The database of epoxide hydrolases and haloalkane dehalogenases: one structure, many functions

Sandra Barth, Markus Fischer, Rolf D. Schmid, Jürgen Pleiss

Corresponding author:

Jürgen Pleiss

Institute of Technical Biochemistry

University of Stuttgart

Allmandring 31

D-70569 Stuttgart, Germany

Phone: (+49) 711-6853191

Fax: (+49) 711-6853196

Email: Juergen.Pleiss@po.uni-stuttgart.de

Keywords:

Sequence-structure-function relationship, sequence alignment, protein structure, phylogenetic analysis, GXGXS-motif

ABSTRACT

The epoxide hydrolases and haloalkane dehalogenase database (EH/HD) integrates sequence and structure of a highly diverse protein family including mainly the Asp-hydrolases of EHs and HDs but also proteins like the Ser-hydrolases non-heme peroxidases, prolyl iminopetidases or 2-hydroxymuconic semialdehyde hydrolases. These proteins have a highly conserved structure, but display a remarkable diversity in sequence and function. 305 protein entries were assigned to 14 homologous families, forming two superfamilies. Annotated multisequence alignments and phylogenetic trees are provided for each homologous family and superfamily. Experimentally derived structures of 19 proteins are superposed and consistently annotated. Sequence and structure of all 305 proteins were systematically analysed. Thus, deeper insight is gained into the role of a highly conserved sequence motifs and structural elements. The EH/HD database is available at http://www.led.uni-stuttgart.de.

Introduction

Epoxide hydrolases (EH, E.C. 3.3.2.3) are ubiquitous enzymes that catalyse the hydrolysis of epoxides to the corresponding vicinal diols; haloalkane dehalogenases (HD, E.C. 3.8.1.5) are bacterial enzymes that cleave carbon-halogen bonds in halogenated aliphatic hydrocarbons; haloperoxidases (E.C. 1.11.1.10) catalyze perhydrolysis of carbonic acids; proline iminopeptidases (E.C. 3.4.11.5) catalyze release of a N-terminal proline from a peptide. These are four examples of the huge family of enzymes with a highly conserved structure, but a remarkable diversity in sequence and biochemical activity. They belong to the α/β hydrolase fold family and consist of two domains: the α/β hydrolase domain and the cap-domain (Holmquist, 2000). It has been shown previously that two loops of variable length can be used to classify the family of EHs (Barth et al., 2004): the NC-loop connects the two domains, the cap-loop is inserted in the cap-domain. Both loops seem to interact with the substrate and are involved in substrate specificity.

Although the enzymes differ in their catalytic residues and in details of the reaction mechanisms, catalysis follows the same scheme (Holmquist, 2000). To identify sequence motifs and structural elements which are common to all members of this diverse family, we systematically compared sequence and structure of all homologous proteins of this family. Despite their obvious difference in sequence and function, there seems to be a similar principle in substrate recognition and catalysis.

Database construction and analysis

The epoxide hydrolases and haloalkane dehalogenases (EH/HD) database is based on the Lipase Engineering Database (LED) (Fischer and Pleiss, 2003) and combines information on sequence, structure, and function of EHs, HDs and related enzyme families. 95 EHs and 18 HDs were collected from Genbank (Benson et al., 2002) by keyword search and sequence similarity using BLAST (Altschul et al., 1997), and analyzed by multisequence alignments

and phylogenetic analysis. Representative sequences were used as a seed for further BLAST searches at a cutoff of $E = 10^{-10}$. Sequence and annotation information were extracted as described previously (Fischer and Pleiss, 2003). Sequence entries with a sequence identity of more than 95% were pooled into one protein entry. Proteins were assigned to homologous families and superfamilies which were determined by phylogenetic analysis using neighbour joining trees provided by ClustalW (Thompson et al., 1994). Homologous families were defined by a high global sequence similarity, while in superfamilies at least the catalytic residues had to be conserved. Available structural information was extracted from the ExPDB (Schwede et al., 2000) and PDB database (Berman et al., 2002). Secondary structure information was created by DSSP (Kabsch and Sander, 1983). For superimposition, structurally conserved residues (residues of the catalytic triad and the oxyanion hole) were used (Pleiss et al., 1998).

Multisequence alignments using ClustalW were performed using the Gonnet 250 score matrix. Multisequence alignments were generated for each superfamily and homologous family using one representative sequence per protein. In the multisequence alignments functionally relevant residues are annotated by colour-coding. The maximum-likelihood method was chosen for phylogenetic analysis of superfamilies and homologous families using Tree-Puzzle (Schmidt et al., 2002), visualized by Phylodendron, and manually edited. The phylogenetic trees demonstrate that the homologous families differ in their sequence diversity: while members of eukaroytic families are highly conserved, microbial families usually show more sequence diversity.

The EH/HD database is accessible at http://www.led.uni-stuttgart.de, “Epoxide hydrolase and haloalkane dehalogenase families”. To access the HTML pages any JavaScript capable WWW browser can be used. Information on protein name, source organism, accession code, link to the corresponding GenBank entry, and a short description of the sequence entry is

provided. Each family is linked to annotated multisequence alignments, phylogenetic trees, and superposed structures.

Substrate specificity

The EH/HD database contains 305 proteins and 397 protein sequences. For 19 proteins, structure information is available (59 entries). 48% of all protein entries are marked as putative in the original databases (Swiss-Prot, PIR, GenBank). All proteins belong to the GX-class of α/β hydrolases, as derived from sequence and structure of the oxyanion hole (Pleiss et al., 2000), and could be assigned to only two superfamilies, soluble and microsomal hydrolases. The soluble hydrolase superfamily contains 267 proteins, 340 sequences, and 57 structures. It consists of 13 homologous families: 8 families of hydrolases with an Asp-based mechanism (5 families of EHs, 2 families of HDs, and 1 family of haloacid dehalogenases) and 5 families with a Ser-based mechanism (1 family of non-heme haloperoxidases, 2 families of meta cleavage compound hydrolases (2-hydroxymuconic semialdehyde hydrolases and 2-hydroxy-6-phenylhexa-2,4-dienoic hydrolases), 1 family of esterases/lipases/peptidases (prolyl-aminopeptidases and a few luciferases and esterases), and 1 family of miscellaneous hydrolases). The microsomal hydrolase superfamily contains 38 proteins, 57 sequences, and 2 structures. In contrast to the soluble hydrolase superfamily, it consists of only 1 homologous family of EHs.

The α/β hydrolases collected in the EH/HD database constitute a diverse group of enzymes with sequence identities below 15 % and very different enzymatic functions, though their structures are highly conserved. While the chemical structure of the substrates differs (epoxides, haloalkanes, carboxylic acids, esters, peptides), the length of NC- and cap-loop seems to be predictive of the shape of the substrates. It has been shown previously that EHs of cluster I and III prefer bulky epoxides: long aliphatic fatty acids (cluster I) or polycyclic aromatic hydrocarbons (cluster III), respectively (Barth et al., 2004). Cluster II hydrolases are

characterized by their preference of small substrates: styrene oxide (bacterial EHs) (Steinreiber and Faber, 2001), 1,3-dibromopropane (HDs) (Bosma et al., 2003), acetate (haloperoxidase) (Picard et al., 1997), and Pro-X (proline iminopeptidase) (Bolumar et al., 2003). Thus, it seems that prediction of substrate specificity by loop length classification can be generalized to all members of this family.

Modular architecture

Although the enzyme families of the EH/HD database are different in sequence and function, they are highly conserved in structure. All structures consist of the modular architecture described for EHs (Barth et al., 2004): N-terminal catalytic domain, NC-loop, cap domain, cap-loop, and C-terminal catalytic domain. Superimposition of the structures of the EH/HD database using the residues of the catalytic triad and the oxyanion hole showed that all β-strands and α-helices of the α/β hydrolase fold are well conserved. The cap domains are also similar in shape and size, and consist of 4 to 5 α-helices arranged in 2 layers. The classification of EHs based on the length of NC- and cap-loop (Barth et al., 2004) was applied to 11 homologous families. Cluster I includes soluble mammalian and plant EHs. All families of HDs, haloperoxidases, proline iminopeptidases, and 2-hydroxymuconic semialdehyde hydrolases belong to cluster II of bacterial EHs. Cluster III includes the microsomal hydrolyse superfamily.

Despite their sequence diversity, all enzymes share three highly conserved sequence motifs: the GXSXG / GXDXG motif of the catalytic nucleophile, the HGX motif of the oxyanion hole, and the GXGXS-motif. The residues of the GXGXS-motif form a structurally highly conserved loop located between the first two β-strands of the α/β hydrolase fold and the cap domain, like a coin in a coin slot. The three elements are tightly linked by hydrogen bonds thus locking the cap-domain to the N-terminal half of the protein. The role of the highly conserved GXGXS-motif is still under discussion. It has been suggested that the first, mostly

aromatic X_1 of the GX_1GX_2S-motif stabilizes the oxyanion hole by interacting with the His of the conserved HGX-motif (Rink et al., 1997). However, this concept only holds for EHs, some HDs, and the non-heme peroxidases, but not for all 14 homologous families, since for many members X_1 is not aromatic. In addition, all GX-type hydrolases contain a conserved HGX motif (Fischer and Pleiss, 2003) in the oxyanion hole but only in members of the EH/HD database the GXGXS-motif is found. We therefore suggest a different role of the GXGXS motif: in hydrolases which contain the first two β-strands and an EH-like cap-domain it stabilises the interface between the α/β hydrolase domain and the cap-domain by a hydrogen network and thus guarantees structural integrity.

ACKNOWLEDGEMENTS

This work was supported by the German Federal Ministry of Education and Research (project PTJ 31/0312702) and by BASF AG.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* **25**, 3389-3402.

Barth,S., Fischer,M., Schmid,R.D., and Pleiss,J. (2003) Sequence and structure of epoxide hydrolases: a systematic analysis. *Proteins,* **in press**.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A., and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res,* **30**, 17-20.

Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S., Fagan,P., Marvin,J., Padilla,D., Ravichandran,V., Schneider,B., Thanki,N., Weissig,H., Westbrook,J.D., and Zardecki,C. (2002) The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr,* **58**, 899-907.

Bohac,M., Nagata,Y., Prokop,Z., Prokop,M., Monincova,M., Tsuda,M., Koca,J., and Damborsky,J. (2002) Halide-stabilizing residues of haloalkane dehalogenases studied by quantum mechanic calculations and site-directed mutagenesis. *Biochemistry,* **41**, 14272-14280.

Bolumar,T., Sanz,Y., Aristoy,M.C., and Toldra,F. (2003) Purification and characterization of a prolyl aminopeptidase from *Debaryomyces hansenii*. *Appl Environ Microbiol,* **69**, 227-232.

Bosma,T., Pikkemaat,M.G., Kingma,J., Dijk,J., and Janssen,D.B. (2003) Steady-state and pre-steady-state kinetic analysis of halopropane conversion by a *Rhodococcus* haloalkane dehalogenase. *Biochemistry,* **42**, 8047-8053.

Fischer,M., and Pleiss,J. (2003) The Lipase Engineering Database: a navigation and analysis tool for protein families. *Nucleic Acids Research,* **31**, 319-321.

Holmquist,M. (2000) Alpha/Beta-hydrolase fold enzymes: structures, functions and mechanisms. *Curr Protein Pept Sci,* **1**, 209-235.

Kabsch,W., and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.

Picard,M., Gross,J., Lübbert,E., Tölzer,S., Krauss,S., van Pée,K.-H., and Berkessel,A. (1997) Metal-free bacterial haloperoxidases as unusual hydrolases: activation of H_2O_2 by the formation of peracetic acid. *Angew Chem Int Ed Engl,* **36**, 1196-1199.

Pleiss,J., Fischer,M., Peiker,M., Thiele,C., and Schmid,R.D. (2000) Lipase engineering database - Understanding and exploiting sequence-structure-function relationships. *Journal of Molecular Catalysis B-Enzymatic,* **10**, 491-508.

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.

Rink,R., Fennema,M., Smids,M., Dehmel,U., and Janssen,D.B. (1997) Primary structure and catalytic mechanism of the epoxide hydrolase from *Agrobacterium radiobacter* AD1. *J Biol Chem,* **272**, 14650-14657.

Rink,R., Kingma,J., Lutje Spelberg,J.H., and Janssen,D.B. (2000) Tyrosine residues serve as proton donor in the catalytic mechanism of epoxide hydrolase from *Agrobacterium radiobacter*. *Biochemistry,* **39**, 5600-13.

Schanstra,J.P., Kingma,J., and Janssen,D.B. (1996) Specificity and kinetics of haloalkane dehalogenase. *J Biol Chem,* **271**, 14747-14753.

Schmidt,H.A., Strimmer,K., Vingron,M., and von Haeseler,A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics,* **18**, 502-4.

Schwede,T., Diemand,A., Guex,N., and Peitsch,M.C. (2000) Protein structure computing in the genomic era. *Res Microbiol,* **151**, 107-12.

Steinreiber,A., and Faber,K. (2001) Microbial epoxide hydrolases for preparative biotransformations. *Curr Opin Biotechnol,* **12**, 552-8.

Thompson,J.D., Higgins,D.G., and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* **22**, 4673-4680.