J. Mol. Biol. (1994) **244**, 125-132

Computer Analysis of Bacterial Haloacid Dehalogenases Defines a Large Superfamily of Hydrolases with Diverse Specificity

Application of an Iterative Approach to Database Search

Eugene V. Koonin and Roman L. Tatusov

National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, U.S.A.

Using an iterative approach to sequence database search that combines scanning with individual amino acid sequences and with alignment blocks, we show that bacterial haloacid dehalogenases (HADs) belong to a large superfamily of hydrolases with diverse substrate specificity. The superfamily also includes epoxide hydrolases, different types of phosphatases, and numerous uncharacterized proteins from eubacteria, eukaryotes, and Archaea. Nine putative proteins of the HAD superfamily with functions unknown, in addition to two known enzymes, were found in *Escherichia coli* alone, making it one of the largest groups of enzymes and indicating that a variety of hydrolytic enzyme activities remain to be described. Many of the proteins with known enzymatic activities in the HAD superfamily are involved in detoxification of xenobiotics or metabolic by-products. All the proteins in the superfamily contain three conserved sequence motifs. Along with the conservation of the predicted secondary structure, motifs I, II, and III include a conserved aspartic acid residue, a lysine, and a nucleophile, namely aspartic acid or serine, respectively. A specific role in the catalysis of the hydrolysis of carbon—halogen and other bonds is assigned to each of these residues.

Keywords: conserved sequence motifs; protein superfamilies; dehalogenases; epoxide hydrolases; phosphatases; enzyme evolution

Bacterial hydrolytic dehalogenases are a group of enzymes that inactivate halogenated aliphatic hydrocarbons by hydrolysis of carbon–halogen bonds and are essential for detoxification of many chlorinated compounds (Weightman *et al.*, 1982; Hardman, 1991). They are divided into haloalkane dehalogenases and haloacid dehalogenases. Nucleotide sequences of genes coding for both types of enzymes have been reported (Janssen *et al.*, 1989; Schneider *et al.*, 1991; Van der Ploeg *et al.*, 1991; Murdiyatmo *et al.*, 1992; Jones *et al.*, 1992; Kawasaki *et al.*, 1992; Barth *et al.*, 1992) and the tertiary structure of a haloalkane dehalogenase (HALO†) has been determined (Franken *et al.*, 1991). Recently, the structure of HALO enzyme–substrate complex has been studied in great detail and it has been shown that the reaction proceeds *via* the formation of a covalent ester intermediate (Verschueren *et al.*, 1993*a*,*b*).

Enzymes that are involved in the detoxification of xenobiotics are likely to be a relatively recent evolutionary invention(s) and it is of interest to find out from what "normal" metabolic enzymes they might have evolved. Initial amino acid sequence comparisons have suggested that the hydrolytic dehalogenases are polyphyletic. The haloalkane dehalogenase from *Xanthobacter* showed surprising sequence similarity to eukaryotic epoxide hydrolases (Janssen *et al.*, 1989) and, subsequently, sequence motifs that are conserved in these proteins have been identified in several other hydrolases (Arand *et al.*, 1994). In addition, structural similarities have been reported to exist between haloalkane dehalogenase and several very different hydrolytic enzymes including dienelactone hydrolase, acetylcholine esterase, carboxypeptidase and triacylglycerol lipase, all of which appear to adopt the so-called α–β hydrolase structural fold (Ollis *et al.*, 1992).

Haloacid dehalogenases have not been studied in comparable detail. Several L-2-haloalkanoic acid dehalogenases from *Pseudomonas* and related bacteria have been found to comprise a highly conserved

Correspondence: E. V. Koonin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, MD 20894, U.S.A.

† Abbreviations used: HALO, haloalkane dehalogenase; HAD, haloacid dehalogenase; NRDB, non-redundant database; ORF, open reading frame.

0022-2836/94/460125-08 $08.00/0

family, which has been reported to be unrelated to haloalkane dehalogenases or any other proteins (Van der Ploeg *et al.*, 1991; Murdiyatmo *et al.*, 1992; Jones *et al.*, 1992). D-2-haloalkanoic acid and D,L-haloalkanoic acid dehalogenases appeared to be unique among known proteins (Kawasaki *et al.*, 1992; Barth *et al.*, 1992).

We concentrate here on the analysis of the amino acid sequence of the haloacid dehalogenases (HADs) using an iterative approach to database search that combines screening with individual sequences and multiple alignment blocks. We show that HADs belong to a large and ubiquitous superfamily of hydrolases with widely different substrate specificites, of which only a few have been functionally characterized. The putative catalytic amino acid residues are identified and their roles in catalysis are predicted.

Amino acid and nucleotide sequences were from the SWISS-PROT, PIR and GenBank databases that are combined in the Non-Redundant sequence DataBase (NRDB) at the National Center for Biotechnology Information (NIH).

Amino acid sequences were compared with the NRDB using programs based on the BLAST algorithm (Altschul *et al.*, 1990). The BLASTP program was used to screen the amino acid sequence database and the TBLASTN program was used to screen the conceptual translation of the nucleotide sequence database in the six reading frames (Altschul *et al.*, 1994). The BLAST algorithm computes the probability of the observed alignments being obtained by chance (P value) using the statistical theory for high-scoring sequence segments (Karlin & Altschul, 1990, 1993). Compositionally biased regions of query sequences that tend to produce spurious hits in database searches were excluded from the analysis using the SEG program (Wootton & Federhen, 1993; Altschul *et al.*, 1994).

Database search for conserved segments similar to multiple alignment blocks was performed using a recently developed iterative procedure, called MoST (Motif Search Tool), a full description of which appears elsewhere (Tatusov *et al.*, 1994). Briefly, the multiple alignment blocks are initially constructed by parsing consistent segments from the ungapped pairwise alignments produced by a BLAST search using the CAP (Consistent Alignment Parser) program. These blocks are converted into position-dependent weight matrices using a method that combines the observed amino acid residue frequencies for each column with *a priori* knowledge of amino acid relationships (Brown *et al.*, 1993). Using these matrices, scores are computed for all segments of the corresponding length in the amino acid sequence database, and the observed distribution of scores is compared with the theoretical distribution. The ratio (R value) of the expected number of sequence segments with a given score to the observed number is then used as a cut-off in database searches.

Multiple alignments were generated using the MACAW program (Schuler *et al.*, 1991).

For the cluster analysis of related sequences, a program called CLUS was written that divides a sequence set into subsets of sequences connected by BLASTP scores above a chosen cut-off.

Protein secondary structure was predicted using the PHD program that has been reported to yield an accuracy of over 70% (Rost & Sander, 1993).

Figure 1 schematically depicts our approach to delineating protein superfamilies that includes multiple, alternate rounds of database search with individual sequences using the BLASTP or TBLASTN programs and block search using the MoST program. The BLAST and MoST searches complement one another. It has been shown that the block search using MoST frequently selects a number of sequences that are not detectable by BLAST at a significant level (Tatusov *et al.*, 1994; Koonin *et al.*, 1994). Conversely, some sequences that contain a deviant version of the conserved motif may be recognized by BLAST but not by MoST.

When the HAD sequences were compared with the NRDB using BLASTP, varying levels of sequence similarity were revealed with two other groups of hydrolases, namely epoxide hydrolases and phosphoglycolate phosphatases. For example, with the *Pseudomonas sp.* DehCl sequence, the P values of 4.2×10^{-8} and 1.6×10^{-3} were observed for the *Alcaligenes eutrophus* phosphoglycolate phosphatase and the rat cytosolic epoxide hydrolase, respectively. Further analysis by repeated cycles of database search using BLAST and MoST as outlined above, resulted in a set of about 50 proteins with three distinct conserved motifs; we designate this set of proteins the HAD superfamily (Figure 2). The inclusion of each sequence in the superfamily was supported by either statistically significant ($P<0.001$) similarity with at least one other member (detected by BLAST) or a significant ($r<0.02$) score in the MoST search (with either motif I or motif II), or both. Sequence conservation in a subset of this superfamily including phosphoglycolate phosphatases from *Alcaligenes eutrophus* and several putative proteins with unknown functions but not the dehalogenases has been recently described (Schaferjohann *et al.*, 1993).

The reactions catalyzed by the known enzymes in the HAD superfamily include hydrolysis of very different molecules (Table 1). Despite this variability in substrate specificity, the presence of three similarly located, highly conserved motifs strongly suggested a functional and evolutionary relationship between all of the proteins. Motifs II and III were closely spaced together whereas the upstream motif I was separated from motif II by a non-conserved spacer widely varying in length. Only two amino acid residues, namely aspartic acid in motif I and lysine in motif II, were strictly conserved in all of the aligned sequences. Secondary structure prediction taking into account multiple alignment of closely related sequences (Rost & Sander, 1993) indicated that motif I and motif II each comprised a β-strand-loop-α-helix unit, with the conserved aspartic acid and lysine located in

the respective loops. Motif III consisted of a hydrophobic β-strand terminating at an aspartic acid (replaced in one sequence by a glutamic acid) or a serine (Figure 2).

The observed pattern of amino acid residue conservation allowed a specific interpretation in terms of the catalytic mechanism. Two different reactions mechanisms with inversion of configuration

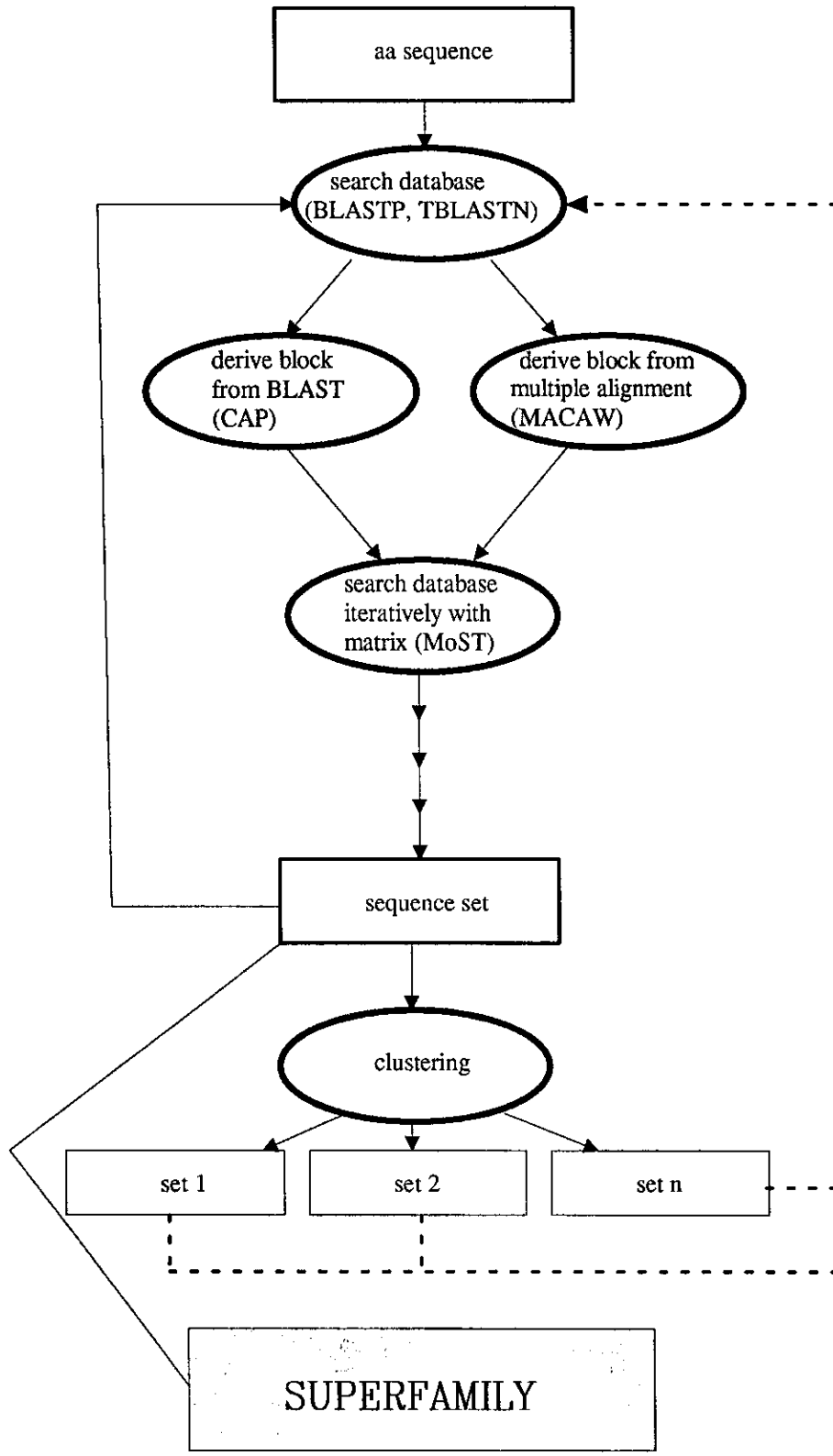

Figure 1. Iterative strategy for identification of protein superfamilies. The MoST search was run iteratively, until convergence. Broken lines indicate optional steps. aa, amino acid.

at the α-carbon have been postulated for haloacid dehalogenases. The first mechanism involves generalized base catalysis, with a nucleophilic group in the enzyme (supposedly a histidine) activating a water molecule, which has been proposed to act as the actual hydrolytic agent (Little & Williams, 1971). A positively charged amino acid residue in the enzyme has been thought to interact with the carboxylic group of the substrate. The second mechanism includes the formation of a covalent acyl-substrate intermediate, involving an acidic amino acid residue, which is subsequently hydrolyzed by an activated water molecule, a process predicted to be facilitated by a nucleophilic residue in the enzyme (Goldman *et al.*, 1968). Similarly to the first mechanism, this scheme includes an essential positively charged residue. Recent site-directed mutagenesis experiments have shown that Asp10 of *Pseudomonas sp.* 2-haloalkanoic acid dehalogenase I (the conserved aspartic acid in motif I of DehCI in Figure 2) is essential for the enzymatic activity, thus providing evidence for the second mechanism (Schneider *et al.*, 1993). Our finding of the three conserved motifs (Figure 2), containing, respectively, an invariant aspartic acid, a lysine and either an aspartic acid or a serine, i.e. a nucleophile, is compatible with this mechanism. We propose that these conserved residues may be directly involved in the catalysis of the hydrolysis of the respective chemical bonds in all the members of the HAD superfamily (Figure 3). This mechanism is supported also by studies of the catalytic mechanism of phosphoglycolate phosphatase that have detected a covalent intermediate, probably involving an acidic amino acid residue (Seal & Rose, 1987). While the function of the conserved Asp in ester intermediate formation has already been predicted (Schneider *et al.*, 1993), the positively charged residue and the nucleophile so far remain unidentified; these residues are plausible targets for further site-directed mutagenesis studies. The formation of the covalent acyl-substrate intermediate relates this mechanism to the mechanism recently described for the haloalkane dehalogenase (Verschueren *et al.*, 1993*a*, *b*), but the participation of a positively charged residue is unique for the HAD superfamily. The negatively charged group of the substrate interacting with this residue is predicted to be the carboxylic group in haloalkanoic acids (Figure 3) and the phosphate in the numerous compounds hydrolyzed by the phosphatases belonging to the HAD superfamily (Table 1).

Cluster analysis based on BLAST scores revealed several distinct families within the HAD superfamily (Figure 2). The haloacid dehalogenases, together with epoxide hydrolases, phosphoglycolate phosphatases, histidinol phosphate phosphatases, and several uncharacterized ORF products comprised one large family; another family included nitrophenyl phosphatases and related putative proteins; three more families contained only functionally uncharacterized (putative) proteins. Phosphoserine phosphatases, trehalose phosphatases and several unknown ORF products did not show significant pairwise similarity to other proteins in the HAD superfamily and were included on the basis of motif conservation alone.

The HAD superfamily is represented by numerous proteins in a single organism, especially in eubacteria. For example, it includes 11 (putative) proteins from *E. coli* (Figure 2), and extrapolating from the currently available fraction of the chromosome sequence (about 60%) to the complete genome, *E. coli*

Figure 2. The three conserved motifs in the HAD superfamily. The alignment was generated by the iterative process depicted in Figure 1. Several related partial sequences encoded by Expressed Sequence Tags (ESTs) from different organisms are not shown. Distinct groups of protein sequences revealed by clustering based on BLASTP probabilities are separated by blanks. Each group includes sequences that could be represented as a connected graph, with each edge associated with a BLAST score greater than 83, which approximately corresponds to $P<0.001$ when the entire NRDB is searched. The consensus is shown as a frequency profile (the frequencies are indicated in the rightmost column) and includes all amino acid residues that were conserved in at least 50% of the aligned sequences; residues that conform to the consensus are indicated by bold type; U designates a bulky aliphatic residue (I, L, V or M) and & designates a bulky hydrophobic residue, either aliphatic or aromatic (I, L, V, M, F, Y or W). NUCLEO designates a nucleophilic residue. Exclamation marks show residues that are predicted to form the hydrolase catalytic triad. Distances from the protein ends and distances between blocks I and II are indicated by numbers; for incomplete sequences, the numbers are shown in parentheses. The indicated secondary structure is the consensus of predictions for individual sequences (a indicates an α-helix; b indicates a β-strand; and l indicates a loop; in positions where the prediction was uncertain, no symbol is indicated). The rightmost field includes the sequence accession numbers in the SWISS-PROT, PIR(p) or GenBank databases. For those ORFs that were extracted from GenBank and are included in the feature tables, the number of the ORF in the respective entry is indicated. Several of the included ORFs are not in the feature tables (indicated by a prime). They have been initially detected by using the TBLASTN program and were subsequently conceptually translated. Some of these ORFs contained apparent frameshift errors and were reconstructed so as to maximize the sequence similarity. Complete information on the location of these ORFs in the respective nucleotide sequences and on the changes that have been made for reconstruction is available from the authors upon request. Abbreviations of organism names: Psp, *Pseudomonas sp.*; Pc, *P. cepacea*; Pp, *P. putida*; Msp, *Moraxella sp.*; Xa, *Xanthomonas authotrophicus*; Ae, *Alcaligenes eutrophus*; Ec, *Escherichia coli*; Dm, *Desulfurococcus mobilis*; Lc, *Lactobacillus casei*; Sc, *Saccharomyces cerevisiae*; Ce, *Caenorhabditis elegans*; Ml, *Mycobacterium leprae*; Sp, *Schizosaccharomyces pombe*; At, *Arabidopsis thaliana*; Bst, *Bacillus stearothermophilus*; Mh, *Mycoplasma hominis*; Mc, *Mycoplasma capricolum*; Bb, *Borrelia burgdorferi*; Lh, *Lactobacillus helveticus*; Bs, *Bacillus subtilis*; Lp, *Lactobacillus plantarum*; Sf, *Streptomyces fradiae*; Ko, *Klebsiella oxytoca*; AcNPV, *Autographa californica* nuclear polyhedrosis virus.

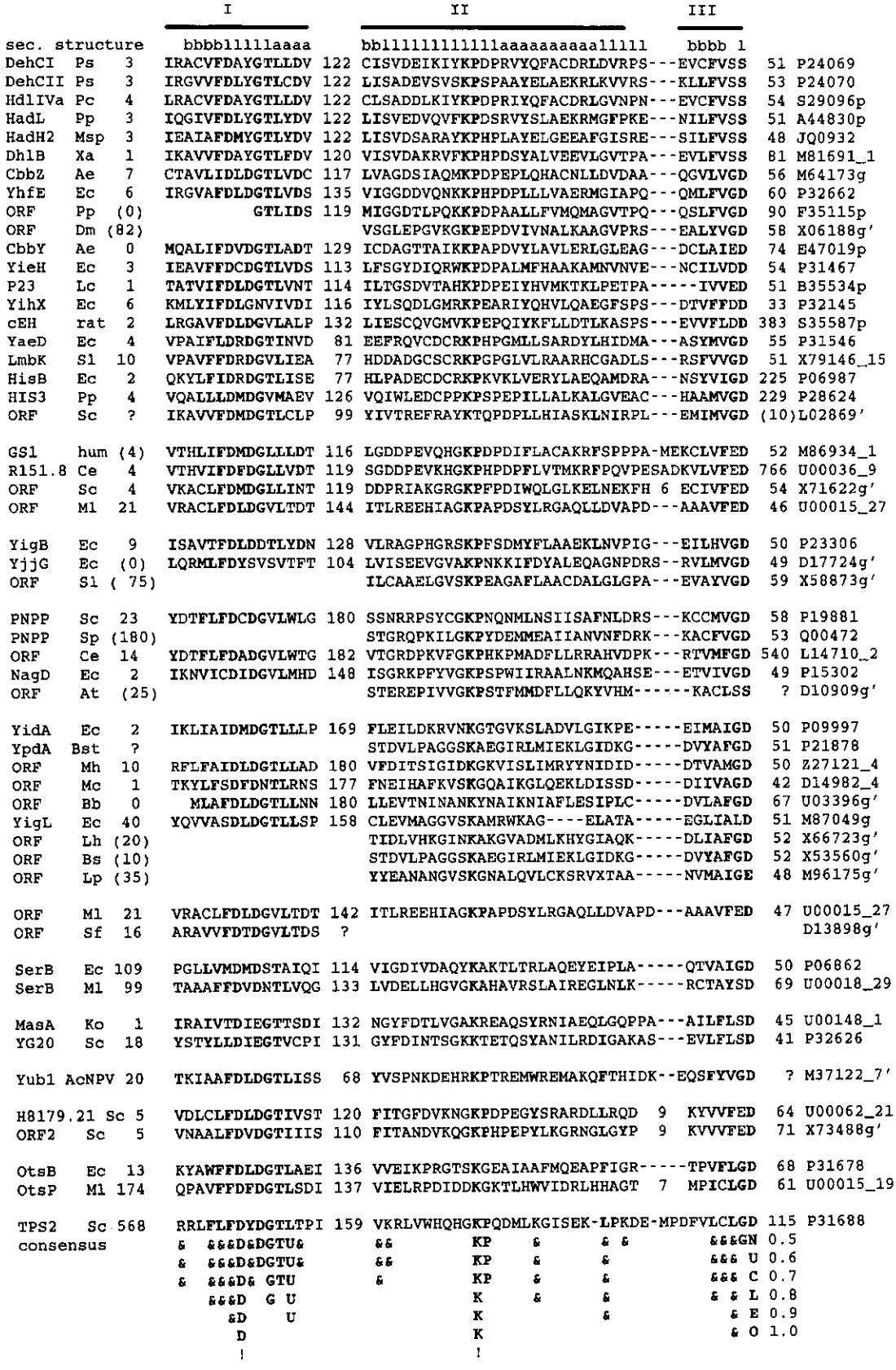
I
II
III
sec. structure bbbblllllaaaa bbllllllllllllaaaaaaaaaaallllll bbbb l
DehCI Ps 3 IRACVFDAYGTLLDV 122 CISVDEIKIYKPDPRVYQFACDRLDVRPS---EVCFVSS 51 P24069
DehCII Ps 3 IRGVVFDLYGTLCDV 122 LISADEVSVSKPSPAAYELAEKRLKVVRS---KLLFVSS 53 P24070
HdlIVa Pc 4 LRACVFDAYGTLLDV 122 CLSADDLKIYKPDPRIYQFACDRLGVNPN---EVCFVSS 54 S29096p
HadL Pp 3 IQGIVFDLYGTLYDV 122 LISVEDVQVFKPDSRVYSLAEKRMGFPKE---NILFVSS 51 A44830p
HadH2 Msp 3 IEAIAFDMYGTLYDV 122 LISVDSARAYKPHPLAYELGEEAFGISRE---SILFVSS 48 JQ0932
DhlB Xa 1 IKAVVFDAYGTLFDV 120 VISVDAKRVFKPHPDSYALVEEVLGVTPA---EVLFVSS 81 M81691_1
CbbZ Ae 7 CTAVLIDLDGTLVDC 117 LVAGDSIAQMKPDPEPLQHACNLLDVDAA---QGVLVGD 56 M64173g
YhfE Ec 6 IRGVAFDLDGTLVDS 135 VIGGDDVQNKKPHPDPLLLVAERMGIAPQ---QMLFVGD 60 P32662
ORF Pp (0) GTLIDS 119 MIGGDTLPQKKPDPAALLFVMQMAGVTPQ---QSLFVGD 90 F35115p
ORF Dm (82) VSGLEPGVKGKPEPDVIVNALKAAGVPRS---EALYVGD 58 X06188g'
CbbY Ae 0 MQALIFDVDGTLADT 129 ICDAGTTAIKKPAPDVYLAVLERLGLEAG---DCLAIED 74 E47019p
YieH Ec 3 IEAVFFDCDGTLVDS 113 LFSGYDIQRWKPDPALMFHAAKAMNVNVE---NCILVDD 54 P31467
P23 Lc 1 TATVIFDLDGTLVNT 114 ILTGSDVTAHKPDPEIYHVMKTKLPETPA-----IVVED 51 B35534p
YihX Ec 6 KMLYIFDLGNVIVDI 116 IYLSQDLGMRKPEARIYQHVLQAEGFSPS---DTVFFDD 33 P32145
cEH rat 2 LRGAVFDLDGVLALP 132 LIESCQVGMVKPEPQIYKFLLDTLKASPS---EVVFLDD 383 S35587p
YaeD Ec 4 VPAIFLDRDGTINVD 81 EEFRQVCDCRKPHPGMLLSARDYLHIDMA---ASYMVGD 55 P31546
LmbK S1 10 VPAVFFDRDGVLIEA 77 HDDADGCSCRKPGPGLVLRAARHCGADLS---RSFVVGD 51 X79146_15
HisB Ec 2 QKYLFIDRDGTLISE 77 HLPADECDCRKPKVKLVERYLAEQAMDRA---NSYVIGD 225 P06987
HIS3 Pp 4 VQALLLDMDGVMAEV 126 VQIWLEDCPPKPSPEPILLALKALGVEAC---HAAMVGD 229 P28624
ORF Sc ? IKAVVFDMDGTLCLP 99 YIVTREFRAYKTQPDPLLHIASKLNIRPL---EMIMVGD (10)L02869'
GS1 hum (4) VTHLIFDMDGLLLDT 116 LGDDPEVQHGKPDPDIFLACAKRFSPPPA-MEKCLVFED 52 M86934_1
R151.8 Ce 4 VTHVIFDFDGLLVDT 119 SGDDPEVKHGKPHPDPFLVTMKRFPQVPESADKVLVFED 766 U00036_9
ORF Sc 4 VKACLFDMDGLLINT 119 DDPRIAKGRGKPFPDIWQLGLKELNEKFH 6 ECIVFED 54 X71622g'
ORF Ml 21 VRACLFDLDGVLTDT 144 ITLREEHIAGKPAPDSYLRGAQLLDVAPD---AAAVFED 46 U00015_27
YigB Ec 9 ISAVTFDLDDTLYDN 128 VLRAGPHGRSKPFSDMYFLAAEKLNVPIG---EILHVGD 50 P23306
YjjG Ec (0) LQRMLFDYSVSVTFT 104 LVISEEVGVAKPNKKIFDYALEQAGNPDRS---RVLMVGD 49 D17724g'
ORF S1 (75) ILCAAELGVSKPEAGAFLAACDALGLGPA---EVAYVGD 59 X58873g'
PNPP Sc 23 YDTFLFDCDGVLWLG 180 SSNRRPSYCGKPNQNMLNSIISAFNLDRS---KCCMVGD 58 P19881
PNPP Sp (180) STGRQPKILGKPYDEMMEAIIANVNFDRK---KACFVGD 53 Q00472
ORF Ce 14 YDTFLFDADGVLWTG 182 VTGRDPKVFGKPHKPMADFLLRRAHVDPK---RTVMFGD 540 L14710_2
NagD Ec 2 IKNVICDIDGVLMHD 148 ISGRKPFYVGKPSPWIIRAALNKMQAHSE---ETVIVGD 49 P15302
ORF At (25) STEREPIVVGKPSTFMMDFLLQKYVHM------KACLSS ? D10909g'
YidA Ec 2 IKLIAIDMDGTLLLP 169 FLEILDKRVNKGTGVKSLADVLGIKPE-----EIMAIGD 50 P09997
YpdA Bst ? STDVLPAGGSKAEGIRLMIEKLGIDKG-----DVYAFGD 51 P21878
ORF Mh 10 RFLFAIDLDGTLLAD 180 VFDITSIGIDKGKVISLIMRYYNIDID-----DTVAMGD 50 Z27121_4
ORF Mc 1 TKYLFSDFDNTLRNS 177 FNEIHAFKVSKGQAIKGLQEKLDISSD-----DIIVAGD 42 D14982_4
ORF Bb 0 MLAFDLDGTLLNN 180 LLEVTNINANKYNAIKNIAFLESIPLC-----DVLAFGD 67 U03396g'
YigL Ec 40 YQVVASDLDGTLLSP 158 CLEVMAGGVSKAMRWKAG----ELATA-----EGLIALD 51 M87049g
ORF Lh (20) TIDLVHKGINKAKGVADMLKHYGIAQK-----DLIAFGD 52 X66723g'
ORF Bs (10) STDVLPAGGSKAEGIRLMIEKLGIDKG-----DVYAFGD 52 X53560g'
ORF Lp (35) YYEANANGVSKGNALQVLCKSRVXTAA-----NVMAIGE 48 M96175g'
ORF Ml 21 VRACLFDLDGVLTDT 142 ITLREEHIAGKPAPDSYLRGAQLLDVAPD---AAAVFED 47 U00015_27
ORF Sf 16 ARAVVFDTDGVLTDS ? D13898g'
SerB Ec 109 PGLLVMDMDSTAIQI 114 VIGDIVDAQYKAKTLTRLAQEYEIPLA-----QTVAIGD 50 P06862
SerB Ml 99 TAAAFFDVDNTLVQG 133 LVDELLHGVGKAHAVRSLAIREGLNLK-----RCTAYSD 69 U00018_29
MasA Ko 1 IRAIVTDIEGTTSDI 132 NGYFDTLVGAKREAQSYRNIAEQLGQPPA---AILFLSD 45 U00148_1
YG20 Sc 18 YSTYLLDIEGTVCPI 131 GYFDINTSGKKTETQSYANILRDIGAKAS---EVLFLSD 41 P32626
Yub1 AcNPV 20 TKIAAFDLDGTLISS 68 YVSPNKDEHRKPTREMWREMAKQFTHIDK--EQSFYVGD ? M37122_7'
H8179.21 Sc 5 VDLCLFDLDGTIVST 120 FITGFDVKNGKPDPEGYSRARDLLRQD 9 KYVVFED 64 U00062_21
ORF2 Sc 5 VNAALFDVDGTIIIS 110 FITANDVKQGKPHPEPYLKGRNGLGYP 9 KVVVFED 71 X73488g'
OtsB Ec 13 KYAWFFDLDGTLAEI 136 VVEIKPRGTSKGEAIAAFMQEAPFIGR-----TPVFLGD 68 P31678
OtsP Ml 174 QPAVFFDFDGTLSDI 137 VIELRPDIDDKGKTLHWVIDRLHHAGT 7 MPICLGD 61 U00015_19
TPS2 Sc 568 RRLFLFDYDGTLTPI 159 VKRLVWHQHGKPQDMLKGISEK-LPKDE-MPDFVLCLGD 115 P31688
consensus & &&&D&DGTU& && KP & & & &&&GN 0.5
& &&&D&DGTU& && KP & & &&& U 0.6
& &&&D& GTU & KP & & &&& C 0.7
&&&D G U K & & & & L 0.8
&D U K & & E 0.9
D K & O 1.0
!
!

Figure 2

Table 1

Proteins with known enzymatic activity in the HAD superfamily

Gene/protein	Organism	*Protein size* (aa)/ quaternary structure	Enzyme	Reaction	Metabolic pathway/ function	References
DehCI, DehCII	*Pseudomonas sp.*	227 × 4 229 × 4	2-Haloalkanoic acid dehalogenase	Monochloroacetate + H_2O = glycolate + HCl[1]	Assimilation of haloalkanoic acids and haloalkanes	Hardman (1991); Schneider *et al.* (1993)
HadL	*P. putida*	227 × 4	l-2-haloalkanoic acid dehalogenase	l-2-Chloropropionate + H_2O = d-lactate + HCl	Assimilation of haloalkanoic acids and haloalkanes	Hardman (1991); Jones *et al.* (1992)
CbbZ	*Alcaligenes eutrophus*	231 × 3	2-phosphoglycolate phosphatase	2-phosphoglycolate + H_2O = glycolate + P_i	Assimilation of 2-phosphoglycolate	Schaferjohann *et al.* (1993)
cEH	Rat	554 × 2	Cytosolic epoxide hydrolase	Epoxide + H_2O = trans-diol	Detoxification of xenobiotics: ??	Knehr *et al.* (1993)
MasA (E1 enzyme)	*Klebsiella oxytoca*	229 × 1	E1 enzyme	2,3-diketo-1-phospho-5-thiomethylpentane + H_2O + O2 = 2-keto-4-methylthiobutyrate + formate + P_i	Salvage pathway of methionine synthesis	Myers *et al.* (1993)
HisB	*E. coli*	355 × 4	Imidazoleglycerol-phosphate dehydratase: histidinol phosphatase	*Imidazoleglycerol-phosphate* = imidazole-oxopropyl-phosphate + H_2O l-histidinol phosphate + H_2O = l-histidinol + P_i	*7th and 9th steps in* histidine biosynthesis	Winkler (1987)
PHO2	*S. pombe*	269 × 2	*p*-nitrophenyl-phosphatase	*p*-nitrophenyl-phosphate + H_2O = *p*-nitrophenol + P_i	Unknown	Yang *et al.* (1991)
SerB	*E. coli*	322 × ?	Phosphoserine phosphatase	3-Phosphoserine + H_2O = serine + P_i	C4 pathway of serine biosynthesis	Stauffer (1987); Ravnikar & Sommerville (1987)
OtsB (PexA) TPS2	*E. coli* Yeast	266 × ? 894 × 1†	Trehalose-6-phosphate Trehalose-6-phosphate synthase/phosphatase	Trehalose-6-phosphate + H_2O = Trehalose-6-phosphate + H_2O = Trehalose + P_i	Trehalose synthesis Trehalose synthesis	Strom & Kaasen (1993)

† TPS2 forms a heterotrimer with TPS1 (trehalose-6-phosphate synthase) and TPS3.

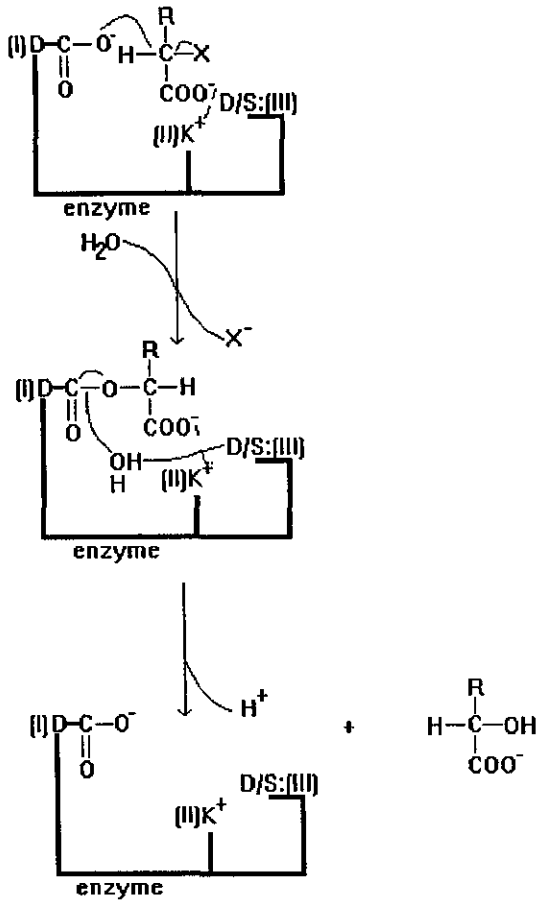

Figure 3. Proposed reaction mechanism for the hydrolases of the HAD superfamily. The scheme is an adaptation of the HAD reaction mechanism B discussed by Van der Ploeg *et al.* (1991). (I)D, (II)K^+ and D/S:(III) indicate the predicted catalytic residues that are located in the respective conserved motifs.

probably encodes about 20 enzymes of this superfamily. For only two of them, namely serine phosphatase (SerB) and trehalose phosphatase (OtsB), the actual enzymatic activity has been reported. Thus, while widespread, the enzymes of the HAD superfamily seem to be under-represented in the current catalogue of biochemical reactions. In addition to the general prediction that all the proteins in this superfamily are hydrolases, specific activity could be predicted for some of the uncharacterized ORF products from sequence similarity (Figure 2). In particular, YhfE, NagD, and other putative proteins in the nitrophenyl phosphatase family are likely to have phosphatase activity, whereas YihX may be an epoxide hydrolase. Remarkably, many of the proteins in the HAD superfamily are "defense" enzymes involved in detoxification of xenobiotics (dehalogenases) or metabolic by-products (phosphoglycolate phosphatase, trehalose phosphatase and epoxide hydrolase). The biological substrates of nitrophenyl phosphatases are not known and these enzymes also may be involved in detoxification reactions. Study of the uncharacterized putative proteins belonging to the HAD superfamily may reveal new biochemical pathways.

Most of the proteins in the HAD superfamily are relatively small, with the characteristic size of 200 to 250 amino acid residues, and appear to consist of the hydrolase domain alone. On the other hand, epoxide hydrolases, histidinol phosphate phosphatases, the yeast trehalose phosphatase and several uncharacterized proteins contain additional domains. Remarkably, in the epoxide hydrolases, the C-terminal domain is another type of hydrolase belonging to the superfamily that also includes the haloalkane dehalogenase (Janssen *et al.*, 1989; Ollis *et al.*, 1992; Arand *et al.*, 1994).

In delineating the HAD superfamily, we applied an iterative computer-assisted strategy combining database screening for pairwise sequence similarity and for similarity to alignment blocks. We believe that this approach will have general application in the analysis of protein sequence databases.

We are grateful to Drs S. Altschul, P. Bork, and D. Lipman for helpful discussions, to Bobby Baum for useful suggestions, and to P. Bork for critical reading of the initial draft of this manuscript.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119–129.

Arand, M., Grant, D. F., Beetham, J. K., Friedberg, T., Oesch, F. & Hammock, B. D. (1994). Sequence similarity of mammalian epoxide hydrolases to the bacterial haloalkane dehalogenase and other related proteins. Implication for the potential catalytic mechanism of enzymatic epoxide hydrolysis. *FEBS Letters*, **338**, 251–256.

Barth, P. T., Bolton, L. & Thomson, J. C. (1992). Cloning and partial sequencing of an operon encoding two *Pseudomonas putida* haloalkanoic acid dehalogenases of opposite stereospecificity. *J. Bacteriol.* **174**, 2612–2619.

Brown, M., Hughey, R., Krogh, A., Mian, S., Sjolander, K. & Haussler, D. (1993). Using Dirichlet mixture priors to derive hidden Markov models for protein families. In *Proc. First Int. Conf. Intelligen Systems Mol. Biol.* (Hunter, L., Searls, D. & Shavlik, J., eds), pp. 47–55, AAAI Press, Menlo Park, CA.

Franken, S. M., Roseboom, H. J., Kalk, K. H. & Dijkstra, B. W. (1991). Crystal structure of haloalkane dehalogenase: an enzyme to detoxify halogenated alkanes. *EMBO J.* **10**, 1297–1302.

Goldman, P., Milne, G. W. A. & Keister, D. B. (1968). Carbon-halogen bond cleavage. III. Studies on bacterial halidohydrolases. *J. Biol. Chem.* **243**, 428–434.

Hardman, D. J. (1991). Biotransformation of halogenated compounds. *CRC Crit. Rev. Biotechnol.* **11**, 1–40.

Janssen, D. B., Pries, F., van der Ploeg, J., Kazemier, B., Terpstra, P. & Witholt, B. (1989). Cloning of 1,2-dichloroethane degradation genes of *Xanthobacter autotrophicus* GJ10 and expression and sequencing of the *dhlA* gene. *J. Bacteriol.* **171**, 6791–6799.

Jones, D. H., Barth, P. T., Byrom, D. & Thomas, C. M. (1992). Nucleotide sequence of the structural gene encoding a 2-haloalkanoic acid dehalogenase of *Pseudomonas putida* strain AJ1 and purification of the encoded protein. *J. Gen Microbiol.* **138**, 675–683.

Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Nat. Acad. Sci. U.S.A.* **87**, 2264–2268.

Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Nat. Acad. Sci., U.S.A.* **90**, 5873–5877.

Kawasaki, H., Tsuda, K., Matsushita, I. & Tonomura, K. (1992). Lack of homology between two haloacetate dehalogenase genes encoded on a plasmid from *Moraxella sp.* strain B. *J. Gen. Microbiol.* **138**, 1317–1323.

Knehr, M., Thomas, H., Arand, M., Gebel, T., Zeller, H. D. & Oesch, F. (1993). Isolation and characterization of a cDNA encoding rat liver cytosolic epoxide hydrolase and its functional expression in *Escherichia coli*. *J. Biol. Chem* **268**, 17623–17627.

Koonin, E. V., Mushegian, A. R., Tatusov, R. L., Altschul, S. F., Bryant, S. H., Bork, P. & Valencia, A. (1994). Eukaryotic translation elongation factor 1 g contains a glutathione transferase domain—study of a diverse, ancient protein superfamily using motif search and structural modeling. *Protein Sci.*, in the press.

Little, M. & Williams, P. A. (1971). A bacterial halidohydrolase, its purification, some properties and its modification by specific amino acid reagents. *Eur. J. Biochem.* **21**, 99–109.

Murdiyatmo, U., Asmara, W., Tsang, J. S., Baines, A. J., Bull, A. T. & Hardman, D. J. (1992). Molecular biology of the 2-haloacid halidohydrolase IVa from *Pseudomonas cepacia* MBA4. *Biochem. J.* **284**, 87–93.

Myers, R., Wray, J. W., Fish, S. & Abeles, R. H. (1993). Purification and characterization of an enzyme involved in oxidative carbon-carbon cleavage reactions in the methionine salvage pathway of *Klebsiella pneumoniae*. *J. Biol. Chem.* **268**, 24785–24791.

Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S. M., Harel, M., Remington, S. J., Silman, I., Schrag, J., *et al.* (1992). The alpha/beta hydrolase fold. *Protein Eng.* **5**, 197–211.

Ravnikar, P. D. & Somerville, R. L. (1987). Genetic characterization of a highly efficient alternative pathway of serine biosynthesis in *Escherichia coli*. *J. Bacteriol.* **169**, 2611–2617.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.

Schaferjohann, J., Yoo, J.-G., Kusian, B. & Bowein, B. (1993). The cbb operons of the facultative chemoautotroph *Alcaligenes eutrophus* encode phosphoglycolate phosphatases. *J. Bacteriol.* **175**, 7329–7340.

Schneider, B., Muller, R., Frank, R. & Lingens, F. (1991). Complete nucleotide sequence and comparison of the structural genes of two 2-haloalkanoic acid dehalogenases from *Pseudomonas sp.* strain CBS3. *J. Bacteriol.* **173**, 1530–1535.

Schneider, B., Muller, R., Frank, R. & Lingens, F. (1993). Site-directed mutagenesis of the 2-haloalkanoic acid dehalogenase I gene from *Pseudomonas sp.* strain CBS3 and its effect on catalytic activity. *Biol. Chem. Hoppe-Seyler.* **374**, 489–496.

Schuler, G. D., Altschul, S. F. & Lipman, D. J. (1991). A workbench for multiple alignment construction and analysis. *Proteins: Struct. Funct. Genet.* **9**, 180–190.

Seal, S. N. & Rose, Z. B. (1987). Characterization of a phosphoenzyme intermediate in the reaction of phosphoglycolate phosphatase. *J. Biol. Chem.* **262**, 13496–13500.

Stauffer, G. A. (1987). Biosynthesis of serine and glycine. In Escherichia coli *and* Salmonella typhimurium. *Cellular and Molecular Biology* (Ingraham, J. L., Low, K. B., Magasanik, B., Schaechter, M. & Umbarger, H. E., eds), pp. 412–418, American Society for Microbiology, Washington, DC.

Strom, A. R. & Kaasen, I. (1993). Trehalose metabolism in *Escherichia coli*: stress protection and stress regulation of gene expression. *Mol. Microbiol.* **8**, 205–210.

Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Nat. Acad. Sci., U.S.A.* in the press.

Van der Ploeg, J., van Hall, G. & Janssen, D. B. (1991). Characterization of the haloacid dehalogenase from *Xanthobacter autotrophicus* GJ10 and sequencing of the dhlB gene. *J. Bacteriol.* **173**, 7925–7933.

Verschueren, K. H., Franken, S. M., Rozeboom, H. J., Kalk, K. H. & Dijkstra, B. W. (1993*a*). Refined X-ray structure of haloalkane dehalogenase at pH 6.2 and pH 8.2 and implications for the reaction mechanism. *J. Mol. Biol.* **232**, 856–872.

Verschueren, K. H., Seljee, F., Rozeboom, H. J., Kalk, K. H. & Dijkstra, B. W. (1993*b*). Crystallographic analysis of the catalytic mechanism of ahloalkane dehalogenases. *Nature (London)*, **363**, 693–698.

Weightman, A. J., Weightman, A. L. & Slater, J. H. (1982). Stereospecificity of 2-monochloropropionate dehalogenation by the two dehalogenases of *Pseudomonas putida* PP3: evidence for two different dehalogenation mechanisms. *J. Gen. Microbiol.* **128**, 1755–1726.

Winkler, M. E. (1987). Biosynthesis of histidine. In Escherichia coli *and* Salmonella typhimurium. *Cellular and Molecular Biology* (Ingraham, J. L., Low, K. B., Magasanik, B., Schaechter, M. & Umbarger, H. E., eds), pp. 395–411, American Society for Microbiology, Washington, DC.

Wootton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149–163.

Yang, J. W., Dhamija, S. S. & Schweingruber, M. E. (1991). Characterization of the specific p-nitrophenylphosphatase gene and protein of *Schizosaccharomyces pombe*. *Eur. J. Biochem.* **198**, 493–497.

Edited by F. Cohen

(Received 14 June 1994; Accepted 23 August 1994)