of least squares with relative weighting using LEONORA, version 1.0. The enzymatic reaction products were identified by comparison of retention times of identical standards, corresponding alcohols, or by mass spectrometry. For gas chromatography−mass spectrometry (GC−MS) analysis, the reaction mixture was extracted with 100 μL of chloroform and injected into a GC−MS system (Hewlett-Packard 6890) equipped with a DB-5MS capillary column. Split injection, isothermic analysis at 60 °C, and scan mode at *m*/*z* 25−200 were used for evaluation of the mass spectra of dehalogenation products.

Construction of Classical QSAR Models. The set of 28 descriptors used for the experimental design was complemented with 54 descriptors computed with VOLSURF, version 2.0 (Multivariate Infometric Analysis), which are potentially useful for modeling of enzyme−substrate association, 24 quantum mechanical descriptors computed with MOPAC, version 6.0 (*21*), for description of the dehalogenation, and three descriptors computed with SAR, version 3.0 (BioByte), for description of substrate desolvation (Table 1). Water and DRY probes with eight energy levels were used in the VOLSURF calculations, while the AM1 Hamiltonian and the PRECISE stopping criteria were used in MOPAC calculations. The models were developed for logarithmically transformed K_m constants by means of partial least-squares projection to latent structures (PLS), as implemented in SIMCA-P, version 10.0 (Ume-Tri). The data matrix was mean-centered and scaled to unit variance prior to PLS analysis.

Construction of COMBINE Models. The crystal structure of the LinB enzyme (PDB entry 1D07) was obtained from the Protein Data Bank. Polar hydrogen atoms were added using WHATIF version 5.0 (*22*). His272 was singly protonated on N_δ in accordance with its catalytic function. Nonpolar hydrogen atoms were added using AMBER version 5.0 (University of California, Berkeley, CA). The structures of the enzyme−substrate complexes were prepared using an automated docking procedure implemented in AUTODOCK version 3.0 (*23*). Grid maps were calculated for the atom types present in the substrates using 81 $\times$ 81 $\times$ 81 grid points and a grid spacing of 0.25 Å. A Lamarckian genetic algorithm was employed for docking with a population of 50 individuals, a maximum number of 1.5×10^6 energy evaluations, a maximum number of generations of 27 000, an elitism value of 1, a mutation rate of 0.02, and a crossover rate of 0.80. The local search was based on a pseudo-Solis and Wets algorithm (*24*) with a maximum of 300 iterations per local search. Fifty docking runs were performed for each enzyme−substrate complex. Calculated substrate orientations from each run were clustered with the clustering tolerance for the root-mean-square positional deviation set to 0.5 Å. Optimal orientations were selected by visual inspection of

enzyme−substrate structures, paying attention to the spatial position of atoms reacting during the dehalogenation reaction. The geometry of the selected enzyme−substrate complexes was optimized using AMBER version 5.0 and the Cornell et al. molecular mechanics force field (*25*). Before minimization, crystallographically resolved water molecules were added to the enzyme−substrate complexes. Water molecules making steric clashes with docked substrate molecules were deleted. One hundred steps of steepest descent were followed by conjugate gradient energy minimization until the root-mean-square value of the potential energy gradient was less than 0.1 kcal mol^{-1} $Å^{-1}$. The nonbonded cutoff was set to 10 Å, and a distance-dependent dielectric constant ($\epsilon = 4r_{ij}$) was employed. The enzyme−substrate interaction energy in the presence of the surrounding solvent together with the change in desolvation energies of the substrate and the enzyme upon binding was estimated. The approach for calculating the electrostatic contributions to the free energies of binding and the changes in enzyme and substrate solvation energies upon binding requires solving the linear form of the Poisson−Boltzmann equation and has already been described in detail (*14*). The statistical method PLS (*26*) was used for identification and ranking of interactions important for the differences in apparent dissociation constants among substrates. The matrix of *X* variables consisted of either 594 columns (van der Waals and electrostatic energy contributions for 296 amino acid residues plus two energy contributions from one catalytic water molecule) or 1753 columns (matrix above plus energy contributions from 1159 crystallographically resolved water molecules) and 25 rows (enzyme−substrate complexes). The dependent variable *y* was represented by 25 logarithmically transformed values of the apparent dissociation constants, K_m. The *X* variables with low-magnitude energies and variance were eliminated from the data matrix (cutoff of 10^{-7}). All PLS models were constructed using the statistical program SIMCA 8.0 (Umetrics). The quality of the models was described by the correlation coefficient (R^2) and by the cross-validated correlation coefficient (Q^2). R^2 is a descriptor of the quality of fit and takes values up to a maximum of 1, which corresponds to a perfect fit. A value higher than 0.5 is generally considered to be statistically significant. Q^2 provides an estimate of the predictive power of a model, with a value higher than 0.4 being generally considered statistically significant.

$$R^2 = 1 - \frac{\sum_i (y_{i\mathrm{calc}} - y_{i\mathrm{obs}})^2}{\sum_i (y_{i\mathrm{obs}} - y_{i\mathrm{mean}})^2}$$

$$Q^2 = 1 - \frac{\sum_i (y_{i\mathrm{pred}} - y_{i\mathrm{obs}})^2}{\sum_i (y_{i\mathrm{obs}} - y_{i\mathrm{mean}})^2}$$

RESULTS

Statistical Experimental Design and Kinetic Characterization. Haloalkane dehalogenase LinB shows an extremely broad substrate range (*1*), which makes selection of compounds for testing a difficult task. Proper selection of potential substrates is important because data homogeneity determines the robustness and validity of structure−activity models. Experimental statistical design is meant to optimize selection of substrates for QSAR modeling by selecting the substrates that represent a broad variety of chemical structures yet are not too different to prevent construction of a mathematical model quantitatively describing relationships between the structure and biological activity (or substrate specificity, as in the present case). PCA was used for comparison of the halogenated compounds in terms of their physicochemical and structural properties. PCA applied on a homogeneous data set of 106 halogenated compounds and 28 molecular descriptors resulted in four significant principal components each explaining 37, 16, 14, and 12% of data variability, respectively. Score plots show clustering of the compounds according to their properties (Figure 1). The first principal component separates compounds by their size (most significant descriptors being MV, SA, MR, TE, MW, MM, M2s, and M3s), whereas the second component does so using electronic and physicochemical properties (LUMO, *D*, and *n*). The third principal component captures the shape and mass of the molecules (M2l, M1l, M3l, and M1s), while the fourth component is made of hydrophobicity and electronic descriptors (logP, HF, EST, and DIP). The principal components derived were used as the design variables in 2^4 fractional design assuming two levels (+ and −) for four different variables (PC1−PC4). The purpose is to select the best representatives for the entire data set. Classification of the compounds according to design variables is presented in Table 2. Overall, 50 compounds were selected for experimental testing to cover all possible classes and enriched by five additional compounds: 1-chloroheptane (7), 1-chlorooctane (8), 1-iodohexane (31), 2-bromobutyrate (106), and 1,3-dibromopropene (238). Steady-state kinetic constants were determined for 25 compounds. Kinetic parameters could not be determined for 18 compounds which did not serve as substrates for LinB, seven compounds that showed high abiotic hydrolysis, and five compounds that exhibited nonlinear kinetics (Table 3).

Construction of QSAR Models. The outputs from the experimental statistical design and PCA enable qualitative structure−activity relationships to be established. The distribution of compounds that are dehalogenated by LinB in comparison to the compounds that resist dehalogenation in the scores plots enables identification of the structural properties important for hydrolytic dehalogenation by this enzyme. Many compounds that undergo dehalogenation by LinB are concentrated on the right side of Figure 1A, indicating that larger molecules are preferred substrates for LinB. Horizontal separation in the same plot also seems to be relevant for activity. Dibrominated substrates with a low energy of the lowest unoccupied molecular orbital (LUMO) positioned at the very top of the figure are dehalogenated by LinB with high catalytic rates. In Figure 1B, the compounds that serve as substrates for LinB are positioned left-most and right-most in the figure, and they are separated by compounds that resist dehalogenation positioned in the middle of the plot. The dependence of the dehalogenation on the shape of the molecules is apparently more complex than the relation to size. The smaller groups of structurally similar compounds are easier to examine (Table 2). Some

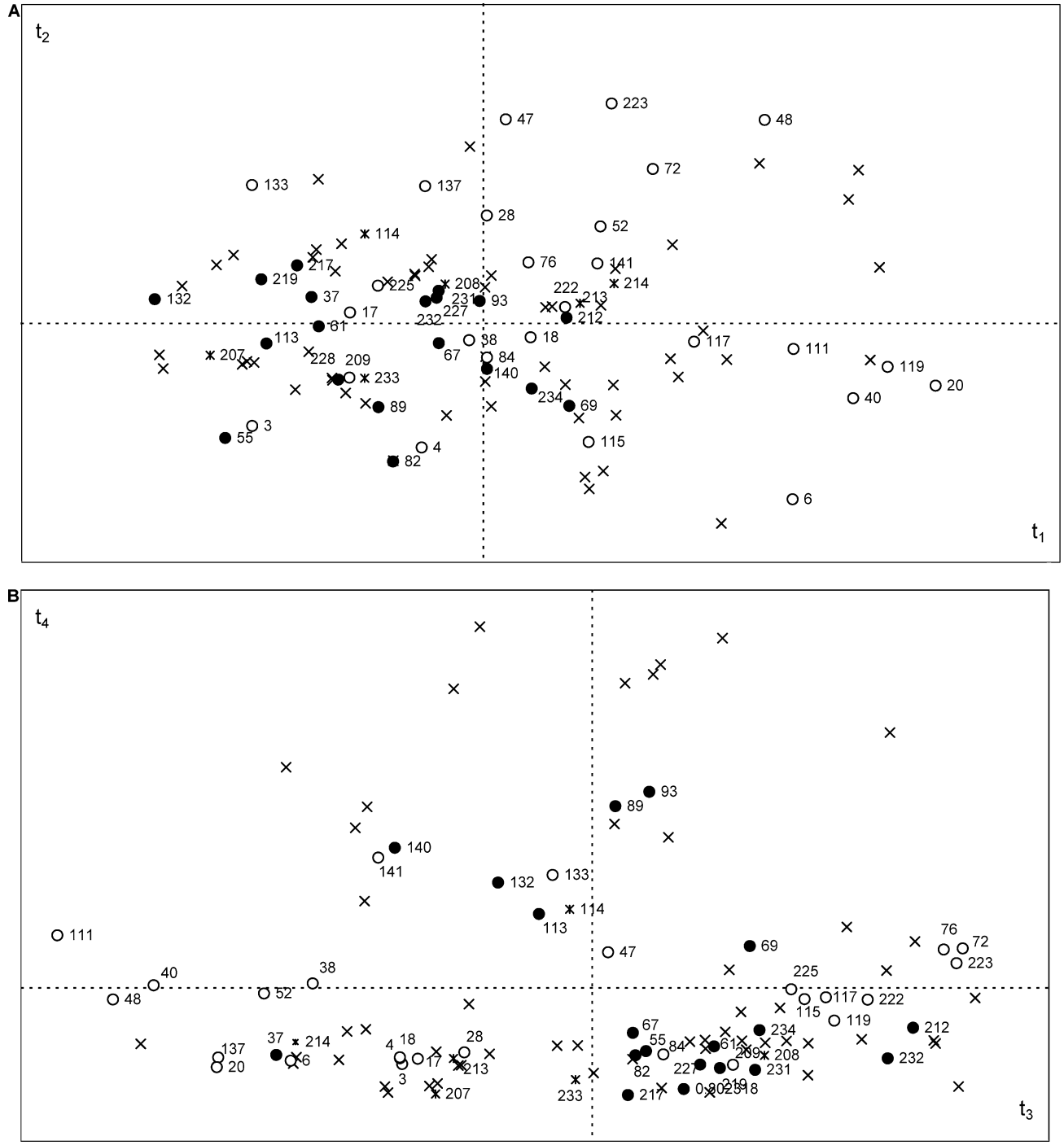

FIGURE 1: Clustering of halogenated compounds based on their molecular structures and physicochemical properties in the scores plot of t_1 vs t_2 (A) and t_3 vs t_4 (B). Compounds dehalogenated by LinB are represented with white circles, compounds not acting as substrates with black circles, compounds with fast hydrolysis with asterisks, and compounds not selected for testing with crosses. The numbering of compounds is presented in Table 4. Dashed lines indicate the position of origin.

of the classes in the table contain primarily compounds that serve as substrates for LinB (classes 1, 5, 10, 11, and 16), while others contain inhibitors (classes 3 and 9). Both classes of inhibitors contain saturated (class 3) or unsaturated (class 9) compounds with substituents at the β-position.

Quantitative analysis of K_m constants was conducted to determine the structural parameters of substrates that are important for binding. It was assumed that the differences in K_m values capture, to a large extent, differences in binding affinities. This concept was proven in the similar analyses conducted for the haloalkane dehalogenase DhlA (*14*, *16*). Log K_m values were initially correlated with one descriptor at a time using linear regression analysis. Significant correlation ($R^2 \geq 0.5$, $N = 25$) was observed for parameters related to the hydrophobicity and size of the molecules: octanol−water partition coefficient (logP; $R^2 = 0.60$), polarizability (POL; 0.58), molar refractivity (MR; 0.55), Randic index (RAN; 0.55), molecular volume:surface ratio (R; 0.53), and molecular volume (MV; 0.50). Two compounds, bis(2-chloroethyl)ether (111) and chlorocyclohexane (115), showed binding affinities significantly higher than those predicted from their hydrophobicity and size. Outlier

Table 2: Classification of the Substrates According to 2^4 Factorial Design Variables

class design compds[a]	1 (−/−/−/−)	2 (−/−/−/+)	3 (−/−/±)	4 (−/±/−)	5 (±/−/−)	6 (−/−/+/+)	7 (−/±/+)	8 (±/−/+)	9 (−/+/±)	10 (±/±)	11 (+/±/−)	12 (+/+/±)	13 (+/±/+)	14 (±/+/+)	15 (−/+/+/+)	16 (+/+/+/+)
	3[b]	38[b]	55[c]	16	5	89[c]	92	40[b]	145	57	28[b]	64	141[b]	69[c]	93[c]	47[b]
	4[b]	88	56	17[b]	6[b]	131	114	90	147	58	29	74		70	95	72[b]
	207	113[c]	61[c]	27	18[b]	134	132[c]	111[b]	208	62	30	116		86	158	75
	233	149	67[c]	37[c]	19	157	133[b]	140[c]	217[c]	63	48[b]	148		96		76[b]
			82[c]	137[b]	20[b]	209[b]		142	219[c]	71	52[b]	212[c]				80
			138	202	39			143	225[b]	83	213	222[b]				155
			144	210	112				227[c]	84[b]	214					223[b]
			146	211					231[c]	115[b]	229					
			221	218					232[c]	117[b]	236					
			228	220						118						
			230	239c						119[b]						
			237c	239t						234[c]						
			237t													

[a] The numbering of tested compounds is in Table 3. [b] Activity observed. [c] No activity observed at 100 mM.

behavior was especially apparent for chlorocyclohexane (115), which has size and hydrophobicity similar to those of its brominated analogue bromocyclohexane (116), but a 1 order of magnitude higher K_m (Table 3). Kinetic measurements were repeated with bis(2-chloroethyl)ether (111), chlorocyclohexane (115), and bromocyclohexane (116), but the determined values did not differ significantly from the original data (not shown).

The PLS method was then applied to investigate whether a combination of structural parameters could better explain variability in K_m. Model M1, based on the octanol−water partition coefficient (logP), surface area (SA), polarizability (POL), dipole moment (DIP), and molecular volume:surface ratio (R), explained 72% quantitative variance (68% cross-validated) in dissociation constants (Table 4). Comparison of the predicted versus observed K_m values revealed poor prediction for chlorocyclohexane and nonlinear dependence of K_m on descriptors used in the model (Figure 2, inset). All attempts to explain the low affinity of binding of chlorocyclohexane by additional descriptors were unsuccessful. Chlorocyclohexane resembles the natural substrate of LinB, 1,3,4,6-tetrachloro-1,4-cyclohexadiene, which must bind in the active site in such a way that a chlorine substituent on one side of the ring is pointing toward the nucleophile Asp108, whereas another chlorine on the other side is pointing in the opposite direction. Chlorocyclohexane may bind in the active site in two different orientations, of which only one is productive and results in chemical conversion of the substrate. Another explanation for the outlier behavior could be that dehalogenation of chlorocyclohexane proceeds by a somewhat different reaction mechanism, with a different rate-limiting step compared to other substrates. Such complex phenomena cannot be properly described by molecular descriptors, and therefore, chlorocyclohexane was excluded from further analysis. The recalculated model M2 showed improved statistical parameters and explained 81% quantitative variance (74% cross-validated) in dissociation constants (Table 4). Two quadratic terms added for two descriptors already present in the model were sufficient to handle model nonlinearity (Figure 2). The final model M3 for dissociation constants can be expressed by the following multiple regression equation: $\log K_m = -0.21641 \times \mathrm{logP} - 0.07611 \times \mathrm{SA} + 0.00026 \times \mathrm{SA}^2 - 0.03836 \times \mathrm{POL} - 1.19829 \times \mathrm{DIP} + 0.33872 \times \mathrm{DIP}^2 - 0.58228\,R + 6.37465$. The validity and robustness of the model were tested by external validation. The predictive ability of the model was estimated by calculating the standard deviation of error in external predictions: values of 0.30 and 0.25 were calculated for the models based on odd-numbered and even-numbered compounds, respectively. These values compare favorably with the standard deviation of error of internal predictions of the final PLS model (0.25), thus providing support for the validity of the developed model.

Construction of COMBINE Models. The substrate set structurally consisted of monohalogenated alkanes up to a chain length of eight carbon atoms, dihalogenated propanes, ethane and pentane, monohalogenated cyclohexanes, monohalogenated ether, and dihalogenated propene. The automated docking procedure provided positionally suitable orientations for 19 substrates: 1-chlorobutane (4), 1-chlorohexane (6), 1-chloroheptane (7), 1-chlorooctane (8), 1-bromopropane (17), 1-bromobutane (18), 1-bromohexane (20), 1,3-dichlo-

Table 3: Steady-state Kinetic Constants of Haloalkane Dehalogenase LinB

no.	substrate	K_m (mM)	SE (mM)	k_{cat} (s^{-1})	SE (s^{-1})	k_{cat}/K_m (mM^{-1} s^{-1})	product
3	1-chloropropane	1.100	0.4440	1.124	0.2342	1.02	1-propanol[e]
4	1-chlorobutane	0.129	0.0070	1.020	0.0257	7.94	1-butanol[e]
6	1-chlorohexane	0.005[a]	0.0025	1.033	0.0402	206.51	1-hexanol[e]
7	1-chloroheptane	0.015	0.0014	3.345	0.3050	222.73	1-heptanol[e]
8	1-chlorooctane	0.021	0.0067	2.750	0.1927	133.17	1-octanol[e]
17	1-bromopropane	0.231	0.0361	5.519	0.3652	23.88	1-propanol[e]
18	1-bromobutane	0.043[a]	0.0053	3.414	0.1370	78.95	1-butanol[e]
20	1-bromohexane	0.010[a]	0.0050	1.050	0.0292	105.05	1-hexanol[e]
28	1-iodopropane	0.110	0.0142	3.935	0.2113	35.79	1-propanol[e]
31	1-iodohexane	0.010[a]	0.0050	2.327	0.0790	232.73	1-hexanol[e]
37	1,2-dichloroethane	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
38	1,3-dichloropropane	0.160	0.0380	0.950	0.0709	5.94	3-chloro-1-propanol[e]
40	1,5-dichloropentane	0.019	0.0069	2.450	0.2580	130.29	5-chloro-1-pentanol[e]
47	1,2-dibromoethane	1.900	0.2081	6.120	0.5332	3.22	2-bromo-1-ethanol[e]
48	1,3-dibromopropane	0.040	0.0056	6.600	0.4121	165.00	3-bromo-1-propanol[e]
52	1-bromo-3-chloropropane	0.210	0.0300	6.886	0.3229	32.79	3-chloro-1-propanol[e]
55	2-chloropropane	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
61	2-bromopropane	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
67	1,2-dichloropropane	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
69	1,2-dichlorobutane	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
72	1,2-dibromopropane	0.140	0.0100	0.843	0.0539	6.02	1-bromo-2-propanol[f]
76	2-bromo-1-chloropropane	0.551	0.0510	1.374	0.0337	2.49	1-chloropropane-2-ol[f]
82	1-chloro-2-methylpropane	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
84	1-bromo-2-methylpropane	0.050	0.0060	1.599	0.0772	31.98	2-methylpropanol[e]
89	3-chloropropanol	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
93	3-bromopropanol	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
106	2-bromobutyrate	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]
111	bis(2-chloroethyl)ether	0.870	0.0334	0.363	0.0551	0.42	2-(2-chloroethoxy)ethanol[f]
113	epichlorohydrine	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
114	epibromohydrine	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]
115	chlorocyclohexane	0.252	0.0450	0.144	0.0250	0.57	cyclohexanol[e]
117	bromocyclohexane	0.021	0.0021	1.335	0.0340	63.29	cyclohexanol[e]
119	(1-bromomethyl)cyclohexane	ND[d]	ND[d]	ND[d]	ND[d]	ND[d]	ND[d]
132	2-chloroacetonitrile	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
133	2-bromoacetonitrile	ND[d]	ND[d]	ND[d]	ND[d]	ND[d]	ND[d]
137	1-bromo-2-chloroethane	ND[d]	ND[d]	ND[d]	ND[d]	ND[d]	ND[d]
140	4-chlorobutyronitrile	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
141	4-bromobutyronitrile	0.207	0.0110	3.701	0.0634	17.90	4-hydroxybutyronitrile[f]
207	3-chloropropene	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]
208	3-bromo-2-methylpropene	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]
209	3-chloro-2-methylpropene	0.340	0.0450	3.080	0.2049	9.06	2-methylpropenol[f]
212	3,4-dichloro-1-butene	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
213	1,4-dichloro-2-butene (trans)	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]
214	1,4-dichloro-2-butene (cis)	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]
217	1-bromopropene	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
219	2-bromopropene	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
222	3-chloro-2-(chloromethyl)-1-propene	0.079	0.0187	8.165	0.0817	103.46	3-chloromethyl-2-propen-1-ol[f]
223	2,3-dibromopropene	ND[d]	ND[d]	ND[d]	ND[d]	ND[d]	ND[d]
225	2,3-dichloropropene	0.542	0.0479	1.632	0.0646	3.01	2-chloropropen-3-ol[f]
227	1-bromo-2-methylpropene	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
231	2-bromo-2-butene (cis)	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
232	2-bromo-2-butene (trans)	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
233	1-chloro-2-butene (trans)	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]	ND[c]
234	1-chloro-3-methyl-2-butene	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]	NA[b]
238	1,3-dibromopropene (cis + trans)	ND[d]	ND[d]	ND[d]	ND[d]	ND[d]	ND[d]

[a] Values not exact due to analytical uncertainties at low concentrations. [b] No activity observed at 100 mM. [c] Not determined due to fast abiotic hydrolysis. [d] Not determined due to nonlinear kinetics. [e] Determined by GC analysis. [f] Determined by GC−MS analysis.

ropropane (38), 1,5-dichloropentane (40), 1,2-dibromoethane (47), 1,3-dibromopropane (48), 1-bromo-3-chloropropane (52), 1,2-dibromopropane (72), 1-bromo-2-methylpropane (84), bis(2-chloroethyl)ether (111), 4-bromobutyronitrile (141), 3-chloro-2-methylpropene (209), 3-chloro-2-(chloromethyl)-1-propene (222), and 2,3-dichloropropene (225). Extended docking (128 runs) had to be used to obtain the reactive conformation of 1-iodohexane (31), chlorocyclohexane (115), and bromocyclohexane (117). No suitable orientations, even in extended docking runs, were found for three substrates: 1-chloropropane (3), 1-iodopropane (28), and 2-bromo-1-chloropropane (76). The orientations for 1-chloropropane and 1-iodopropane were therefore prepared from the selected orientation of 1-bromopropane (18) by manual exchange of halogen atoms and energy minimization. The complex with 2-bromo-1-chloropropane was prepared in the same manner from the selected complex with 1,2-dibromopropane (72). The substrate orientations obtained from automated docking spatially formed one cluster occupying a position that was properly suited for the dehalogenation reaction. During energy minimization, some substrates drifted away from their original positions and formed

Table 4: QSAR and COMBINE Models

model	approach	scaling	centering	freezing	objects	variables	A	R^2	Q^2
M1	QSAR	unit variance	+	NA[a]	25	5	1	0.72	0.69
M2	QSAR	unit variance	+	NA[a]	24	5	2	0.81	0.74
M3	QSAR	unit variance	+	NA[a]	24	5[b]	3	0.94	0.90
M4	COMBINE	none	−	−	23	595	2	0.85	0.79
M5	COMBINE	none	+	−	23	595	1	0.49	0.34
M6	COMBINE	unit variance	−	−	23	595	2	0.86	0.82
M7	COMBINE	unit variance	+	−	23	595	1	0.63	0.54
M8	COMBINE	pareto	−	−	23	595	2	0.86	0.82
M9	COMBINE	pareto	+	−	23	595	1	0.56	0.42
M10	COMBINE	none	−	+	23	595	2	0.86	0.81
M11	COMBINE	none	+	+	23	595	1	0.54	0.42
M12	COMBINE	unit variance	−	+	23	595	2	0.85	0.81
M13	COMBINE	unit variance	+	+	23	595	1	0.59	0.49
M14	COMBINE	pareto	−	+	23	595	2	0.87	0.82
M15	COMBINE	pareto	+	+	23	595	1	0.57	0.46
M16	COMBINE	none	−	±[c]	23	595	3	0.95	0.91

[a] Not applicable. [b] Cross-terms were used for descriptors SA and DIP. [c] Molecules 7 and 8 frozen.

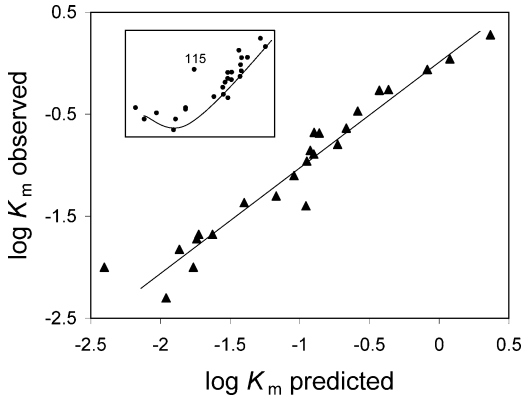

FIGURE 2: Quality of predictions for dissociation constants from the classical QSAR model visualized in the predicted vs observed plot. The statistical parameters of the final classical QSAR model (model M3) are as follows: $R^2 = 0.94$, $Q^2 = 0.90$, $N = 24$, and $A = 3$. The inset presents the plot obtained for the initial model (model M1) showing the underestimated prediction for chlorocyclohexane (115) and nonlinear relationships for long-chain substrates ($R^2 = 0.72$, $Q^2 = 0.68$, $N = 25$, and $A = 1$). The axes of the inset (log K_m predicted vs log K_m observed) are not shown for clarity.

a separate cluster that was less well suited for the nucleophilic attack. An alternative approach for complex refinement was therefore applied by which the substrate molecules were kept in their docked orientations and only the protein atoms were allowed to adapt (see the next paragraph).

Initial COMBINE models were built for the set of 25 substrates. Each row in the *X* matrix described the interaction energies of each substrate with the amino acid residues of LinB and the solvent molecules on a per residue or solvent molecule basis. The *y* column represented the logarithm of the apparent dissociation constant for each substrate. The effect of several conditions on the statistical quality of models was assessed: (i) data pretreatment, (ii) conformational freezing of the substrate, (iii) presence of solvent molecules, and (iv) object selection.

(i) Several different types of data pretreatment were applied to the *X* matrix. The statistical parameters of COMBINE models based on the data with different pretreatment schemes are summarized in Table 4. All models employing centering (M5, M7, M9, M11, M13, and M15) performed poorly. The centering unifies the distribution of individual variables around zero. Such a procedure is apparently not suitable for interaction energies which carry physically meaningful information about the protein−ligand complex. The statistical quality of all noncentered models was comparable. It was therefore decided to use the data without any pretreatment for construction of the final model.

(ii) Comparison of models with (M10−M15) and without freezing (M4−M9) of the substrate molecules during energy refinement of the enzyme−substrate complexes revealed that the models without freezing possessed somewhat better statistical parameters. Chemometric analysis of both types of models, together with careful inspection of the structures of enzyme−substrate complexes, led us to propose a combined model (M16) based on completely relaxed structures for all the complexes except for those of the two longest substrates in the set: 1-chloroheptane (7) and 1-chlorooctane (8). These molecules make close contacts with the tunnel residues which results in their drift from the reactive position during minimization.

(iii) The influence of solvent on the COMBINE model was initially tested by inclusion of 1159 crystallographic water molecules in the data matrix; i.e., the interaction energies of the water molecules with the substrate acted as the objects in the PLS model, and later implicitly by adding desolvation energies. Water molecules significantly increased the complexity of the PLS model but did not improve its statistical quality (data not shown). Therefore, all water molecules were excluded from further analysis except for the catalytic water molecule bound near the catalytic triad. Since addition of desolvation energies of both the substrate and the enzyme to the *X* matrix did not improve the models either, these variables were not incorporated into the final model.

(iv) Two outlying objects, bis(2-chloroethyl)ether and chlorocyclohexane, were systematically identified in the PLS models. These two molecules have the lowest k_{cat}/K_m with LinB of all 25 substrates analyzed in this study and probably differ from the rest in their binding mode and/or the kinetics of their mechanism of dehalogenation. Note that chlorocyclohexane was also identified as an outlier in the classical QSAR study relating K_m values to the physicochemical properties of the substrate molecules.

The optimized parameters were used for the construction of the final model M16: the *X* matrix was uncentered and unscaled; the *y* vector was logarithmically transformed;

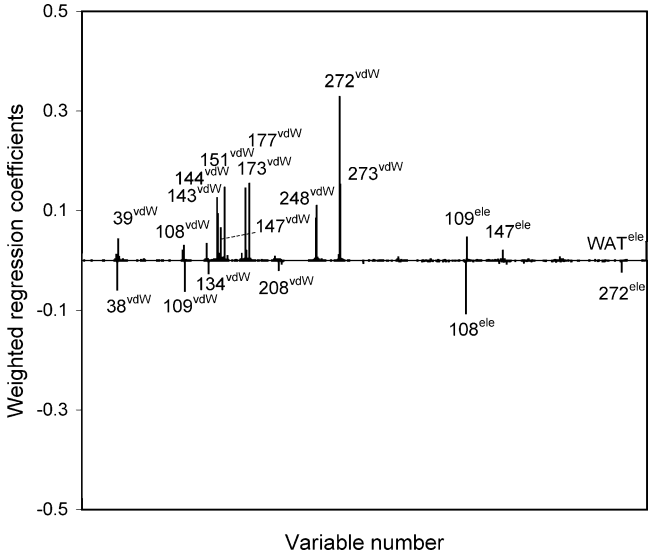

FIGURE 3: Relative importance of energetic contributions for substrate specificity quantified by a plot of weighted regression coefficients from the COMBINE model. Selected variables (energy contributions) are numbered according to the LinB sequence. The statistical parameters of the final COMBINE model are as follows: $R^2 = 0.92$, $Q^2 = 0.89$, $N = 23$, and $A = 2$.

enzyme−substrate complexes were completely relaxed except for 1-chloroheptane and 1-chlorooctane; solvent molecules were not modeled either implicitly or explicitly; two outliers, bis(2-chloroethyl)ether and chlorocyclohexane, were excluded from the analysis. The final model explained 95% quantitative variance (91% cross-validated) in dissociation constants closely resembling the quality of the final QSAR model (Table 4). Significant interactions for substrate specificity of LinB were identified by calculating weighted regression coefficients (Figure 3). Twenty *x* variables (interaction energies) have been assigned as the most important contributions based on the coefficient values: 15 of them correspond to van der Waals terms, while five correspond to electrostatic terms. These coefficients also provide information about the direction of the effect: 14 have a positive sign versus six with a negative value. The chemometric meaning of the signs is that positive coefficients relate to favorable contributions, whereas negative coefficients assign unfavorable contributions. The scores plot (Figure 4A) displays the distribution of objects (substrates) according to the first and second principal components (PC) of the model. The first PC separates compounds horizontally into three groups: (i) 1-chlorohexane (6), 1-bromohexane (20), 1-iodohexane (31), and 1,5-dichloropentane (40), (ii) 1-chloropropane (3), 1-bromopropane (17), 1-iodopropane (28), and 2,3-dichloropropene (225), and (iii) substrates not included in any of two former groups. The second PC separates objects vertically, also into three groups: (i) 1-chloroheptane (7) and 1-chlorooctane (8), (ii) 1-chloropropane (3), 1-bromopropane (17), 1-iodopropane (28), and 2,3-dichloropropene (225), and (iii) substrates left over. The loadings plot (Figure 4B) shows the distribution of variables according to the extent of their contribution to the individual PCs. The most significant contribution to the first PC is provided mainly by van der Waals interaction energies 38^{vdw}, 39^{vdw}, 108^{vdw}, 109^{vdw}, 143^{vdw}, 151^{vdw}, 169^{vdw}, 173^{vdw}, 177^{vdw}, 207^{vdw}, 208^{vdw}, 211^{vdw}, 248^{vdw}, 272^{vdw}, and 273^{vdw} and four electrostatic interaction energies (108^{ele}, 109^{ele}, 272^{ele}, and WAT^{ele}). The most important contribution to the second PC is provided by van der Waals interaction energies 38^{vdw}, 109^{vdw}, 143^{vdw}, 173^{vdw}, 177^{vdw}, 208^{vdw}, 272^{vdw}, and 273^{vdw} and three electrostatic interaction energies (108^{ele}, 147^{ele}, and 272^{ele}).

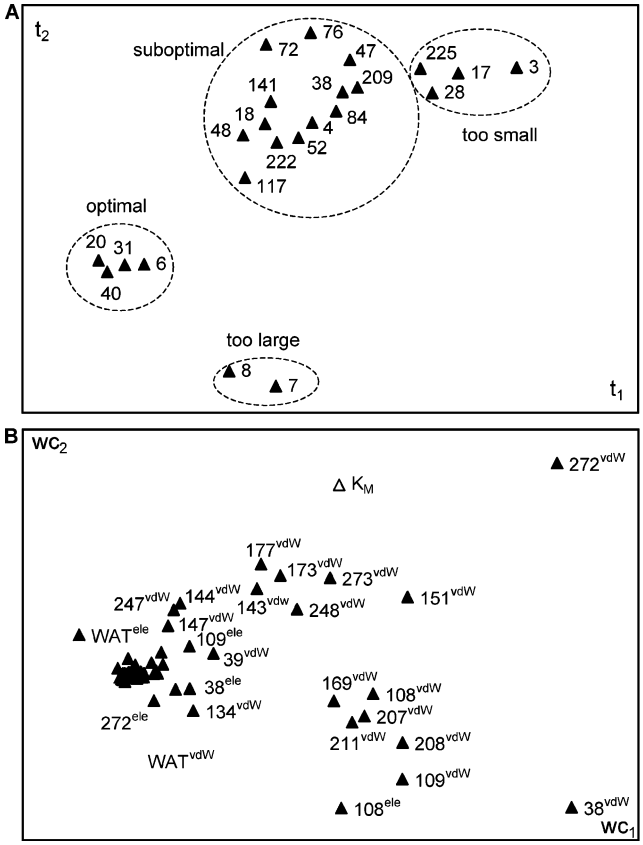

FIGURE 4: Clustering of halogenated compounds (A) based on their intermolecular interactions with amino acid residues (B) in the scores plot of t_1 vs t_2 (A) and wc_1 vs wc_2 (B). The numbering of compounds is presented in Table 4. Variables (interaction energies) are numbered according to the LinB sequence.

The variable importance in projection (VIP) parameter quantifies the overall importance of each variable in the model and as such is well suited for identification of the amino acid residues representing the best candidates for site-directed mutagenesis. The 24 energy contributions with the highest VIP are shown on Figure 5 and are (i) first-shell residues (Asn38, Asp108, Trp109, Glu132, Ile134, Phe143, Phe151, Phe169, Val173, Trp207, Pro208, Ile211, Leu248, and His272), (ii) tunnel residues (Pro144, Asp147, Leu177, and Ala247), and (iii) second-shell residues (Pro39 and Phe273).

DISCUSSION

Enzyme substrate specificity can be viewed as the range of small organic ligands that serve as substrates for a given enzyme. The substrate specificity of an enzyme is said to be narrow when only a few different substrate molecules are converted by it, whereas it is defined as broad if a large variety of compounds serve as substrates. A substrate molecule must bind to the enzyme active site and be converted to a product, leaving the enzyme structure essentially intact. Enzyme substrate specificity is therefore a function of the structure of both the substrates and the protein. Relationships between the molecular structure of

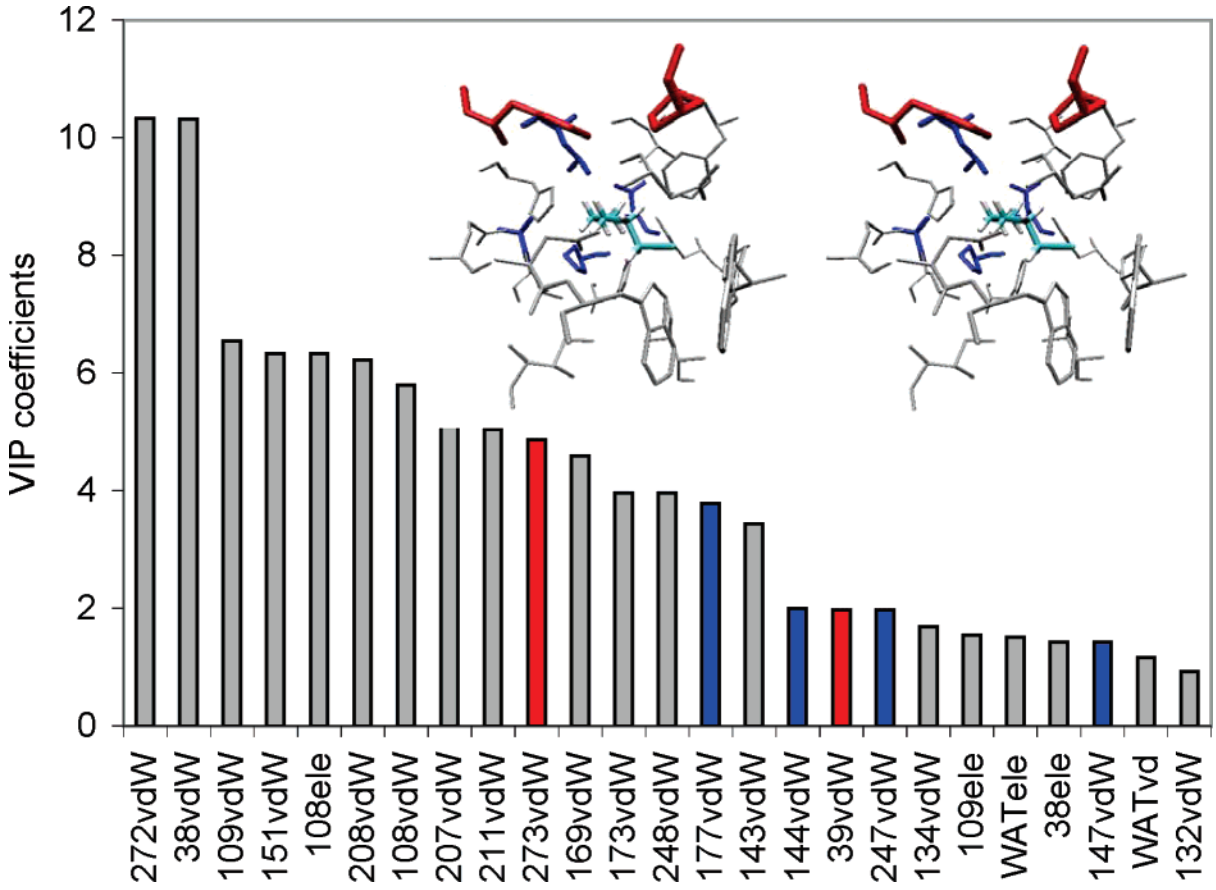

FIGURE 5: Key residues determining the substrate specificity of haloalkane dehalogenase LinB deduced from the VIP plot. Selected variables (energy contributions) are numbered according to the LinB sequence. The interaction energy contributions shown in red correspond to the second-shell residues (Pro39 and Phe273); interaction energy contributions shown in blue correspond to the tunnel residues (Pro144, Asp147, Leu177, and Ala247), and all other interaction energy contributions correspond to the first-shell residues. The inset presents a stereoview of the same residues in the LinB crystal structure.

substrates and enzymes should therefore be studied in parallel to provide a detailed understanding of this complex phenomenon.

Analysis of Substrate Specificity Using the Classical QSAR. A qualitative analysis was first conducted to distinguish compounds which serve as substrates for LinB from those which cannot be dehalogenated by this enzyme. The size and shape of the molecules, energy of the lowest unoccupied molecular orbital and the substitution pattern, were identified as important molecular properties for dehalogenation by this chemometric method. The size and shape of the molecules are important for binding of substrates in the enzyme active site. Molecules of improper size cannot bind efficiently or at all in the active site. A detailed quantitative analysis of the relationship between the size of the molecules and their binding affinities is described below. The energy of the lowest unoccupied molecular orbital is important for the reaction step. The compounds with high energy do not easily undergo nucleophilic attack by the enzyme's catalytic aspartic acid. The substitution pattern can play a role both in binding and in catalysis. Those compounds with the leaving halogen atom in the β-position or adjacent to a bulky substituent do not easily undergo S_N2 dehalogenation for steric reasons.

Quantitative analysis of dissociation constants using the classical QSAR approach revealed that binding affinity correlates with hydrophobicity, molecular surface, dipole moment, and molecular volume:surface ratio of tested substrates. Hydrophobic, polarizable molecules with large surface areas bind with high affinity. Small dipole moments and large volume:surface ratios lower the binding affinity of selected substrates. The binding affinity increases with an increase in molecular size until it reaches an optimum beyond which the binding affinity drops with a further increase in size. This nonlinear relationship was mathematically described by addition of quadratic terms for surface area, and dipole moment variables in the QSAR model and structurally correspond to the anatomy of the binding pocket of LinB which is buried in the protein interior (*8*). Docking experiments conducted with LinB and substrates of varying chain length confirmed this proposal. We expect that such nonlinear relationships will be common for broad-range specificity enzymes with buried active sites accepting ligands with varying length, such as lipases (*27*, *28*).

Analysis of Substrate Specificity Using COMBINE. Two outliers were detected during the initial modeling of 25 substrates of the haloalkane dehalogenase LinB using the COMBINE approach: bis(2-chloroethyl)ether and chlorocyclohexane. They show the lowest k_{cat}/K_m, and the latter compound was also detected as an outlier in the classical QSAR analysis, despite the fact that predictions in classical QSAR and COMBINE models are based on radically different descriptors. We propose that these two molecules may bind to the active site in different binding modes and/or be converted by a kinetically different dehalogenation mechanism compared to the rest of the substrates. This

proposal is supported by comparison of transient kinetics data measured with two analogous substrates, chlorocyclohexane and bromocyclohexane, suggesting that the hydrolysis of the alkyl−enzyme intermediate formed by dehalogenation of chlorocyclohexane is surprisingly 33 times slower than that of bromocyclohexane (*29*).

The COMBINE model primarily described differences in binding affinities caused by variability in chain length and complementarity with the active site. The mode of binding of various substrates to the active site of the LinB enzyme is highly similar, positioning the leaving halogen atom at the intersection of two halogen-stabilizing residues, Asn38 and Trp109, and the nucleophilic carbon atom near the attacking oxygen of Asp108. Binding affinity is seen to increase with increasing substrate size up to a chain length of six carbon atoms and then decreases. The optimal length corresponds to a group of long-chain substrates (1-chlorohexane, 1-bromohexane, 1-iodohexane, and 1,5-dichloropentane). Most unfavorable is the group of monosubstituted propanes (1-chloropropane, 1-bromopropane, 1-iodopropane, and 2,3-dichloropropene), and the suboptimal length corresponds to the rest of the substrates. The two longest substrates in the set, 1-chloroheptane and 1-chlorooctane, showed drift from the reactive position during the minimization procedure due to their close contacts with the tunnel residues. Complementarity with the active site is dominated by the energy contributions from Asn38 and His272. These two amino acid residues are located on opposite sides of the active site and directly interact with the substrate molecules bound in the Michaelis complex. Asn38 forms the bottom of the active site and together with Trp109 stabilizes a halogen atom by a hydrogen bond (*8*, *30*), while His272 is the base of the catalytic triad (*31*) and forms the first point of contact for the substrates entering the active site pocket via the entrance tunnel. His272 exhibited extremely strong van der Waals interaction with all studied substrates. The importance of this residue for binding of small ligands near the opening of the entrance tunnel was noted earlier in crystallographic studies (*32*, *33*). Distinction of hexanes from the excessively long 1-chloroheptane and 1-chlorooctane was achieved by freezing these molecules in the reactive position during complex refinement. Using this methodology, we could capture nonlinear relationships between the chain length of the substrate molecules and their binding affinities. Separation of substrate molecules primarily according to their size is due to contributions from His272, Leu177, Val173, and Phe273. These residues are located on the side of the active site opposite to the halide-binding pocket, i.e., in the direction of the tunnel, and make more favorable van der Waals interactions with long substrates than with short ones. The opposite trend, i.e., repulsion with long substrates, holds for the electrostatic interaction energies of Asn38 and Asp108 as well as for the van der Waals interaction energies of Trp109, Pro208, and Ile211. It is conceivable to assume that placing the long substrates in the active site will bring their hydrocarbon chains close to the wall of the tunnel, thus maximizing van der Waals interactions. On the opposite side of the active site, the short substrates could freely occupy the best positions near the amino acid residues located at the bottom of the active site. Positional differences were observed for the halogen-stabilizing residue Asn38 in different enzyme−substrate complexes. This amino acid displays high flexibility of both side chain and main chain atoms, thus allowing good accommodation of the active site to different substrates.

Identification of Specificity-Determining Second-Shell and Tunnel Residues. The tunnel and the second-shell residues (Pro39, Pro144, Asp147, Leu177, Ala247, and Phe273) are the natural targets for substitutions since their replacement will not lead to loss of functionality by disruption of the active site architecture. The relevance of such a proposal has already been proven by mutants constructed previously using directed evolution and side-directed mutagenesis techniques. The equivalents of tunnel residue L177 and second-shell residue Phe273 and in DhaA (Cys176 and Tyr273, respectively) were identified as hot spots for specificity of this enzyme in the directed evolution toward dehalogenation of 1,2,3-trichloropropane (*34*). The importance of L177 for specificity has been demonstrated by independent directed evolution (*35*), cumulative mutagenesis (*36*), and saturated mutagenesis (*37*) experiments. Highly rigid Pro39 is adjacent to Asn38, which is functionally one of the most important residues of LinB. Asn38 is involved in (i) halogen binding, (ii) transition-state and product stabilization, and (iii) coordination of the catalytic water molecule.

Comparison of Substrate Specificities for DhlA and LinB. The COMBINE models constructed for DhlA (*14*) and LinB (this study) were compared. In both models, only a limited number of protein residues (6−8%) contributed significantly to the explanation of variability in K_m. In addition, van der Waals interaction energies dominated over electrostatic interaction energies. Significant contributions provided by specific amino acid residues correspond well with the composition of the enzyme active site. Different halide-stabilizing residues (Trp125 and Trp175 in DhlA and Asn38 and Trp109 in LinB) are known to be employed in substrate binding in different dehalogenases (*30*), and these were correctly identified by the models. It is interesting to note the difference in contributions provided by the catalytic base located at equivalent positions in both proteins at the opening of the entrance tunnel. His289 in DhlA is significantly less important than His272 in LinB which relates to the different orientation of the active site pocket. The pocket of DhlA is approximately orthogonal to the entrance tunnel, while in LinB, it is in line with it. Since the catalytic base forms the tunnel opening, it makes a direct contact with the substrates bound to the active site in LinB, but not in DhlA. These differences reflect molecular adaptation of haloalkane dehalogenases to their specific roles in their bacterial hosts.

ACKNOWLEDGMENT

Dr. Gabriele Cruciani and Dr. Manuel Pastor (University of Perugia, Perugia, Italy) are kindly acknowledged for providing us with VOLSURF.

REFERENCES

1. Damborsky, J., Rorije, E., Jesenska, A., Nagata, Y., Klopman, G., and Peijnenburg, W. J. G. M. (2001) Structure-specificity relationships for haloalkane dehalogenases, *Environ. Toxicol. Chem. 20*, 2681−2689.
2. Stucki, G., and Thuer, M. (1995) Experiences of a large-scale application of 1,2-dichloroethane degrading microorganisms for groundwater treatment, *Environ. Sci. Technol. 29*, 2339−2345.

3. Swanson, P. E. (1999) Dehalogenases applied to industrial-scale biocatalysis, *Curr. Opin. Biotechnol. 10*, 365−369.
4. Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S. M., Harel, M., Remington, S. J., Silman, I., Schrag, J., Sussman, J. L., Verschueren, K. H. G., and Goldman, A. (1992) The α/β hydrolase fold, *Protein Eng. 5*, 197−211.
5. Damborsky, J., and Koca, J. (1999) Analysis of the reaction mechanism and substrate specificity of haloalkane dehalogenases by sequential and structural comparisons, *Protein Eng. 12*, 989−998.
6. Franken, S. M., Rozeboom, H. J., Kalk, K. H., and Dijkstra, B. W. (1991) Crystal structure of haloalkane dehalogenase: An enzyme to detoxify halogenated alkanes, *EMBO J. 10*, 1297−1302.
7. Newman, J., Peat, T. S., Richard, R., Kan, L., Swanson, P. E., Affholter, J. A., Holmes, I. H., Schindler, J. F., Unkefer, C. J., and Terwilliger, T. C. (1999) Haloalkane dehalogenase: Structure of a *Rhodococcus* enzyme, *Biochemistry 38*, 16105−16114.
8. Marek, J., Vevodova, J., Kuta-Smatanova, I., Nagata, Y., Svensson, L. A., Newman, J., Takagi, M., and Damborsky, J. (2000) Crystal structure of the haloalkane dehalogenase from *Sphingomonas paucimobilis* UT26, *Biochemistry 39*, 14082−14086.
9. Mani, S. V., Connell, D. W., and Braddock, R. D. (1991) Structure activity relationships for the prediction of biodegradability of environmental pollutants, *Crit. Rev. Environ. Control 21*, 217−236.
10. Peijnenburg, W. J. G. M., and Damborsky, J. (1996) *Biodegradability prediction*, pp 143, Kluwer Academic Publishers, Dordrecht, The Netherlands.
11. Damborsky, J., Lynam, M., and Kuty, M. (1998) Structure-biodegradability relationships for chlorinated dibenzo-p-dioxins and dibenzofurans, in *Biodegradation of Dioxins and Furans* (Wittich, R.-M., Ed.) pp 163−226, R. G. Landes Co., Austin, TX.
12. Lewis, D. F. (2003) P450 structures and oxidative metabolism of xenobiotics, *Pharmacogenomics 4*, 387−395.
13. Ortiz, A. R., Pisabarro, M. T., Gago, F., and Wade, R. C. (1995) Prediction of drug binding affinities by comparative binding energy analysis, *J. Med. Chem. 38*, 2681−2691.
14. Kmunicek, J., Luengo, S., Gago, F., Ortiz, A. R., Wade, R. C., and Damborsky, J. (2001) Comparative binding energy analysis of the substrate specificity of haloalkane dehalogenase from *Xanthobacter autotrophicus* GJ10, *Biochemistry 40*, 8905−8917.
15. Tomic, S., and Kojic-Prodic, B. (2002) A quantitative model for predicting enzyme enantioselectivity: Application to *Burkholderia cepacia* lipase and 3-(aryloxy)-1,2-propanediol derivatives, *J. Mol. Graphics Modell. 21*, 241−252.
16. Kmunicek, J., Bohac, M., Luengo, S., Gago, F., Wade, R. C., and Damborsky, J. (2003) Comparative binding energy analysis of haloalkane dehalogenase substrates: Modelling of enzyme-substrate complexes by molecular docking and quantum mechanic calculations, *J. Comput.-Aided Mol. Des. 17*, 299−311.
17. Tomic, S., Bertosa, B., Kojic-Prodic, B., and Kolosvary, I. (2004) Stereoselectivity of *Burkholderia cepacia* lipase towards secondary alcohols: Molecular modelling and 3D QSAR approach, *Tetrahedron Asymmetry 15*, 1163−1172.
18. Stewart, J. J. P. (1990) MOPAC: A semiepirical molecular-orbital program, *J. Comput.-Aided Mol. Des. 4*, 1−45.
19. Nagata, Y., Miyauchi, K., Damborsky, J., Manova, K., Ansorgova, A., and Takagi, M. (1997) Purification and characterization of haloalkane dehalogenase of a new substrate class from a γ-hexachlorocyclohexane-degrading bacterium, *Sphingomonas paucimobilis* UT26, *Appl. Environ. Microbiol. 63*, 3707−3710.
20. Nagata, Y., Hynkova, K., Damborsky, J., and Takagi, M. (1999) Construction and characterization of histidine-tagged haloalkane dehalogenase (LinB) of a new substrate class from a γ-hexachlorocyclohexane-degrading bacterium, *Sphingomonas paucimobilis* UT26, *Protein Expression Purif. 17*, 299−304.
21. Stewart, J. J. P. (1990) *MOPAC Manual*, version 6.0, Quantum Chemistry Program Exchange, Bloomington, IN.
22. Vriend, G. (1990) WHAT IF: A molecular modeling and drug design program, *J. Mol. Graphics 8*, 52−56.
23. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, *J. Comput. Chem. 19*, 1639−1662.
24. Solis, F. J., and Wets, R. J. B. (1981) Minimization by random search techniques, *Math. Oper. Res. 6*, 19−30.
25. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc. 117*, 5179−5197.
26. Geladi, P., and Kowalski, B. R. (1986) Partial Least-Squares Regression: A Tutorial, *Anal. Chim. Acta 185*, 1−17.
27. Bertolini, M. C., Schrag, J. D., Cygler, M., Ziomek, E., Thomas, D. Y., and Vernet, T. (1995) Expression and characterization of *Geotrichum candidum* lipase I gene. Comparison of specificity profile with lipase II, *Eur. J. Biochem. 228*, 863−869.
28. Pleiss, J., Fischer, M., and Schmid, R. D. (1998) Anatomy of lipase binding sites: The scissile fatty acid binding site, *Chem. Phys. Lipids 93*, 67−80.
29. Prokop, Z., Monincova, M., Chaloupkova, R., Klvana, M., Nagata, Y., Janssen, D. B., and Damborsky, J. (2003) Catalytic mechanism of the haloalkane dehalogenase LinB from *Sphingomonas paucimobilis* UT26, *J. Biol. Chem. 278*, 45094−45100.
30. Bohac, M., Nagata, Y., Prokop, Z., Prokop, M., Monincova, M., Koca, J., Tsuda, M., and Damborsky, J. (2002) Halide-stabilizing residues of haloalkane dehalogenases studied by quantum mechanic calculations and site-directed mutagenesis, *Biochemistry 41*, 14272−14280.
31. Hynkova, K., Nagata, Y., Takagi, M., and Damborsky, J. (1999) Identification of the catalytic triad in the haloalkane dehalogenase from *Sphingomonas paucimobilis* UT26, *FEBS Lett. 446*, 177−181.
32. Oakley, A. J., Prokop, Z., Bohac, M., Kmunicek, J., Jedlicka, T., Monincova, M., Kuta-Smatanova, I., Nagata, Y., Damborsky, J., and Wilce, M. C. J. (2002) Exploring the structure and activity of haloalkane dehalogenase from *Sphingomonas paucimobilis* UT26: Evidence for product and water mediated inhibition, *Biochemistry 41*, 4847−4855.
33. Streltsov, V. A., Prokop, Z., Damborsky, J., Nagata, Y., Oakley, A. J., and Wilce, M. C. J. (2003) Haloalkane dehalogenase LinB from *Sphingomonas paucimbilis* UT26: X-ray crystallographic studies of dehalogenation of brominated substrates, *Biochemistry 42*, 10104−10112.
34. Bosma, T., Damborsky, J., Stucki, G., and Janssen, D. B. (2002) Biodegradation of 1,2,3-trichloropropane through directed evolution and heterologous expression of a haloalkane dehalogenase gene, *Appl. Environ. Microbiol. 68*, 3582−3587.
35. Gray, K. A., Richardson, T. H., Kretz, K., Short, J. M., Bartnek, F., Knowles, R., Kan, L., Swanson, P. E., and Robertson, D. E. (2001) Rapid evolution of reversible denaturation and elevated melting temperature in a microbial haloalkane dehalogenase, *Adv. Synth. Catal. 343*, 607−616.
36. Nagata, Y., Prokop, Z., Marvanova, S., Sykorova, J., Monincova, M., Tsuda, M., and Damborsky, J. (2003) Reconstruction of mycobacterial dehalogenase Rv2579 by cumulative mutagenesis of haloalkane dehalogenase LinB, *Appl. Environ. Microbiol. 69*, 2349−2355.
37. Chaloupkova, R., Sykorova, J., Prokop, Z., Jesenska, A., Monincova, M., Pavlova, M., Tsuda, M., Nagata, Y., and Damborsky, J. (2003) Modification of activity and specificity of haloalkane dehalogenase from *Sphingomonas paucimobilis* UT26 by engineering of its entrance tunnel, *J. Biol. Chem. 278*, 52622−52628.

BI047912O